© 2023 Khanh Linh Hoang

NATURAL LANGUAGE PROCESSING TO SUPPORT EVIDENCE QUALITY ASSESSMENT OF CLINICAL LITERATURE

BY

KHANH LINH HOANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Associate Professor Halil Kilicoglu, Chair and Director of Research
Professor Bertram Ludäscher
Associate Professor Jana Diesner
Associate Professor Richard David Boyce

Abstract

Evidence Synthesis is the process of synthesizing information from clinical literature to translate the research findings into patient care and healthcare policy. Throughout the evidence synthesis process, a critical yet challenging step is the quality assessment of clinical studies. Quality in research can be considered through two aspects: methodological quality which concerns how rigorously a research is designed and conducted, and reporting quality which describes how transparently a piece of scientific work is reported as a publication. This thesis explores natural language processing (NLP) approaches to support evidence quality assessment of clinical studies. Specifically, I consider different levels of information granularity used for evidence assessment, and implemented three machine learning developments: (1) Classification of evidence types from clinical publications based on study designs, (2) Classification of sentences from randomized controlled trials (RCTs) with checklist items recommended in reporting guidelines, (3) Extraction of fine-grained methodological characteristics from RCTs to assist methodological quality assessment. Applications of these NLP approaches range from assisting authors in checking their manuscripts for compliance with reporting guidelines and supporting journal editors and peer reviewers in assessing papers (pre-publication) to assisting systematic reviewers in synthesizing evidence and meta-researchers in studying research rigor and transparency (post-publication).

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Evidence synthesis

Beginning in from the late 80s and early 90s, there has been a pronounced shift in healthcare towards **Evidence-based Medicine** (EBM), which brings research evidence, clinical judgement, and patient values preferences together to support decision-making in patient care (Figure 1.1) [8]. Before EBM, health professionals relied on the advice of more experienced colleagues, often taken at face value, their intuition, and on what they were taught as students. Experience is subject to flaws of bias and what we learn as students can quickly become outdated. EBM is a medical paradigm that emphasizes on the integration of the best research evidence with clinical expertise and patients' values [8]. The idea is that rather than relying on clinical experience alone for decision making, health professionals need to use clinical experience together with other types of evidence-based information, such as information available from the scientific literature. For example, physicians can use their clinical skills and prior experience to identify each patient's unique clinical situation, and based on the evidence findings that is learned from existing literature of the same situations, they can come up with potential interventions tailored for the patient conditions accordingly [9].

The process of obtaining and integrating scientific evidence from literature is called **Evidence Synthesis** [10], which is designed to help physicians to synthesize all available evidence for a given clinical question in a systematic method. **Evidence**, in general, can be defined as *"information which supports or contradicts a hypothesis or a claim derived from scientific research"* [11]. In EBM context, the evidence synthesis process often takes form of systematic reviews or meta-analyses, where relevant existing clinical studies are considered as "evidence" to answer a specific clinical question. Both kinds of research try to answer a defined research

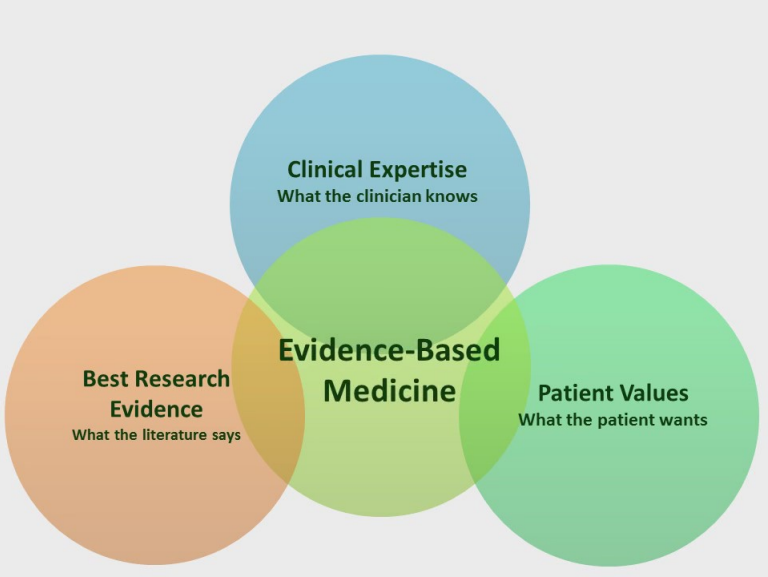

Figure 1.1: Evidence-based Medicine Components [1]

question by collecting and summarising all empirical evidence that fits pre-specified eligibility criteria [12]. However, a meta-analysis is different from a systematic review that it uses statistical methods to summarise results of clinical studies. The process of conducting systematic reviews or meta analysis consists of five fundamental steps [13] (Figure 1.2):

(1) *Ask and formulate a clinical question*: A question arises out of a clinical situation and is asked by clinicians. Systematic reviewers/meta-analysts need to construct an appropriate clinical question that can be answered through evidence synthesis. The question is often constructed using PICO framework, which stands for Participants, Interventions, Comparison (often combined with Interventions) and Outcomes [14]. Using this framework helps a clinician articulate the important parts of the clinical question most applicable to the patient and facilitates the searching process by identifying the key concepts for an effective search strategy [14].

(2) *Find relevant papers from biomedical literature to acquire evidence*: systematic reviewers/meta-analysts search for relevant clinical studies (which resulted in publications) from multiple resources, screen through the study papers (often via the paper abstracts) and select the ones that are relevant and potentially contain evidence to answer the clinical question, also called included studies.

(3) *Assess clinical research quality*: systematic reviewers/meta-analysts examine clinical studies for their validity and clinical usefulness. More specifically, they look at the full-text clinical papers, extract information and use it to assess quality of the studies.

(4) & (5) Implement and evaluate: systematic reviewers/meta-analysts apply the evidence to answer the original clinical question by synthesizing outcomes from the included studies and published the synthesized

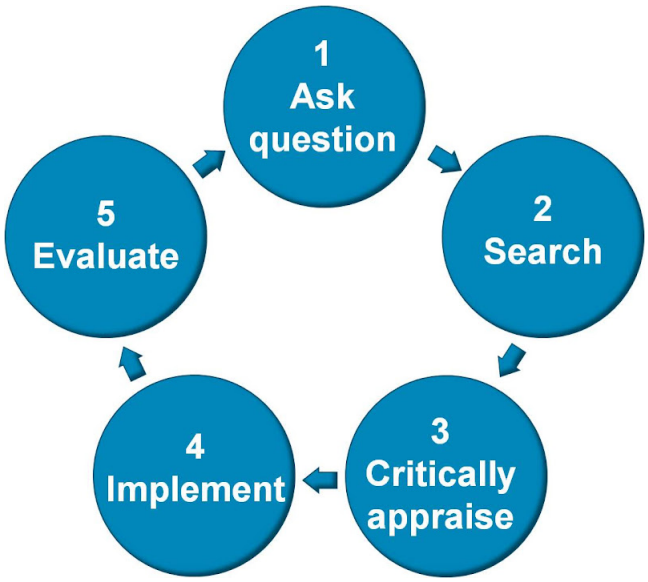

Figure 1.2: Steps to practice EBM in the form of systematic review [2]

evidence in a final systematic review paper.

An example for an evidence synthesis workflow could be: to discover new effective pharmaceutical treatments for COVID-19, physicians may ask a specific clinical question: *"Can chloroquine be used for the treatment of COVID-19?"*. A synthesis process then is conducted by a group of researchers in the form of a systematic review study to answer this question. The process starts with online searches for relevant clinical studies on the use of chloroquine in patients with COVID-19 from bibliographic databases and other resources. Following screening of titles and abstracts, only a small number of studies is included into the final review. These studies are then assessed for their quality. Assessing the quality of individual studies within the synthesis process enables researchers to determine how much confidence they can have of the study findings. To do so, researchers look at the full-text of these studies and extract specific information to make quality assessments (e.g., study design, sample size, intervention groups, etc.). Once researchers are confident about quality of these studies, they can make recommendations about using chloroquine for COVID-19 treatment which are synthesized from the studies' findings [15].

1.1.2 Evidence quality assessment

Throughout the evidence synthesis process, the most critical and yet challenging step is assessing quality of the included clinical studies– **Evidence Quality Assessment (EQA)** (also called **Risk of Bias Assessment** or **Critical Appraisal**). Physicians need reliable information about what might harm or help patients when they make healthcare decisions. Research involves gathering data, then collating and analysing it to produce

meaningful information. However, not all clinical research is rigorously designed and implemented. Many studies are biased and their results are false [16]. This can lead us to draw false conclusions [17] [18]. So, how can we tell whether a piece of research has been done properly and that the information it reports is reliable and trustworthy? How can we decide what to believe when research on the same topic comes to contradictory conclusions? This is where evidence quality assessment helps. According to the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group, the quality of "evidence", in this context a clinical study, *"reflects the extent to which physicians can be confident that a finding from the study is adequate to support a particular recommendation"*. EQA is designed to help reviewers to decide whether studies have been undertaken in a way that makes their findings reliable, make sense of the results, and know what these results mean in the context of the decision they are making. During this process, researchers/reviewers often try to answer questions about the clinical study which will allow them to assess the validity, the usefulness and clinical applicability of the study, or to recognize any potential for bias, so as to eliminate irrelevant or weak studies. Examples of questions often asked by reviewers are: *What type of research question is being asked?*, *Was the study design appropriate for the research question?*, *Did the study methods address the potential sources of bias?*, *Were the statistical analyses performed correctly?*, Or *do the data justify the conclusions?* [19]. To answer these questions, reviewers need to look for some certain information from the publication of the clinical study and use it as criteria to assess quality of the study accordingly. The best practice of evidence assessment is for at least two domain experts to independently assess each included study and then to reach consensus about the final assessment. Various guidelines and assessment tools have been developed to provide a structured approach to the process of critical appraisal for reviewers [20].

Quality in clinical research can be considered through two dimensions: **methodological quality** and **reporting quality** [21] [22]. Methodological quality concerns how rigorously a research was designed and conducted [23]. Reporting quality describes how transparently a piece of scientific work is reported in a publication [24].

In terms of methodological quality, the most basic information used to assess methodological quality of a clinical study is its study design, which refers to the general plan by which a study was or is to be carried out [25]. The lack of certain qualities in the study design may make the study prone to bias, thus weakening findings and conclusions of the clinical study. Therefore, different designs of studies yield different strengths of evidences. For example, the evidence pyramid is widely used as the most generic guideline for levels of evidence quality based on different study designs (Figure 1.3). Each level of the pyramid is a kind of research design from which the evidence comes. The higher the level of research design in the pyramid, the more reliable its evidence is considered.

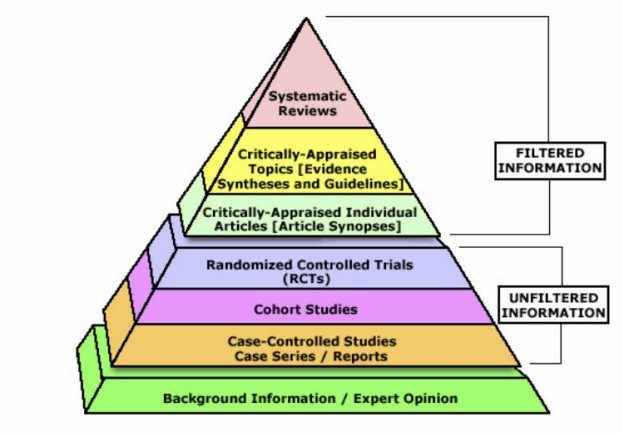

Figure 1.3: Evidence Pyramid [3]

Study design information is often explicitly mentioned in titles and abstracts of clinical studies. For this reason, it is usually not difficult for researchers/systematic reviewers to identify the information. However, it is not enough to make assumptions about the quality of a study based purely on the design. Indeed, besides study design, the actual execution of the study methodology is also important. Subsequently, depending on the study design, different and more fine-grained information will be used to assess evidence quality. For example, in a **randomized controlled trial (RCT)**, a study design that randomly assigns participants into an experimental group or a control group, information to assess evidence quality is often randomization related, such as sequence generation, allocation concealment or blinding [18]. On the other hand, for observational studies, a study design that researchers observe the effect of a risk factor, treatment or other intervention without trying to change who is or is not exposed to it, information to assess evidence quality is often procedure related, such as eligibility criteria, control confounding and follow-up [26]. Even though fine-grained methodological criteria by which the quality of a study is assessed will vary according to its study design, some general methodological characteristics such as significance level, power, drop-out rate to calculate samples size, underpin the evaluation of studies of many different designs [27].

Reporting quality can be assessed based on reporting guidelines which are presented in the form of a checklist of recommendations, often specific for each study design [28]. These guidelines provide detailed lists of information items recommended by the experts to be presented in the published paper in order to make the study easier to evaluate and reproduce. The most intuitive way to assess reporting quality is to compare the information items in these professional recommendations with information presented in the full-text paper of a clinical study. By cross-checking the information items reported in the actual publications vs. the information items recommended by the checklists, journal editors will be able to assess the adherence

to reporting guidelines of the studies and can then require the authors to report the missing items as needed. Examples of such guidelines are CONSORT for RCTs [29], STROBE for observational studies [30], ARRIVE for pre-clinical animal studies [31], PRISMA for systematic reviews [32], all developed under the Enhancing the Quality and Transparency Of Health Research (EQUATOR) Network [33]. Reporting guidelines have been positively viewed, in the form of endorsement by medical journals (for example, the CONSORT Statement is endorsed by over 600 journals [34]). Some journal editorials indicate their support, while others institute mandatory submission of a guideline checklist and/or flow diagram along with manuscript submission.

Methodological quality and reporting quality are not always and necessarily related. A well-designed study can be poorly reported, while a well-written paper may resulted from a poorly designed and implemented study. An example of assessing methodological quality versus reporting quality of clinical studies is to consider two RCTs, both were conducted to investigate whether hydroxychloroquine could reduce COVID-19 severity in adult patients [35], [36]. The first study involved hospitalized 11,197 patients and was conducted at 176 hospitals in the United Kingdom [35]. The second study involved 667 hospitalized patients with suspected or confirmed COVID-19 and also was conducted at 55 hospitals in Brazil [36]. Considering two methodological characteristics: sample size and settings, we can see that the second study has a much smaller sample size compared with the first study. In addition, even though both studies used multi-center settings, the second study was conducted on a smaller scale (fewer hospitals) than the first one. Therefore, in terms of methodological quality, the results of the first study are more likely to be generalizable, thus have better quality, than the second study. On the other hand, in terms of reporting quality, it is undetermined which study has better reporting quality since both of them report this information in their publications.

1.1.3 Problems with assessing clinical research quality

For researchers synthesizing evidence, assessing the rigor (methodological quality) and transparency (reporting quality) of clinical research from the literature in a timely manner is an extremely difficult task. There are multiple factors contributing to this. However, the most outstanding difficulties come from two aspects: the size and growth of medical literature, and the extensive domain expertise needed to fulfill the task.

Problem 1: Size and growth of the medical literature

An obvious and worsening barrier to the implementation of evidence synthesis is the fastly growing body of medical literature. Hundreds of thousands of medical research papers are being published every year (at a rate of at least one every 26 seconds [37]), making it almost impossible for researchers to keep up with the research progress even within a narrow research topic. A simple search query of clinical papers about

COVID-19 on PubMed, the primary bibliographic database in the biomedical domain, shows that about 290,000 research on this topic were published in the last 3 years[1]. The amount of literature needs to be screened in a typical systematic review is often large: in a typical systematic review, over 2000 abstracts need to be reviewed in order to find 15 relevant studies [38]. It often takes months (or even years) to complete a thorough systematic review (a mean of 67 weeks from deposit of a protocol to publication of the review [39]). And by the time the review is done, new relevant evidence might be released from newly published research that makes findings from the review obsolete. Peer review process is being used by conferences and journals as the medium to assess research quality [40]. Peer reviewers are expected to examine papers on potential shortcomings. However, with the increase in published papers, the need for peer reviewers has skyrocketed and the amount of time required for a thorough review is high [41]. The time a peer reviewer spends on reviewing a manuscript is often limited and consequently, it is not uncommon for a reviewer to miss crucial shortcomings of the manuscript, allowing the dissemination of low-quality publications [42].

Problem 2: Extensive domain knowledge needed

In assessing quality of a clinical research, while some methodological characteristics are often explicit and easy to recognize, such as "study design", others are hard to identify as they require intensive domain knowledge to fulfill the task. This is equivalent not only to the ability to assess quality based on the surface information, but also being able to identify underlying quality issues by cross-examining information at different level of granularity.

For example, in assessing methodological rigor, a common weakness is the use of inadequate sample size. For scientific and ethical reasons, the sample size for a trial needs to be planned carefully, with a balance between medical and statistical considerations [43]. Ideally, a clinical study should have a sample size that is large enough to have a high probability (the probability that the test correctly rejects the null hypothesis) to detect statistically significant differences between treatment outcomes. Despite the necessity of having sufficient sample size, many published randomised trials have low statistical power [44], or fail to calculate or report power analyses [45]. An analysis of 136,212 clinical trials between 1975 and 2014 shows that even though statistical power of clinical trials increased over time, the number of trials with power more than 80 was still low [46]. The consequence of this is overestimation of effect size and low reproducibility of results, which eventually undermine the reliability of the studies. That being said, to assess the sufficiency of sample size of a clinical study, only looking at the reported sample size is not sufficient. Researchers need to consider multiple statistical factors, such as power value, P-value or dropout rate [47] in order to examine if the sample size calculation is done properly and correctly.

[1] https://www.ncbi.nlm.nih.gov/research/coronavirus/

Or, to assess reporting transparency, another common quality issue is the inconsistency in how the methodology is described in a clinical publication. For example, a study claims to be "double-blind", but only indicates one object who is involved the study (such as "patient"), being blinded. In an analysis of 622 RCTs done by Penic et al. in 2020, even though 62% trials explicitly claimed to be "double-blind" randomized trials, the blinded objects (e.g. patients, healthcare providers or data collectors) were only reported in 14% of the papers. The study suggested that the "double-blind" term is overused and ambiguous [48]. As a result, the term should not be the sole reason for researchers to make judgments in terms of blinding quality. In fact, key individuals involved in RCTs should be reported and considered as factors for quality judgement.

Experts often need to reference consensus methodology and reporting quality guidelines (such as CONSORT checklist for RCTs or STROBE for observational studies), and discussion between assessors is needed to ensure that these factors have been considered appropriately [49]. However, despite the existence of assessment guidelines as well as the endorsements from the publishing journals to encourage authors to use them, the adoption rates of these guidelines from authors are still inadequate [50].

In recent years, the scientific community has witnessed the rise of **Meta-research** (also called **Meta-science**), a relatively new field that has its roots in traditional meta-analysis and systematic reviews which aims to study research itself [51]. Meta-researchers who conduct meta-research studies have the main purpose of evaluating certain methodological aspects of published research, which includes methodological and reporting quality characteristics from published research. However, since meta-research studies assess quality of previous research from multiple aspects, meta-researchers are facing the same time-consuming and intensive human labor issues which make it very difficult to scale those studies.

1.1.4 Biomedical Natural Language Processing and its application in clinical literature quality assessment

The intense cost in time and effort has led to the development of computer support tools for the evidence synthesis process. Previous research found that systematic reviewers typically use software such as EndNote, Reference Manager, RefWorks, and Excel to manage references [52]. Some commercial products are designed as end-to-end support tools: DistillerSR[2] and Covidence[3] primarily provide an integrated environment for data capture and management, for tasks such as harvesting search results from databases, screening studies, and providing questionnaires for manual data extraction for quality assessment. Another end-to-end tool, EPPI-Reviewer[4], provides (and continues to develop) advanced features such as automatic term reorganization,

[2]https://www.evidencepartners.com/
[3]https://www.covidence.org/
[4]https://eppi.ioe.ac.uk/cms/

and document clustering and classification, using machine learning and data mining. Nevertheless, in a previous interview study with 16 systematic reviewers, we identified a gap between the technology support available and what technology is being used by our reviewers [53]. Despite the existence of the above end-to-end advanced technology support, most steps are still done manually, making the review process more time-consuming and inefficient than it needs to be [53]. We also found that automation is increasingly proposed for systematic review researchers, often involving machine learning (ML) and natural language processing (NLP). Such an approach can also play a role in supporting evidence synthesis process in general, and in assessing quality of clinical research in particular [53].

Biomedical text mining and natural language processing (BioNLP) is a research domain that deals with processing data from journals, medical records, and other biomedical documents (in which, a large part of them are clinical publications) to understand biomedical text and extract knowledge from it. The applications of BioNLP range from identification of biological entities (such as proteins, genes, chemical compounds, drugs, or disease names), to classification of biomedical documents based on their contents and topics, or to support information retrieval by identifying documents and concepts matching search queries [54]. As for evidence synthesizing in particular, BioNLP has been used to develop informatics systems that support or automate the processes of systematic review or each of the tasks of the systematic review [55]. In a most recent review of automation for systematic reviews, Dinter in 2020 listed 41 primary research working on automating different steps in the systematic reviewing process, including: searching, screening and selecting relevant studies, data extraction, and study quality assessment [56]. Based on the review results, *"selection of primary studies"* step was automated most often; whereas, automation of *"evaluation of the selected primary studies"* had fewest number of relevant studies.

Although there is less work on automated research quality assessment, Kilicoglu in 2018 in a review of biomedical text mining for research rigor and integrity proposed several directions text mining tools can help [57]. Two directions are relevant to quality assessment: (1) managing information overload by summarizing and aggregating knowledge derived from the publications including claims, hypotheses, supporting evidence; and (2) assessing adherence to reporting guidelines by assessing a manuscript against the relevant reporting guidelines and flagging reporting quality issues [57]. Marshall and Wallace, similarly, suggested the use of NLP tools to expedite evidence quality assessment process which entails both a data extraction task (identifying snippets of text in the article as relevant for bias assessment) and a text classification task to predict an article as being at high or low risk of bias [58]. In fact, in the last 5 years, a growing number of NLP models have been developed to help researchers assess the quality of evidence extracted from clinical literature, divided into two categories. The first category of models focuses on automatic prediction of the evidence quality levels in a more direct manner (*text classification*) [59], [60]. For example, given a clinical

paper, text classification models directly predict the level of risk of bias of the study (either high or low) purely based on the text of the paper. The second group of automation focuses on automatic extraction of information items (such as study design, participants, interventions) that can be used to assess the evidence quality from the abstracts or full texts of the related clinical papers (*information extraction*) [61], [62].

Despite the effort, current NLP approaches to assist clinical research quality expose four major shortcomings:

(1) Most information extraction tools focus on a limited set of information, mostly PICO characteristics such as inclusion and exclusion criteria of patients, interventions or outcome measures of a clinical study [62]. However, PICO framework is mainly used by practitioners of EBM to form clinical questions and facilitate literature search [63]. Therefore, some aspects of the PICO characteristics are relevant for quality assessment purpose (e.g., number of participants), they are not sufficient to address the question of research quality, instead more focusing on what the study is about.

(2) Although some extraction tools have started to consider fine-grained information such as study design, sample size, none of the extraction tools provides a comprehensive extraction of methodological information that can be used for evidence quality assessment.

(3) Prediction tools using text classification focus on making risk of bias assessment directly. Their output is a risk of bias level (e.g. high, medium or low risk of bias) without having an explanation of why or what criteria the systems use to make their decision [60]. While the predictions are sometimes supported by highlighting relevant sentences like in Marshall et al. [59], the rationale between the predictions can still be opaque.

(4) While existing automation approaches are sufficient to assist researchers in assessing reporting quality at high level (e.g. by identifying sentences that describe certain information), the lack of fine-grained information might prevent researchers/reviewers from identifying methodological flaws and inconsistencies that are not easily identified in the clinical papers. For example, in a study to examine methodological issues of 80 RCTs which was done manually by Altman and Dore in 1990, the authors found a information mismatch issue between randomization block size and the number of patients in each treatment group described in the following quotes *"1 trial of 30 patients used inappropriately large blocks of 20"* and *"When blocking is used without stratification the maximum difference between the numbers in the two groups should be half the block size; this was not the case in 2 trials."* [64]. In such cases, capturing that a sentence is about randomization or that the article has high risk of bias is not sufficient to identify the mismatch. Instead identifying fine-grained information such as type of randomization (in this case is "block randomization") and corresponding attributes (in this case is "block size of 20") is needed to detect the underlying methodological issues and, consequently, assist with methodological quality assessment.

1.2 Thesis statement

Understanding the pain points that researchers and other stakeholders (journals, peer reviewers, systematic reviewers, meta-researchers) have been experiencing with assessing quality of clinical research– the deluge of information from the bulk of clinical literature and the intensive domain knowledge need; along with the drawbacks of current NLP attempts to address these two problems that we discussed above, in this thesis, I study different NLP approaches that can assist researchers and other stakeholders to perform the evidence quality assessment task. Toward that objective, two main research questions which I try to answer through the thesis are:

- ***Research question 1: What information do biomedical researchers and other stakeholders need to assess evidence quality?***
- ***Research question 2: How can we use NLP and ML techniques to automatically extract them?***

To answer these questions, I have specifically focused on two main quality aspects of clinical research: **methodological quality** and **reporting quality**, and identify information that can be used to assess these quality aspects. Figure 1.4 shows summary of the thesis focus, which is evidence quality assessment, along with the stakeholders who can be beneficial from the computer support that the thesis tries to develop. Subsequently, I present three research studies I have conducted and contributed to during my doctoral studies, in which three NLP tasks are proposed and developed to automatically extract the identified information.

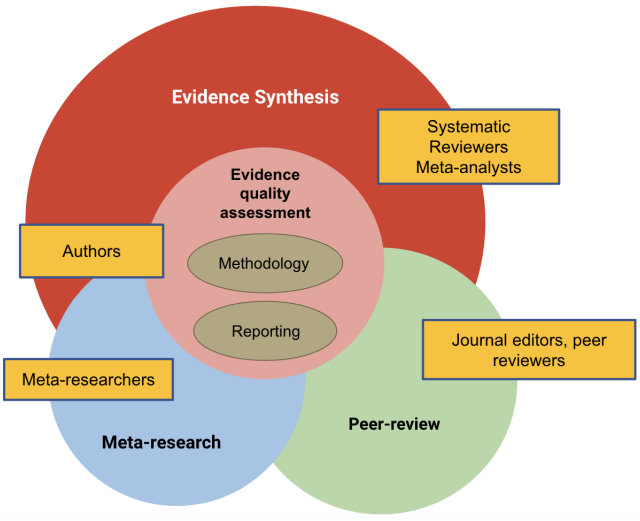

Figure 1.4: Summary of research and the thesis focus

Chapter 2

Literature Review, Current Issues and Thesis Solutions

In this chapter, there are three main contents:

- First, I will review different levels of information granularity that can be use for methodological and reporting quality assessment: starting from the most coarse-grained information which can be identified at publication level, to more fine-grained information which can be identified in the full text publication of a clinical study.
- Second, I then will review in depth the existing computer support for evidence quality assessment, including workflow-based computer support and NLP-based computer support.
- Last but not least, I will discuss in details the shortcoming of the existing NLP-based approaches and describe solutions proposed and developed in this thesis to address those shortcomings.

2.1 Criteria to assess quality of clinical research

Criteria for assessing the quality of primary research emerged in the late 1980s with the rise of EBM. This set the stage for the development of a variety of scales, guidelines and checklists for quality assessment, often associated with particular study designs [65], [66]. Even though the existing checklists and guidelines are varied in terms of criteria and grading scheme [66], as mentioned previously, assessment factors mainly involve two related quality aspects: **methodological** and **reporting** [22]. In a nutshell, **methodological quality** means assessing methodology rigor of a study and often relates to the study design and conduct

of the research [67]; and **reporting quality** means assessing reporting transparency and is in regards of how well a piece of scientific work is reported as an article published by a scientific journal [68]. Often time, the scales are designed to assess methodological quality, in which they give researchers a quantitative index of the likelihood that the reported methodology and results are free of bias. The checklists, on the other hand, are most useful for reporting quality assessment, in which they provide researchers with guidelines as to what information should be included in reporting clinical research. Notice that these two aspects are mutually inclusive since methodological quality only can be assessed through what has been reported in the actual paper of a clinical study. Therefore, sometimes, the two aspects are being assessed simultaneously throughout the assessment process.

2.1.1 Methodological quality assessment

Methodological quality is the extent to which the design and conduct of a trial are likely to have prevented systematic errors and biases. Therefore, assessing methodological quality of clinical research enables researchers determines how well a clinical study was designed and executed, thus to determine how much confidence they can have of the study findings. To assess methodological quality, different levels of methodological information must to be considered, which may vary depending of the type of study and on the subject of research. In the Introduction chapter, we briefly discussed criteria to assess methodological quality such as study designs (through the evidence pyramid). In this sections, we will review in depth criteria for quality assessment at different levels of granularity, as well as the existing scales/guidelines that have been created to assist reviewers to perform the task.

Coarse-grained criteria

The most coarse-grained criteria to assess quality of a clinical study is through its **Study Design**. According to the Clinical Trials Dictionary, study design is *"the general plan by which a study was or is to be carried out, including details on the nature of the study population and data collection procedures and, when appropriate, other details, such as for treatment procedures in the case of trials"* [69]. Study design is considered as the most coarse-grained information for quality assessment because: (1) it could be determined at individual study level, and (2) different study designs indicate different levels of evidence quality.

According to the evidence pyramid (Figure 1.3), there are two groups of evidence quality levels: evidence from primary articles that appear in peer-reviewed journals and can be found by searching databases– called "unfiltered information"; and evidence from filtered resources that summarize and appraise evidence from several studies– called "filtered information". Each group is further divided into different levels– each

ascending level represents a different type of study design and corresponds to increasing rigor, quality, and reliability of the evidence. As we ascend through these different study designs, we become more confident that their results are accurate, have less chance of statistical error, and minimize bias from confounding variables that could have influenced the results. For example, in the group of filtered information study designs, a systematic review would provide the most trustworthy evidence, higher than critically-appraised topics (CATs) and articles (which is often a short summary of evidence or summary of an individual article from the literature). In the group of unfiltered information, randomized controlled trials provide, in general, stronger evidence than observational studies. Rigorous observational studies provide stronger evidence than uncontrolled case report or case series, in which clinicians only examine patients' medical records for exposure and outcome. And finally, the foundation level of the pyramid consists of background information and expert opinion. This is the lowest level of scientific quality and is not considered evidence as such.

Note that in the pyramid, the "filtered information" group contains study designs that are results of evidence synthesis process, and the "unfiltered information" group contains study designs that are input of the process. In this thesis, since we are only looking at study designs that are being used for evidence synthesis, therefore we will only look into the study designs belong to the "unfiltered information" group in further details.

From an epidemiological standpoint, "unfiltered information" group consist of primary research studies with no external appraisal or interpretation provided. And if the research is conducted with human subjects, it is called **clinical study**. In this thesis, since our focus are clinical studies - with human subjects, terms such as "evidence quality", "biomedical research quality" or "clinical research quality" are used interchangeably and meant the same thing: assessing quality of clinical studies as inputs of evidence synthesis process.

There are two major types of clinical study designs: **observational** and **experimental**. Observational studies are hypothesis-generating studies in which the aim is observation without altering or influencing that which is observed, while experimental studies (also called interventional studies) are hypothesis testing studies, in which an experimental treatment is conducted with specified procedure [70]. Then underneath each type, more specific study designs are defined depending on methodological characteristics. Observational study design is further divided into two groups of study designs: analytical, which include case-control and cohort studies, and descriptive which includes cross-sectional studies and case reports. Experimental study design includes clinical trials which involve subjects with a disease and place them into different treatment groups, field trials which involve subjects without a disease being placed in different treatment groups, and community trials which is also known as cluster trials, involve groups of individuals with and without disease who are assigned to different intervention/experimental groups. Clinical trials are further divided into randomized clinical trials and non-randomized clinical trials [70] (Figure 2.1). Since RCT is the most common study

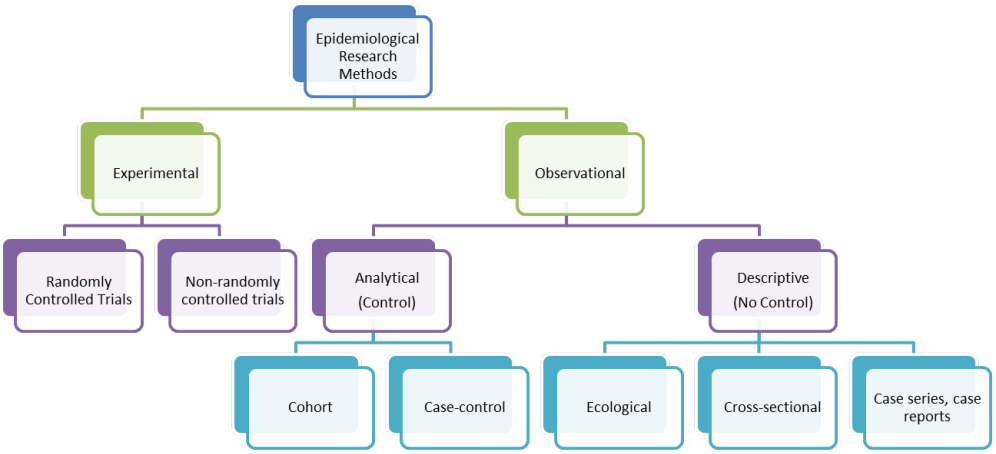

Figure 2.1: Hierarchy of study designs [4]

design from the Experimental branch, and is considered as "gold standard" of clinical research, from now on I will mostly use **RCT** as the representative study kind for the experimental design.

Fine-grained criteria

Different study designs are prone to sources of systematic errors and biases. Therefore, to assess methodology quality from a clinical study, criteria at more fine-grained levels will be varied, depending on the study design of the study. As different scales and guidelines also developed with different quality assessment criteria specialized to the study designs.

RCT, which is generally considered as the gold standard of experimental study design, therefore, has most established appraisal guidelines for it. In RCT study design, patients are randomly assigned into an experimental group or a control group [71]. To assess quality of RCTs, reviewers often look for key methodological factors that reflect the potential bias of the study. Note that the terms "quality" and "bias" is used interchangeably to grade the methodological quality of RCTs. A risk of bias can arise from critical flaws in methodological design, unreliable or non-reproducible methods or improper or incomplete statistical analysis. For RCTs, the randomization design helps to reduce the risk of bias when testing the effectiveness of new treatments. Therefore, quality assessment factors include how the randomization was implemented and how people who are involved in the study are prevented from being aware of the treatment process. Those factors are: randomization generation (how the randomized treatment assignments are generated); allocation concealment (how the random assignment process is concealed among involved parties), and blinding (how the patients or care providers are kept unaware of the treatment process) [72]. In early days, Moher et al. listed twenty-five scales and nine checklists which were developed to assess the quality of RCT studies [66].

Domain	Description	Review authors' judgement
Sequence generation.	Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.	Was the allocation sequence adequately generated?
Allocation concealment.	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrolment.	Was allocation adequately concealed?
Blinding of participants, personnel and outcome assessors *Assessments should be made for each main outcome (or class of outcomes).*	Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Was knowledge of the allocated intervention adequately prevented during the study?
Incomplete outcome data *Assessments should be made for each main outcome (or class of outcomes).*	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), reasons for attrition/exclusions where reported, and any re-inclusions in analyses performed by the review authors.	Were incomplete outcome data adequately addressed?
Selective outcome reporting.	State how the possibility of selective outcome reporting was examined by the review authors, and what was found.	Are reports of the study free of suggestion of selective outcome reporting?
Other sources of bias.	State any important concerns about bias not addressed in the other domains in the tool. If particular questions/entries were pre-specified in the review's protocol, responses should be provided for each question/entry.	Was the study apparently free of other problems that could put it at a high risk of bias?

Figure 2.2: Domains and criteria to assess risk of bias from Cochrane Handbook [5]

More recently, Ma et al. in 2020 reviewed and compared fourteen assessment guidelines and scales for RCTs [73]. Among those, perhaps the most popular guideline is "Risk of Bias Assessment" developed by Cochrane Collaboration, an international network of professionals who produces high-quality and accessible systematic reviews [18]. In this guideline, different domains and criteria to assess risk of bias of RCT studies are specified in details. Based on these criteria, assessors can create a risk of bias assessment template and try to answer a list of questions regarding these characteristics before making final judgments (e.g. high risk of bias or low risk of bias). Figure 2.2 shows an example of domains, criteria and signaling questions recommended in the Cochrane Handbook from its risk of bias assessment guideline [5]. Along with the guideline, Cochrane also developed a computer-based tool to support reviewers with the assessment process, in which we will discuss further in the computer support section.

Another popular quality assessment guideline is The Grading of Recommendations Assessment, Development, and Evaluation (GRADE), which was developed by multiple international groups of guideline developers, methodologists and clinicians to provide a common, transparent and sensible system for grading the quality of evidence and the strength of recommendations [17]. While Cochrane's guideline mainly focuses on risk of bias assessment in the context of systematic review, the GRADE approach provides a different scheme to rate the quality of evidence which focuses more on study design and methodological aspects of a RCT. According to GRADE, the methodological quality of RCTs could be affected by several design and execution factors, including: lack of concealment, intention to treat principle violated, inadequate blinding, loss to follow-up, early stopping for benefit, selective outcome reporting [17]. Note that some criteria to assess evidence quality

of RCTs recommended by GRADE are similar to those in the Cochrane Handbook (such as "adequate if stating the use of any type of blinding"), some of them are outcome-specific and are at a more fine-grained level. For example, consider the quality assessment of an RCT about the effects of an intervention on acute spinal injury with two different outcome measures: all-cause mortality and motor function. The patients of the study were blinded throughout the whole treatment process. However, the outcome assessors were not blinded for any outcomes. According to the Cochrane signaling questions regarding blinding, the quality of evidence based on blinding methods can be considered high since the blinding mechanism is reported (in this case patients were blinded). However, according to GRADE, blinding methods need to be examined specifically by the objects and by the outcome measures. In this case, blinding of outcome assessors is less important for the all-cause mortality outcome measure, but crucial for the motor function outcome measure. Therefore, based on the GRADE assessment scheme, the quality for the motor function outcome may be downgraded .

Fine-grained criteria to assess observational studies, including cohort studies, case-control studies, cross-sectional studies, and case reports, are different from experimental studies because in this study design, researchers only observe the effect of a treatment or other intervention on patients without trying to interfere with the treatment process [74]. Since there is no randomization factor, assessment criteria of observational studies are involved the risk of failure to develop and apply appropriate eligibility criteria (inclusion of control population), flawed measurement of both exposure and outcome, failure to adequately control confounding, and incomplete or inadequately short follow-up [74]. Similar to experimental studies, scales and checklists have been developed to specify domains and criteria to assess quality of observational studies. These guidelines are also divided to further sub design types such as the Critical Appraisal Skills Programme (CASP) checklists for cohort study vs. case-control study, the Scottish Intercollegiate Guidelines Network (SIGN) critical appraisal checklists for cohort vs. case-control study, the National Institutes of Health (NIH) quality assessment tool for observational cohort and cross-sectional studies.

2.1.2 Reporting quality assessment

Incomplete reporting and lack of transparency in clinical papers can affect the research quality assessment because important methodological details may be missing. When key elements such as randomization/blinding procedures (in a RCT study) are missing, it can be difficult to assess the rigor and reliability (methodological quality) of a study for evidence synthesis. Similar to methodological quality, multiple checklists and guidelines have been developed to improve transparency and accuracy of reporting of clinical research depending on study designs of the studies such as CONSORT for RCTs [75] (and other CONSORT extensions for other

RCT design types such as cluster RCT), ARRIVE for preclinical animal studies [76], and PRISMA for systematic reviews [77], and STROBE for observational studies [30].

Among these, Consolidated Standards of Reporting Trials (CONSORT) is the most well-known checklist that provides a comprehensive list of 25 information items that should be reported in a RCT study. Many of these information items correspond to the criteria to assess methodological quality of RCT such as trial design, randomization generation, allocation concealment, or blinding [34]. The main CONSORT Statement is based on the standard two-group parallel design. However, there are several variations to the standard trial methodology, including different design aspects (e.g., cluster), interventions (e.g. herbals) and data (e.g. harms). Therefore, to help improve the reporting of these trials, the main CONSORT Statement has been extended and modified by members of the CONSORT group for application in these various areas. This results in multiple CONSORT extensions, each corresponding to a specific sub-design of RCT such as cluster trials, non-inferiority and equivalence trials, pragmatic trials, etc. [75]. While the majority of information items in these extensions are similar to each other, some specific fine-grained methodological items are specified corresponding to each type of design. For example, the information item regarding trial design (item 3a) in the standard CONSORT checklist is described as *"Description of trial design (such as parallel, factorial) including allocation ratio"*, while the same information item in the CONSORT cluster trial checklist is described as *"Definition of cluster and description of how the design features apply to the clusters"* [78]. The difference between a standard RCT and a cluster randomized trial is that cluster trials randomize interventions to groups of patients (e.g., families, medical practices) rather than to individual patients. Therefore, a proper reporting of the 3a item in a cluster trial should contain the information of criteria for grouping patients. An example of such description could be: *"We report the rationale and design of a planned pragmatic,* ***cluster*** *randomized, double-blinded trial, with 1500 newly diagnosed individuals with COVID-19 infection, together with up to one close household contact each (1200 contacts), randomized to* ***either vitamin D3 (loading dose, then 3200 IU/day) or placebo*** *in a 1:1 ratio and a* ***household*** *cluster design."* [79].

Similar to the CONSORT statement for the reporting of randomized trials, the Strengthening of the Reporting of Observational Studies in Epidemiology (STROBE) statement was developed with recommendations to improve the quality of reporting observational studies. The STROBE statement consists of a checklist of 22 items. Many of those are methodological information items which are the same in the CONSORT checklist for RCTs, such as study design, participants (eligible criteria) or statistical methods. The list contains methodology-related information items that are specific for observational studies such as exposures, predictors, potential confounders, and effect modifiers [30]. STROBE also developed different versions of the checklist corresponding to sub-types of observational study design, including cohort study, case-control study

and cross-sectional study. Fine-grained information items for each sub-design type are specified accordingly. For example, the information item regarding statistical methods in the STROBE checklist for cohort study is described as *"explain how loss to follow-up was addressed"*, while the same item in the STROBE checklist for case-control study is described as *"explain how matching of cases and controls was addressed"*.

To a certain extent, reporting quality is related and more general than methodological quality. To make methodological quality assessment, researchers often have to rely on how it is reported. For this reason, many criteria used to assess methodological quality are also being used to assess reporting quality. For example, typical methodological information of RCTs such as "randomization", "blinding" or "allocation concealment" are all listed in the CONSORT checklist (as for reporting quality) [29] as well as the GRADE guideline (as for methodological quality) [17]. However, even if a clinical study is well-reported, it does not mean it is a rigorous study. To assess reporting quality, researchers can just purely reply on the text and make a binary judgement if a certain information is reported or not. In contrast, to assess methodological quality, assessors often need to consider multiple methodological characteristics at the same time and at a more detailed level. For example, "blinding", a well-known aspect used to assess both reporting and methodological quality, refers to the concealment of group allocation from one or more individuals involved in a clinical research study (most commonly a RCT) and is a critical methodological feature to reduce risk of bias after randomization [80]. Considering two RCT studies on the same topic: one reported the usage of a "double-blind" method in which both patients and physicians were blinded during the treatment process, and one was an open-label RCT which means it employs no blinding during the study. From the reporting quality assessment point of view, as long as the authors of these two studies properly report blinding information on their publications, it can be considered a good reporting quality. However, from the methodological quality assessment point of view, the former is considered to have higher methodological quality since it employed a blinding mechanism to reduce risk of bias [80].

2.2 Computer support for clinical research quality assessment

The high cost, in terms of time and effort, of evidence quality assessment has led to the development of computer support tools. In terms of computer support for evidence synthesis (in a form of systematic reviewing process) broadly, some commercial products are designed as end-to-end tools, such as DistillerSR [81] and Covidence [82] or EPPI-Reviewer [83] which all provide an integrated environment for data capture and management along with advanced features (such as automatic term reorganization, and document clustering and classification). In this section, I will only review non-commercial tools that have been developed to assist reviewers assessing quality of evidence, which can be divided into two main categories: tools that construct

the assessment workflow and provide templates needed throughout the workflow; and tools that automate some steps of the process using NLP approaches.

2.2.1 Workflow-based computer support

Workflow-based computer support focuses on assisting reviewers by mimicking the assessment workflow and providing centralized platforms that contain templates needed (such as list of questions for reviewers) for the quality assessment. The templates are mainly designed based on existing quality guidelines or checklists to ensure that reviewers do not miss any of the required criteria. In the review of methodological quality (risk of bias) assessment tools, Ma et al. listed 27 workflow-based tools developed by different organizations associated with different study designs [73]. These include tools to assess RCT studies (such as: Risk of Bias tool from Cochrane [18] [5], The Effective Practice and Organisation of Care (EPOC) tool [84]); tools to assess non-RCT studies (such as: the Canada Institute of Health Economics (IHE) Quality Appraisal Tool [85]); tools to assess observational studies such as the NIH quality assessment tool for observational cohort and cross-sectional studies [86].

Among these tools, the Cochrane Risk of Bias tool for randomized trials (which is called RoB) is the most commonly recommended tool for RCTs [18]. A revised revision of this tool (RoB 2.0) was published in 2019 [5]. The RoB 2.0 tool consists of five bias domains (as shown in Figure 2.2) and is suitable for assessing individually-randomized, parallel-group, and cluster-randomized trials. It provides a template interface with lists of signaling questions corresponding to each domain for users to answer. Then underlying reasoning algorithms will combine the answers and use them to make bias judgements. Similar to RoB, ROBINS-I is a tool for assessing risk of bias, however designed for non-randomised studies of interventions [87]. ROBINS-I has the same approach as RoB by providing a template with signaling questions and running an algorithm behind the scene to formulate risk of bias judgements for each of bias domains, informed by answers to the signalling questions. Nevertheless, tools like RoB or ROBINS-I are not fully automated since users' answers for the signaling questions are needed to run the algorithm.

2.2.2 NLP-based computer support

To support researchers in evidence synthesis, NLP has been offered as a potential solution to automate some of the steps throughout this process [58]. Two common categories of NLP-based automated computer support have been developed: ***information extraction*** and ***text classification***. ***Information extraction*** is applied to develop models that can automatically identify snippets of text needed for the synthesis process in general and for quality assessment in particular (e.g. extract the number of patients randomized from a

clinical trial paper). ***Text classification*** is applied to develop models that can automatically categorize texts (in the form of abstracts, full-text, sentences or fragments of sentences) into predefined categories (e.g. determining whether a paper reports an RCT or an observational study). For quality assessment purposes, application of classification is often to predict evidence quality levels (e.g. high quality evidence or low quality evidence) of a clinical study.

Automatic information extraction

Conceptually, information extraction is considered as a sub-task of the evidence quality assessment process. In this task, information items from titles, abstracts or full-text of clinical papers are extracted and represented in a structured template, so that they can be used by researchers during the evidence quality assessment. The information extracted could be in the form of a whole paragraph, a specific sentence or a snippet of text within a sentence.

In a literature review of automatic information extraction that supports systematic reviews, Jonnalagadda et al. (2015) listed 26 NLP models which support automatic information extraction from RCTs [61]. Note that not all of these data extraction tools are designed to support evidence quality assessment specifically, but rather the evidence synthesis process in general (which includes other tasks such as searching for relevant articles). Therefore, the tools in Jonnalagadda et al.'s review extract different information items that could be used for different purposes. The majority of them focus on extracting PICO-related information. However, only some, not all, PICO-related information is in regard to study methodology, thus can be used for quality assessment. Some examples of fine-grained methodological information belonging to the PICO framework that could be helpful for quality assessment are information about participants such as settings, sample size or information about interventions such as treatment duration, doses, etc. In a review of automation technologies to support systematic reviews in 2019, Marshall et al. specifically discussed the two most popular data extraction tools that support the evidence synthesis process, ExaCT and RobotReviewer. The authors also highlighted the fact that tools which facilitate the screening process are widely accessible and usable, while information extraction tools are still at the piloting stage or require higher amounts of human input [58]. Most recently, in 2020, Schmidt et al. started a living review protocol with the purpose of reviewing all available information extraction methods for systematic review semi-automation [62]. The team published the first results of the review in May 2021, in which they reviewed 53 information extraction tools/models that support systematic review. In addition, according to Systematic Review Toolbox, which is a web-based catalogue of tools that support various tasks within the systematic review and wider evidence synthesis process (last updated in 2021), there are a total 51 software/tools that support the information extraction task [88]. Among those, there are 7 tools that indicate "automatic" or "automation" in their descriptions.

Table 1 shows a summary review of 56 automatic information extraction tools from Jonnalagadda et al. in 2015, Marshall et al. in 2019, and Schmidt et al. in 2021 reviews, and additional relevant studies from the literature. The table contains the following information:

- Study Reference: reference to the publication.
- Year: year of the publication, sorted in ascending order.
- Level of information:
  - Sentence level*: the system classifies each sentence into different categories of information item.
  - Entity level: the system extracts chunks of text (could be words or phrases) that contains a specific information item.

  *Note that some studies tried to extract information at sentence level. Thus, the task is "sentence classification", in which models are built to classify sentences into different categories of information. However, conceptually, such task is still considered as a information extraction.
- Information Items: what are the information extracted from the work of the publication.
- Study Design: what kind of study designs used as data from the work of the publication.
- Full text (Y or N): binary field indicates if the work used full text of publications as data (N means it only used either Title or Abstract, or both).
- NLP Methods: description of NLP techniques, ML models used.

Table 2.1: Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	**Year**	**Level of information**	**Information Items**	**Study Design**	**Full text (Y or N)**	**NLP Methods**
Demner -Fushman et al. [89]	2005	Sentence Entity	PICO Problem	RCT, Cohort, Case series, Case control, Diagnostic test, Other	N	Rule-based Naive Bayes

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Demner -Fushman et al. [90]	2006	Sentence	Outcomes	RCT	N	Rule-based Naive Bayes
Fiszman et al. [91]	2007	Entity	Intervention, Comparison	RCT, Comparative Studies	N	Rule-based
Chung et al. [92]	2007	Sentence	Aim, Methods, Participants, Results, Conclusion	RCT	N	Conditional Random Field (CRF) Support Vector Machine (SVM)
Hara et al. [93]	2007	Sentence Entity	Disease, Treatment, Patients	RCT	N	Rule-based, CRF, SVM
Xu et al. [94]	2007	Sentence Entity	Participants, Demorgraphic, Number of participants	RCT	N	Rule-based, Naive Bayes, Hidden markov model (HMM)
Dawes et al. [95]	2007	Entity	Patient; Exposure; Comparisons, Outcome; Duration of follow up; Results		N	Rule-based

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
De Bruijn et al. [96]	2008	Sentence Entity	Eligibility criteria, Intervention parameters (dosage, frequency, duration), Sample size, Start and end date of enrollment, Primary and secondary outcomes, Relevant time points, Funding information, Publication details (date, authors).	RCT	N	Rule-based SVM
Hansen et al. [97]	2008	Entity	Participants	RCT	N	Rule-based, SVM
Chung et al. [98]	2009	Sentence	Intervention, Participants, Outcome Measures	RCT	N	CRF, SVM
Chung et al. [99]	2009	Entity	Intervention arm	RCT	N	Rule-based, CRF
Summerscales et al. [100]	2009	Entity	Treatements Groups Outcomes	RCT	N	CRF

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Kiritchenko et al. [101]	2010	Sentence Entity	Eligibility criteria, Intervention parameters (dosage, frequency, duration), Sample size, Start and end date of enrollment, Primary and secondary outcomes, Relevant time points, Funding information, Publication details (date, authors).	RCT	Y	SVM, CRF, Rule-based
Boudin et al. [102]	2010	Sentence	PICO	RCT	N	SVM, Naive Bayes Random Forest
Xu et al. [103]	2010	Entity	Exposure	Cohort, Case series, Case control, Other	N	Rule-based
Boudin et al. [104]	2010	Sentence Entity	PICO	RCT	N	Rule-based, SVM, Naive Bayes, Random Forest

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Lin et al. [105]	2010	Entity	Intervention, Number of Participants, Age, Setting, Design, Enrollment dates, Funding Organization	RCT, Cohort	Y	Regex, CRF
Kim et al. [106]	2011	Sentence	PICO Study Design	RCT	N	CRF
Summerscales et al. [107]	2011	Sentence Entity	Groups Outcomes Group Size	RCT	N	Regex, CRF, Naive Bayes; Decision Tree; Regression
Huang et al. [108]	2011	Sentence	PICO	RCT	N	Naive Bayes; Decision Tree; Regression
Verbeke et al. [109]	2012	Entity	PICO Study Design	RCT	N	SVM, HMM
Zhao et al. [110]	2012	Sentence Entity	Participants (Sex, Age, Race, Condition) Results, Intervention, Study Design, Research Goal.	RCT	Y	CRF
Amini et al. [111]	2012	Sentence	PICO, Design	RCT	N	CRF, SVM, Naive Bayes; Decision Tree; Regression.
Zhu et al. [112]	2012	Entity	Participants (Age, Gender, Race)	RCT	N	Regex

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Hsu et al. [113]	2012	Sentence Entity	Hypothesis, Statistical Methods, Outcomes and estimation, Generalizability.			SVM, CRF
Huang et al. [114]	2013	Sentence	PICO	RCT	Y	Naive Bayes; Decision Tree; Regression
Kelly et al. [115]	2013	Entity	Age of Subjects, Duration of Study, Ethnicity of Subjects, Gender of Subjects, Health Status of Subjects, Number of Subjects.	RCT	N	Rule-base regex
Hassanzadeh et al. [116]	2014	Sentence	PICO Study Design,	RCT	N	CRF
Karystianis et al. [117]	2014	Entity	Participant, Outcomes, Design, Exposure, Other	Cohort, Case series, Case control, Other	N	Regex,
Chabou et al. [118]	2015	Sentence	PICO	RCT	N	Regex, CRF
Blake et al. [119]	2015	Entity	Outcomes	RCT, Animal studies	Y	Regex, SVM, Naive Bayes, Decision Tree, Regression.
Suwarningsih et al. [120]	2015	Sentence	PICO	RCT	N	Regex
Wallace et al. [121]	2016	Sentence	PICO	RCT	N	Regex, CRF, SVM

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Raja et al. [122]	2016	Sentence Entity	Participants (Age, Gender, Condition or disease) Intervention Outcomes	RCT, Other	N	Regex, Naive Bayes; Decision Tree; Regression.
Basu et al. [123]	2016	Sentence	Eligibility Criteria	RCT	Y	SVM
Bui et al. [124]	2016	Sentence, Entity	PICO Number of participants	RCT	Y	Regex, SVM
Raja et al. [122]	2016	Sentence, Entity	Participant (Condition or disease), Intervention, Comparison	RCT	Y	Regex, APIs metadata retrieval
Singh et al. [125]	2017	Entity	PICO	RCT	Y	CNN
Marshall et al. [126]	2017	Sentence, Entity	Participant (Condition or disease), IC, O, Age, Gender, Randomisation, Blinding, Design, Eligibility criteria, Race, Other	RCT	Y	CNN, SVM, distance supervision
Karystianis et al. [127]	2017	Entity	Participants, Country, Exposure, Outcomes	Cohort, Cross sectional survey, Case control	N	Rule-base regex
Lucic et al. [128]	2017	Entity	Outcomes	RCT, Animal studies	Y	Rule-base regex, SVM
Jin et al. [129]	2018	Sentence	Background, Participants, Interventions, Outcomes, Study Design, Others.	RCT	N	CNN, RNN, BERT-based Transformer Models

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Jin et al. [130]	2018	Sentence	Aims, Participants, Interventions, Outcomes, Methods (Design and Settings), Results, Conclusion.	RCT	N	CRF, Bidirectional Long short-term memory (Bi-LSTM)
Nye et al. [131]	2018	Entity	PICO	RCT	N	CRF, Bi-LSTM & CRF
Demner -Fushman et al. [132]	2018	Entity	IC (Drug dose; strength; route; frequency; duration), Other	RCT, Animal studies	Y	CRF, LSTM
Chabou et al. [133]	2018	Sentence	PICO	RCT	N	Regex, CRF
Baladron et al. [134]	2018	Entity	Number of participants	RCT	N	Regex, Naive Bayes, Decision Tree, Regression
Brockmeier et al. [135]	2019	Entity	PICO	RCT	N	CRF, Bi-LSTM & CRF
Kang et al. [136]	2019	Sentence	PICO	RCT	Y	BiLSTM & CRF
Xia et al. [137]	2019	Sentence	PICO	RCT	N	SVM
Guo et al. [138]	2019	Entity	Drug	RCT	N	SVM
Norman et al. [139]	2019	Sentence Entity	Diagnostic tests: index and reference standard, Participants (Condition or disease)	Diagnostic test	Y	BERT models, PDF extraction
Brassey et al. [140]	2019	Entity	Participants, Intervention, Number of Participants	RCT	N	Rule-base regex

Table 2.1 (cont.): Summary of 56 automatic information extraction tools that support clinical research quality assessment

Study Reference	Year	Level of information	Information Items	Study Design	Full text (Y or N)	NLP Methods
Yuan et al. [137]	2019	Sentence	Aims, Participants, Interventions, Outcomes, Methods, Results, Conclusion.	RCT	N	SVM
Marshall et al. [141]	2020	Entity	PICO, Number of participants, Design	RCT	N	CRF, LSTM, CNN, SVM
Schmidt et al. [62]	2020	Sentence, Entity	PICO	RCT	N	BERT models, PDF extraction

According to the summary table, 53 out of 56 studies extracted information from RCTs. Majority of the studies took text from abstracts, only 14 studies extracted information from the full texts. 22 studies extracted information at entity level only, 19 studies extracted information at sentence level, 15 studies extracted information at both levels. 50 out of 56 studies extracted at least one kind of information in the PICO framework (either Participants, Intervention, Comparison, Outcomes, or all), in which, fine-grained information regarding characteristics of participants (such as: age, gender, race, ethnicity, etc.) is extracted by the highest number of the studies. Only small number of studies extracted information that can be used for quality assessment. In particular, only 5 studies extracted "Study Design" information [106], [109], [110], [116], [130]; of which three work on the same dataset created by Kim et al. in 2011 known as PIBOSO-NICTA dataset [106]. Five other studies also extracted more fine-grained methodological information, such as "Duration of the Treatments" [95], [96], [101], [115]; 2 studies extracted "Sample Size" [96], [101]; 1 study extracted statistical analysis methods [113]. The two most comprehensive and well-known data extraction systems are ExaCT [101] and RobotReviewer [121]. We are going to discuss these two tools in details later in this section. Two newer studies which work on a similar problem, extracting PICO information items, however, at a more granular level, were published by Nye et al. in 2018 [131] and Kang et al. in 2019 [136]. In both research, a information extraction model was developed on a new corpus of RCT papers in which PICO information elements were manually annotated at mention level.

In term of machine learning algorithms, Support Vector Machine (SVM) and Conditional Random Field

(CRF) are the most popular ML algorithms that were used for data extraction, which are used by 20 and 21 studies respectively. Interestingly, a large number of studies, 29, used rule-based methods (often with regular expression) either alone or in combination with ML methods to identify and extract information. Commonly used features along with these ML algorithms include lexical features (such as n-grams), syntactic features (such as speech of tags) and positional features (such as position of sentences in a paragraph or position of words in a sentence). Some studies applied deep learning method (such as bidirectional long short-term memory– BiLSTM) in combination with CRF which takes word embeddings [130]–[132], [135], [141], [142] or contextual embeddings (such as BERT) as inputs [62], [130], [139]. In terms of performances, all of the studies used popular performance metrics including precision, recall and F1 scores to report the tools/models' performances. Since the way the studies set up these problems (sentence classification vs. entity extraction) are different, it is hard to compare the models' performances from one to another. Overall, performances reported in these studies were all in the 70s or 80s range.

In the next section I will review in depth two data extraction tools: ExaCT and RobotReviewer. The reasons are two-fold: (1) ExaCT, perhaps, is the tool that extracts the most comprehensive list of both PICO-related and methodology-related information (21 items in total), in which some of them, such as study design, sample size, can be used for evidence quality assessment; (2) RobotReviewer is the only tool that not only extracts PICO-related information, but also automatically detects (and highlights) sentences.

ExaCT is designed to extract key trial characteristics (e.g. eligibility criteria, sample size, drug dosage, primary and secondary outcomes, early stopping, etc.) from full-text journal articles reporting on RCTs [101]. More specifically, ExaCT consists of two parts: an information extraction (IE) engine that searches the article for text fragments that best describe the trial characteristics, and a web browser-based user interface that allows human reviewers to assess and modify the suggested selections. The IE engine uses a statistical text classifier to locate the sentences with the highest probability of describing a trial characteristic. Then, the IE engine's second stage applies "weak" rules to these sentences to extract text fragments containing the target answer (for example, start date of enrollment is extracted as the first string that looks like "a date", sample size is an integer number with a reference to people, funding organization is a sequence of words with the first letter capitalized). The same approach is used for all 21 trial characteristics selected for ExaCT. In their evaluation, the ExaCT team found that, averaged over 21 information elements, both precision and recall were 80%. While ExaCT is able to capture a wide range of information from an RCT study, it is not designed specifically to support the quality assessment purpose. Some extracted information such as funding number, funding organization, DOI, authors, are not directly relevant to study quality. In contrast, the tool does not capture important randomization-related information that reviewers need when assessing evidence quality such as randomization generation or concealment allocation.

RobotReviewer provides two features: PICO extraction [121] and risk of bias prediction [59]. Regarding the data extraction feature, RobotReviewer takes RCT reports in PDF format, automatically retrieves

sentences which describe PICO-related information and represents them into three groups of information: population, intervention and outcomes. Besides PICO, RobotReviewer also extracts sentences that describe the conduct of randomization from RCTs and groups them into three categories: random sequence generation, the allocation concealment, and blinding. This information is relevant to biases and eventually is used to assess risk of bias of RCT studies according to the Cochrane Handbook. According to internal evaluations conducted by the RobotReviewer team, in terms of PICO extraction, the tool achieved high precision (≥ 0.88) for all PICO items (the team did not report recall).

Both ExaCT and RobotReviewer show some advantages over other NLP tools/models:

- ExaCT extracts a wide range of information that could be used for multiple purposes throughout the evidence synthesis process, while RobotReviewer is the only tool that explicitly extracts randomization-related information that could be used for quality assessment.
- Different from other systems, which were trained mainly on abstracts of RCTs, both ExaCT and RobotReviewer classification models were trained on text from the full-text articles. Compared with abstracts, full-texts of articles are richer in terms of contents which would give the models better/more representative training examples.
- Compared with other tools/methods (which are mostly one-off methods), both ExaCT and RobotReviewer were more mature and used in practice. Both tools provide user-friendly interfaces and try to incorporate humans into the process. In the ExaCT case, the tool provides a user-friendly interface that shows candidate sentences and allows users to decide if each of the sentences is relevant or not (through a checkbox). In the RobotReviewer case, the tool also provides a separate interface that shows the full-text articles and actively highlights sentences relevant for predictions.

Automatic risk of bias prediction

Besides data extraction, another NLP approach to assist evidence quality assessment is to develop models that directly predict the level of biases based on text from the articles. More specifically, such NLP systems will take input which are full-text articles, abstracts or a paragraph from the full-text, and use them as textual resources for features to train machine learning models. The models then predict outputs which are quality levels, such as low quality or high quality. This approach can be done as a text classification task with or without data extraction. According to the Systematic Review Toolbox, there are 32 software/tools that support the quality assessment (risk of bias assessment) task. Only four of them mentioned "automatic" in the tool descriptions. And only two tools (RobotReviewer and TrialsStreamer) support quality prediction.

RobotReviewer is the only NLP tool that starts to tap into both aspects of data extraction and risk of bias prediction. The tool not only attempts to identify information at the sentence level that can be used to assess risk of bias of a clinical study, but also tries to predict the levels of risk of bias at document level [59],

[141]. Following the guideline for risk of bias assessment provided by Cochrane, RobotReviewer considers six domains and uses them for risk of bias level prediction. Six separate machine learning models were developed separately for prediction. The models were trained on a corpus of 2200 clinical trial papers which was originally derived from the Cochrane Database of Systematic Reviews using a distant supervision method. The papers were labeled as being at low, high or unclear risk of bias for each domain, and sentences were labeled as being informative or not. The risk of bias prediction features of RobotReviewer were evaluated multiple times, both by the tool creators themselves and also by others [143]–[145]. In the internal evaluation done by the tool creators, F1 scores range from 0.57 - 0.75 for the six information items. In the three external evaluations, RobotReviewer was used to predict risk of bias levels of RCT studies and compared with human judgments. The first evaluation study showed that the mean level of agreement between RobotReviewer and human researcher assessment was 72% [143]. The second evaluation found that RobotReviewer's reliability was moderate, ranging from 0.34-0.48 for different information items [144]. The third evaluation study showed that RobotReviewer yielded a moderate degree of agreement with human reviewers: Cohen's kappa for randomization was 0.52, for allocation concealment was 0.60, and for blinding of personnel/patients was 0.43 [145]. Notably, blinding of outcome assessors had only slight agreement (0.04). In a study comparing RobotReviewer performances with human, Jardim et al. found that RobotReviewer had equal performance to humans, though participating reviewers were not interested in modifying standard procedures to include automation [146]. Both data extraction and risk of bias prediction features from RobotReviewer were adopted and used to empower a new product named TrialStreamer, a automatically updated database of clinical trial reports, from the same research group [141]. TrialStreamer integrates different features to automatically annotate RCT studies including extraction of PICO elements (from RR), prediction of risk of bias levels (also from RR) and extraction of sample size. All this information is then used to represent the RCTs in a structured representation and curated in a database for subsequent use [1].

Similar to RobotReviewer, Millard et al. in 2016 developed a machine learning model to assist risk-of-bias assessments for systematic reviews [60]. Millard's model took the same approach with RobotReviewer by implementing two models: (1) a sentence model to identify relevant sentences. In particular, sentences from full-text articles are input to a classification model and predicted with labels "relevant" (or "not relevant") if they belong (or not belong) to either randomization, allocation concealment and blinding group of information; (2) a article-level model to predict the risk-of-bias value of each article. The scores output by the model are used to rank articles by predicted risk of bias. The two models were trained and evaluated separately. According to the results, sentences can be successfully ranked by relevance with area under the receiver operating characteristic (ROC) curve (AUC) ≥ 0.98. This is useful to assist reviewers by indicating which parts of the article text are particularly relevant to risk of bias. They were also able to rank articles according to risk of bias with AUC ≥ 0.72, which would help reviewers to prioritize articles to look at first (e.g from low to high risk of bias).

[1] https://trialstreamer.ieai.robotreviewer.net/

2.3 Problems with the existing NLP-based approaches and thesis solutions

2.3.1 Problems

Based on the review of existing literature, there are several shortcomings of NLP approaches in addressing the automation of evidence quality assessment:

- Most information extraction tools focus on a limited set of information– mostly PICO, which is more important for identifying relevant articles than assessing their quality. RobotReviewer seems to be the only tool that extracts specific information items (randomization generation, allocation concealment and blinding) that are used for evidence quality assessment of RCTs.

- The methodological information extracted from the existing tools are mostly at sentence or document level. More specifically, both RobotReviewer and Millard's model identify randomization, allocation concealment and blinding information at sentence level. For example, given the sentence *"Participants and the research assistant were blinded to group allocation at baseline; and, this could be maintained at 6 and 13 weeks."*, these models are able to predict that the sentence discusses blinding; however, they are unable to capture who was blinded (participants and the research assistant) and when they were blinded (at baseline). This is because these models do not capture fine-grained information such as who was blinded (e.g. *Participants* and *the research assistant*) and how they were blinded (e.g. *at baseline* and *be maintained at 6 and 13 weeks*). The lack of fine-grained information makes it difficult and time-consuming for researchers/reviewers from identifying methodological flaws (such as measuring outcomes on individual patient in a cluster trial), weaknesses (such as under-powered sample size) or inconsistencies (such as the mismatch between randomization block size and number of patients) while assessing methodology quality of clinical studies.

- Prediction tools make evidence quality judgment directly without explanation of the criteria that lead the systems to make the judgment. For example, given a RCT article, RobotReviewer and Millard's model extract textual features from the article, and based on these features, the models give predictions of high/low or undecided risk of bias directly to users. In the case of Millard's model, this may result in users missing the insights of how the predictions are actually driven and what are the reasons for such assessment decisions. In the case of RobotReviewer, although supporting sentences are highlighted, once again, the lack of fine-grained information that would be helpful to detect methodological flaws could be missed. In fact, in an evaluation study in which we assessed the use of RobotReviewer risk of bias assessment feature by asking an expert to use the tool in his on-going systematic review project, we found that even though RobotReviewer successfully captured sentences that contain "blinding" information in general pretty well (such as sentences that mentioned "double-blind"), the tool only

successfully captured sentences that contain "blinding object" information 4 out of 10 times. Most of the work on data extraction from published evidence has been for RCT reports; extracting data from other types of research (e.g. non-randomized treatment studies or cross-sectional observational studies) are also important [147].

- Lack of adequate data: Data extraction or prediction tools are based on supervised models that rely on manually labelled data in order to "learn" to do a task. Yet, in the biomedical domain, manually labeling data is expensive and time-consuming. Most of existing data sets are specially designed for PICO information. None of them comprehensively contains methodological characteristics for quality assessment. On the other hand, majority of the existing models used text from titles ad abstracts only, which is limited and misses out important methodological details which only reported in full-text. In addition, the existing tools such as ExaCT or RR both faced the problem of sparsity of training data to train their models. ExaCT, for example, was trained on a small set (132 total) of full-text articles. RobotReviewer was trained using a much larger dataset, but the "labels" were induced semi-automatically, using a distant supervision strategy which is known for noisy data. Nye et al. in 2018 released a new corpus of 5,000 richly annotated abstracts of medical articles describing clinical randomized controlled trials in which fine-grained information of PICO elements were annotated mainly by non-expert workers. Even though such a corpus could be helpful to resolve data scarcity problems, the quality of the annotation is still questionable (e.g. the average Cohen's kappa agreement F1 score between expert annotators –medical students– was 67%, and between the non-expert annotators was 53%).

2.3.2 Thesis solutions

I hypothesize that NLP tools could assist researchers in evidence quality assessment by automatically identifying fine-grained, explicitly stated, methodological characteristics of a clinical study and representing them in a structured representation that computers could reason with. Existing NLP computer support for evidence quality assessment such as RobotReviewer achieves comparable results with human assessments and it has been shown that it helps to reduce the time required for conducting evidence synthesis. Nevertheless, given information (highlighted sentences) provided by RobotReviewer, researchers would not be able to identify potential methodological flaws, weaknesses and inconsistencies between methods and results of a clinical study, which requires further analysis of fine-grained information.

My overall research goal is investigate and develop NLP methods to assist researchers to assess evidence quality from clinical studies. Towards this goal, I investigate NLP methods to extract information from clinical publications that can be used for evidence quality assessment at different levels of granularity. In particular, I not only extract quality information at document level (as we called it "coarse-grained" granularity) and sentence level (as we called it "medium-grained" granularity), but also take a further step

comparing with current approaches by looking at quality information at term level (as we called "fine-grained" granularity). To my best knowledge, currently there are no automated approaches using NLP and ML to capture methodological information at term granular level to support quality assessment. My thesis will present three research toward that objective as shown in Figure 2.3.

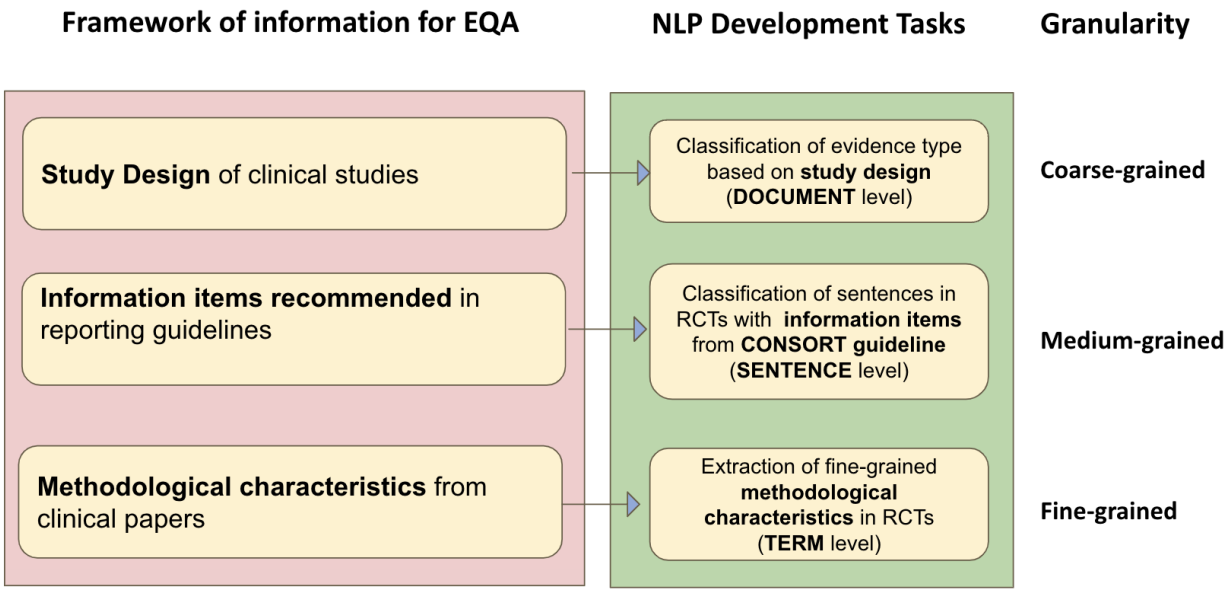

Figure 2.3: Hierarchy of study designs

- **Research 1: Automatic extraction of study design to support methodological quality assessment**
  - **What is the task**: A development of classification models to automatically classify clinical papers into different evidence types based on study designs. This development was built based on the consideration that study design is the most coarse-grained criteria to assess evidence quality.
  - **Level of information**: Document level
  - **Input/Output**: The model takes clinical papers as inputs, and predicts study designs of the papers as outputs.
  - **How does it help**: Identifying study design of an article is particularly helpful because it helps to specify which methodological information should be considered next in the pipeline of evidence quality assessment.
- **Research 2: Automatic classification of CONSORT checklist items to support reporting quality assessment**
  - **What is the task**: A development of classification models to map sentences from RCT papers to CONSORT checklist items as a step toward computer supported reporting transparency assessment.

- **Level of information**: Sentence level
- **Input/Output**: The model takes sentences from RCTs as inputs, and predicts what information the sentences describe, using CONSORT items as predicting labels (e.g. sentences about sample size determination or blinding, or trial design).
- **How does it help**: This model takes one step further by considering more fine-grained information that is used to assess reporting transparency of RCTs. Results of such a model would be helpful for further methodology quality assessment because if the paper is not transparent, we have little to judge whether the study is rigorously done.

- **Research 3: Automatic extraction of methodological characteristics to support methodological quality assessment**
  - **What is the task**: A development of information extraction systems to extract fine-grained methodological information items from RCT papers and represent them in a structured representation.
  - **Level of information**: Term level
  - **Input/Output**: The model takes full text of RCTs as inputs, and extract methodological characteristics in the format of text mentions that can be used for quality assessment and map them into a hierarchical data model.
  - **How does it help**: This model takes one step further by considering lowest level of information granularity, at mention level. Extracted information can be represented in a structure that can be reasoned and queried for quality assessment purposes later on.

The three developments are connected in such a way that they address different levels of information granularity used for evidence assessment: starting from the most coarse-grained information, study design, which can be identified at the document (individual paper) level; then going to more fine-grained yet still at the sentence level within papers; and finally the most fine-grained entity-level information which can be identified through the text mentioned within sentences. Multiple users, who work in different fields but have the same focus on research quality, can all benefit from these developments. Those include: systematic reviewers who do evidence synthesis, journal editors and peer reviewers who review papers, or authors themselves who might need to authorize their manuscripts before publishing. Extracting information in different levels of information granularity is particularly helpful because at each stage of the evidence quality assessment process, different stakeholders might have different needs. For example, systematic reviewers at the screening step might be only interested in a certain type of clinical study (e.g. RCT), thus can use "study design" as the information for better retrieving relevant publications. Or, journal editors at reviewing step, might be interested in assessing how much a submitted manuscript in compliance with reporting guidelines,

thus information reported at sentence level could be used. Finally, extracting fine-grained methodological information allows reviewers to identify any methodological weaknesses, inconsistencies or flaws that are not explicit and easy to spot from the lengthy full text articles. More than that, the extraction of fine-grain methodological characteristics from clinical study could also be beneficial beyond the quality assessment purpose by enabling semantic searching of the literature based on methodological characteristics of the articles.

Chapter 3

Natural language processing preliminaries

Before going into details of the research presented in this thesis, in this chapter, I will review some basic natural language processing (NLP) and machine learning (ML) concepts that I used in my research. Readers who are already familiar with these concepts may decide to skip this chapter. More specifically, in this chapter, I will:

- First, discuss the broad NLP paradigms that I used: (1) rule-based approaches; (2) traditional supervised ML approaches; (3) representation learning-based approaches.
- Second, review the NLP methods and techniques that are applied for two specific tasks: ***Text Classification*** and ***Information Extraction***, which are the two categories of NLP-based automatic computer support for EQA as discussed in chapter 2, and also the focus of my research. For each task, I will discuss: (1) Task description; (2) NLP/ML models ; and (3) Evaluation metrics to assess performances of NLP/ML models.

3.1 General NLP paradigm

NLP includes many different techniques for interpreting human language, ranging from statistical and machine learning methods to rules-based and algorithmic approaches. While supervised and unsupervised learning, and specifically deep learning, are now widely used for modeling human language, there is also a need for syntactic and semantic understanding and domain expertise that are not necessarily present in these machine learning approaches. Here, we are going to discuss further three approaches that are used in the field for different downstream applications in general, and have been applied for this thesis in particular. Those are:

(1) rule-based; (2) supervised machine learning approach; and (3) representation learning approach.

3.1.1 Rule-based NLP approach

A rule-based NLP system is commonly comprised of a set of rules defined by human which tend to focus on pattern-matching or parsing depending on the NLP task that it is designed for, and can be tailored for domain-specific use cases. For text classification, rule-based approaches classify text into organized groups by using a set of handcrafted linguistic rules based on characteristics of the text (e.g., lexical, syntactic, semantic). These rules instruct the system to use semantically relevant elements of a text to identify relevant categories based on its content. In general, each rule consists of an antecedent or pattern and a predicted category. For information extraction, a dictionary of terms/phrases or several rules are created based on the existing knowledge-base or vocabulary for an information type. In the subsequent step, these dictionary terms are tagged in the text using a string exact match or a variation term that follows the defined rule. Figure 3.1 shows the input/output flow of a typical rule-based NLP model. Rule-based approaches, though are the oldest approaches to NLP, have been proven to work well in many downstream NLP tasks and in various domains (especially the domain that has unique language characteristics, a sub-language [148], and a large number of knowledge sources). The advantages of rule-based approaches are: they can be flexibly and incrementally developed, they do not require massive training corpus, and they tend to have high performance in specific use cases (high precision). However, rule-based models often suffer performance degradation when generalized (low recall), require skilled developers and linguists.

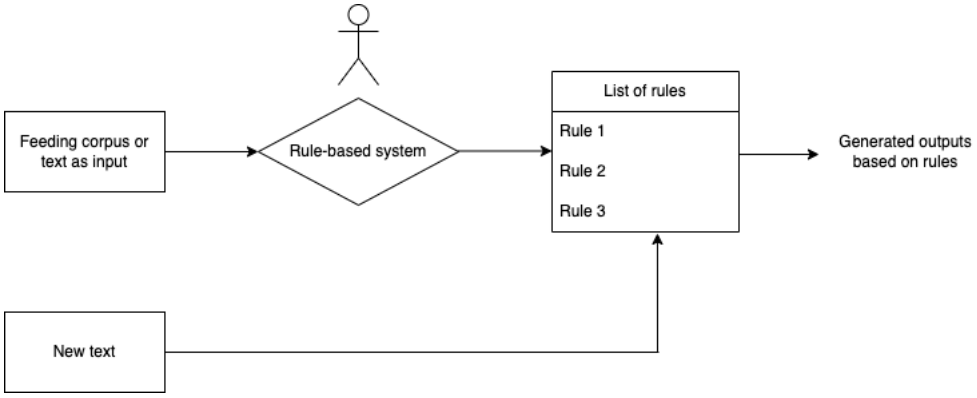

Figure 3.1: Rule based system input/output flow

3.1.2 Traditional supervised machine learning

A machine learning-based NLP system relies on more modern "statistical inference" techniques. There are two main types of ML-based approaches: supervised machine learning and unsupervised machine learning. Supervised learning is trained using data that is labeled (or tagged), and during training, those models learn

the best mapping function between a known data input and expected known output. Unsupervised learning is a learning approach using unlabeled data which means no labeled data is required for training. In this section, we will discuss further supervised machine learning approach, which is the focus of the research in this thesis.

As mentioned, this machine learning approach is called "supervised" because its way of learning from training data mimics the same process of a "teacher" supervising the end-to-end learning process. In supervised machine learning, a batch of text documents are tagged or annotated with examples of what the machine should look for and how it should interpret that aspect. These documents are used to "train" a statistical model, which is then given untagged text to analyze. For text classification, by using pre-labeled examples as training data, supervised machine learning algorithms can learn the different associations between pieces of text, and that a particular output (as labels) is expected for a particular input (as text). For information extraction, the ML-based systems use statistical-based models for detecting the named entities. These models try to make a feature-based representation of the observed data. Figure 3.2 shows the general design of a supervised machine learning system, which typically contains three main components:

- Feature extractor: is the component to transform each text into a numerical representation in the form of a vector. This process is also referred as "feature engineering" process. Note that this component is required for traditional ML algorithms, but not for representation learning approaches (in which we will discuss in the next section) since it can be automatically learnt during the training process.

- Machine learning algorithm: is the component that learns a model from the features that are defined. Once the model is trained with enough training samples, the trained model can be applied to make predictions on unseen data.

- Trained ML model: The same feature extractor is used to transform unseen text to feature sets, which can be fed into the model to get predictions on labels.

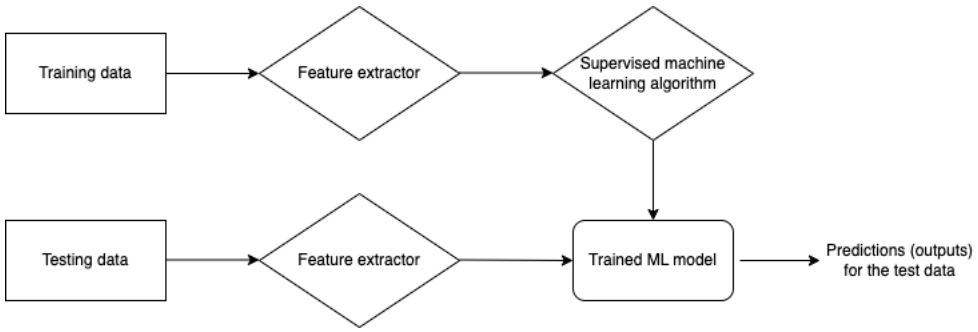

Figure 3.2: General design of a supervised machine learning system

Depending on downstream tasks, different ML algorithms can be used. Examples of some most popular traditional supervised NLP machine learning algorithms are: Naive Bayes, Logistic Regression, and Support

Vector Machines; Maximum Entropy and Conditional Random Field (for named entity recognition/information extraction).

3.1.3 Representation learning approach

Representation learning is a class of machine learning approaches that allows a system to discover the representations required for feature detection or classification from raw data. The concept of "representation learning" often appears in conjunction with deep learning which is a machine learning method based on neural network architectures with multiple layers of processing units [149]. In a deep learning model, the requirement for manual feature engineering is reduced by allowing a machine to be fed with raw data and to automatically discover latent representations and processing needed for text classification or information extraction. The approach thus avoids the feature engineering process, yet is able to learn complex and intricate features from data [150]. Figure 3.3 shows the comparison between typical (more traditional) supervised ML model (as shown in Figure 3.2) vs. deep learning model.

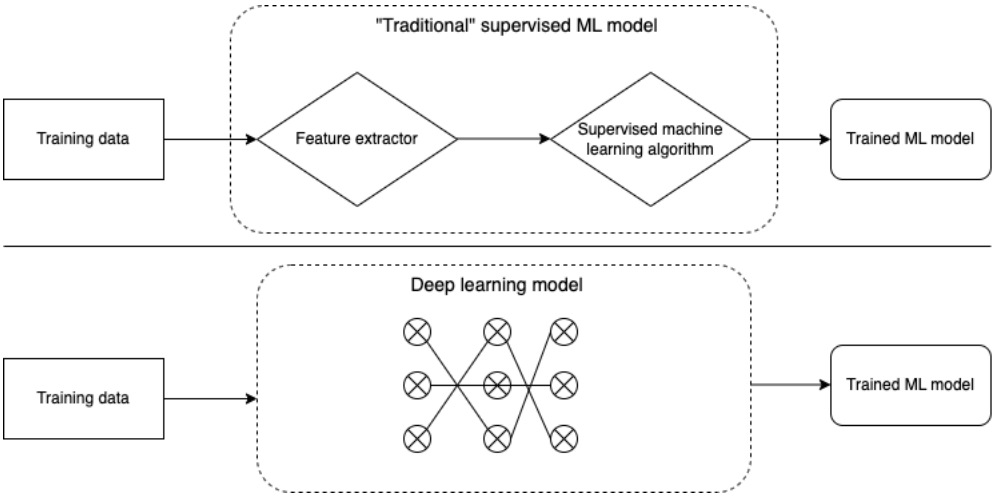

Figure 3.3: Comparison between traditional supervised ML model vs. deep learning model

Since deep learning models do not require human defined features as inputs, the most common way to represent an input instance (e.g. text) to a deep learning model is by its embedding vector. By definition, embedding vector is a low dimensional dense vector, where each dimension represents a latent feature and words that appear in similar context are expected to be represented with similar vectors [151]. Several approaches could be applied to produce word embedding:

(1) Static embeddings: In static embedding methods, a global vocabulary that contains all of the unique words from large corpora will be built. Then, similar representations are learnt for the words that appear in similar contexts. Two most popular architectures to learn the underlying word representations are Continuous Bag of Words (CBOW) and Skip-gram [152]. In the CBOW model, the distributed representations of context (or surrounding words by defining a fixed window size of words) are combined to predict the word in the

middle; while in the Skip-gram model, the distributed representation of the input word is used to predict the context [152]. Two popular pre-trained static embeddings models which were trained by CBOW and Skip-gram methods are Word2Vec [153] and GloVe [154].

Static embeddings like Word2Vec or GloVe can be used for transferring word embedding weights and using them to train deep learning models. The problem of such methods is that the words' contextual meaning is ignored. For example, only one representation is learnt for the "bank" word even though it can appear in two different contexts: *"I went to a bank to withdraw money"* and *"I walked along the bank of the river."*.

(2) Contextual embeddings: Contextual embedding methods are used to learn sequence-level semantics by considering the sequence of all words in the documents. Contextual embeddings assign each word/phrase a representation which is a vector based on its context, thereby capturing uses of words across varied contexts (e.g. in the example above, the contextual embeddings of the word "bank" will be different in each sentence depending on the context of the sentence in which the word appears). The most convenient way to obtain contextual embeddings is through pretrained language models which were trained on large corpora such as BERT, which is a linguistic representation model developed as an encoder of the Transformer model from Google AI [155] or ELMo, which is a deep contextualized word representation developed by AllenNLP [156]. Figure 3.4 shows the pipeline from raw text to embedding vector as input of deep learning model.

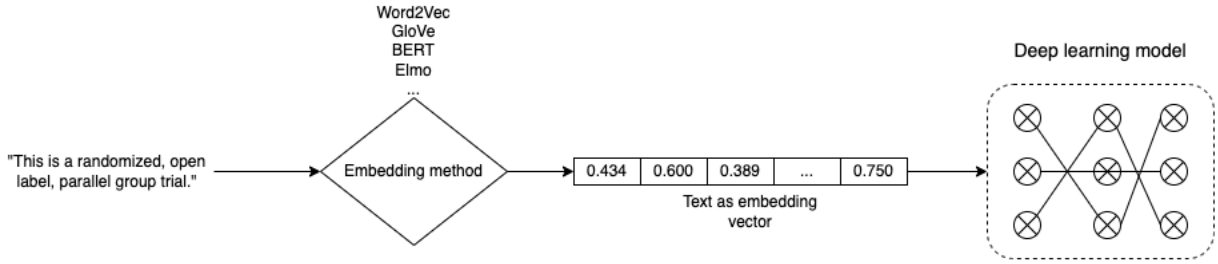

Figure 3.4: Pipeline from raw text to embedding vector as input of deep learning model

In the next section, we will discuss in details the BERT pretrained model and its variants since we used those as contextual embeddings sources for our research in this thesis.

BERT pretrained language model

Before introducing the BERT pretrained language model and discussing in details how it was developed, we will briefly introduce Transformer which is the architecture in which the BERT model was originally built upon.

Transformer is a type of neural network model that contains two main components: an encoder and a decoder as shown in the Figure 3.5 [155]. The input words of the model are represented using some form of embedding. The encoder maps an input sequence of symbol representations to a sequence of representations. Each encoder has two sub-layers: a multi-head self attention mechanism on the input vectors, and a simple, position-wise fully connected feed-forward network. Then the decoder generates an output sequence of

symbols one element at a time. Each decoder has three sub-layers: a masked multi-head self attention mechanism on the output vectors of the previous iteration, a multi-head attention mechanism on the output from encoder and masked multi-headed attention in decoder, and a simple, position-wise fully connected feed-forward network. The advantage of Transformer model over other deep learning models is the attention mechanism. Attention allows models to focus on parts of their input sequence while they predicted the output sequence. Multi-headed attention expands this concept by calculating attention in parallel multiple times. This allows the model to attend a word to multiple sub-structures within a given sentence at once [155].

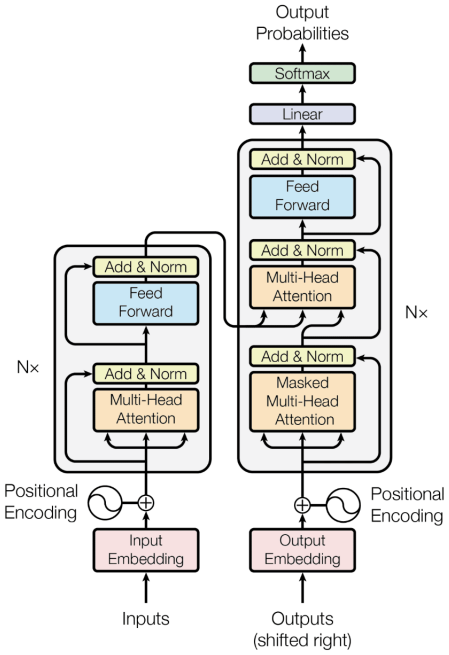

Figure 3.5: The Transformer model architecture: Input embeddings are passed to an attention layer which are then passed to a feed forward layer. The output of the encoder is passed to the decoder, which also includes an encoder-decoder attention layer [155]

BERT or Bidirectional Encoder Representations from Transformers, a pretrained language model, was developed as an encoder of the Transformer architecture and has been used to create state-of-the-art models for a wide range of NLP tasks [7]. We usually create a language model by training it on some unrelated task but tasks that help develop a contextual understanding of words in a model. The BERT model implementation improves standard Transformer by training the model through two different tasks: masked language modeling and next-sentence prediction. The Masked Language Model (MLM) randomly masks certain elements of the input, and the objective is to predict the original masked word based solely on its context. The objective of the MLM allows the representation to merge the left and right context, which allows us to pre-train a deep bidirectional Transformer. In addition to the masked language model, BERT uses a next sentence prediction task that jointly pre-trains the text pair representations. In this task, two sentences— A and B — are chosen for pre-training (50% of the time B is the actual next sentence that follows A; and 50% of the time B is

a random sentence from the corpus). The model was trained to predict if two sentences are next to each other given context surrounding them. With this design, outputs of the pretrained BERT is rich context aware representation of each token in the text that can be used for downstream tasks. There are different biomedical-focused variants of BERT model, such as BioBERT [157] or PubMedBERT [158], which used the same Transformer architecture with the original BERT. While BioBERT continues to pre-train the original BERT, the PubMedBERT is pretrained from scratch using abstracts and full-text articles from PubMed. By having a pre-trained model that integrates both general and biomedical domain corpora, developers and practitioners could now encapsulate biomedical terms that could specifically be used for biomedical NLP tasks. Given the success of BERT model and the domain specific expertise that BERT variant models provide, in this work, we used both BioBERT and PubMedBERT to develop our systems.

Thus far in this chapter, we have discussed the general NLP and ML concepts that can be applied for any downstream tasks. In the next section, I will discuss in details how these methods are used for two specific tasks, which are also the focus of our research: ***text classification***, and ***information extraction***.

3.2 Text classification

3.2.1 Task definition

Text classification is one of the fundamental tasks in NLP with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. The task is to assign a label from a set of predefined categories to unstructured text of arbitrary length (e.g., a document, a sentence). For example, sentiment analysis is one of the most common text classification tasks. In which, given a text (e.g. a movie review), a sentiment analysis classification system tries to predict the sentiment of the text to either "positive", "negative" or "neutral". Automatic text classification applies ML and NLP to automatically classify text in a faster, more cost-effective, and more accurate manner than human annotation.

3.2.2 Methods for text classification

The general architecture of ML classification models is similar to the general supervised learning model that we have discussed. It contains 3 main components: feature extractor, ML algorithm and the trained model. Some of the most popular machine learning algorithms that have been used for text classification task include the Naive Bayes family of algorithms, logistic regression, support vector machines (SVM)– which are known as "traditional ML" approaches; and neural network which is known as the "deep learning" approach. Depending on which approach is used (traditional or deep learning), the feature extraction methods can be different.

Traditional ML algorithms for text classification

Traditional ML algorithms learn from the data, where choice of algorithm and features (inputs) to be fed into algorithms are made by subject matter experts, who engineer features to be passed into the models. Some of the most common traditional algorithms used for text classification are: Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) [159]. Next, I discuss the SVM algorithm in details since we used it for some research in this thesis.

The SVM is a model used for both classification and regression problems though it is mostly used to solve classification problems. In a nutshell, SVM draws a line or "hyperplane" that divides a space into two sub-spaces. One subspace contains vectors (labels) that belong to a group, and another subspace contains vectors that do not belong to that group. The optimal hyperplane is the one with the largest distance between each label. The idea of SVM algorithm is shown in Figure 3.6. SVM is based on geometrical properties of the data while other algorithms such as logistic regression is based on statistical approaches. Therefore, SVM often works well with unstructured and semi-structured data like text and images.

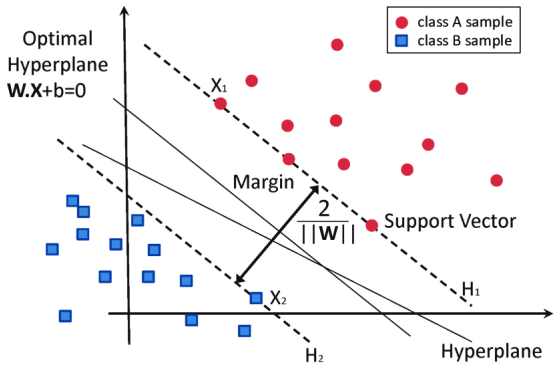

Figure 3.6: SVM algorithm [6]

Traditional ML algorithms require human defined the features that can be used to train models. The most basic (and popular) features for traditional ML algorithms such as SVMs are bag-of-words and bag-of-ngrams. The bag-of-words (BOW) is a feature extraction method that turns arbitrary text into fixed-length vectors by counting how many times each word appears. This process is often referred to as vectorization. For example, consider a corpus that contains the following words: *["This", "is","a", "randomized", "controlled", "trial"*], and we wanted to vectorize the text *"randomized controlled trial"* we would have the following vector representation of that text: (0, 0, 0, 1, 1, 1). Bag-of-ngrams method does the same idea but not on single word, rather on n-grams. An n-gram is a contiguous sequence of n items from a given sample of text or speech. N-gram of size 1 is referred to as a "unigram" which contains one word only; size 2 is a "bigram" which contains 2 words; size 3 is a "trigram" which contains 3 words. For example, in the above text, we would have the following bigrams: *"this is", "is a", "a randomized", "randomized controlled"*, and *"controlled trial"*. That way, Bag-of-ngrams features take into account the co-occurrences of words in the text instead of

considering the words independently like BOW approach.

Deep learning algorithms for text classification

Several deep learning models have been proposed in the past decade for text classification. Most popular models are Convolutional Neural Networks (CNNs) [160] and Recurrent Neural Networks (RNNs) [161]. A CNN model is a feed forward neural network using filtering and pooling, while a RNN model is a recurring network that feeds the results back to the network. CNNs are preferred in interpreting visual data, sparse data or data that does not come in sequence (e.g. images). RNNs, on the other hand, are designed to recognize sequential or temporal data (e.g. text). They make better predictions considering the order or sequence of the data as they relate to previous or the next data nodes. LSTM– Long Short Term Memory, which is a special kind of RNNs, capable of learning long-term dependencies, is another popular deep learning algorithm for text classification [162]. A LSTM layer consists of a set of recurrently connected blocks, known as memory blocks. Each one contains one or more recurrently connected memory cells and three multiplicative units named as: forget gate (to decide how much of the past you should remember), input gate (to decide how much of this unit is added to the current state), output gate (decides which part of the current cell makes it to the output) [162]. With this design, LSTM model address the limitation of a regular RNN model which is not able to keep track of long-term dependencies.

Nevertheless, the appearance of the Transformer model with and attention mechanisms [155], which can generate contextualized word vectors as discussed before, has been a significant turning point in the development of text classification and other NLP tasks. Many researchers have studied text classification models which used the Transformer model as a transfer learning resource, which achieves better performance than the above models. Transfer learning is the reuse of a pre-trained model on a new problem. Deep learning text classification models can use the pretrained embeddings from Transformer-based language model such as BERT as the inputs of a new output layer which can be a feed-forward neural network with sigmoid or softmax function (sigmoid is used for binary classification methods where we only have 2 classes, while softmax applies to multiclass problems). This process is called "fine-tuning", which essentially is a transfer learning technique. Fine-tuning is the process of using the weights of pre-trained model (such as BERT) as initialization for a new model trained on data from the same domain. Figure 3.7 shows the fine-tuning process from source model to target model. This process consists of four main steps:

- Use a pretrained neural network model as the source model (such as PubMedBERT).
- Create a new neural network model as the target model. This copies all model designs and their parameters on the source model except the output layer. We assume that these model parameters contain the knowledge learned from the source dataset and this knowledge will also be applicable to the target dataset.

- Add an output layer to the target model, whose number of outputs is the number of categories in the target dataset. For example, for text classification task, output layer could be a fully connected layer with softmax. Then randomly initialize the model parameters of this layer.
- Train the target model on the target dataset. The output layer will be trained from scratch, while the parameters of all the other layers are fine-tuned based on the parameters of the source model.

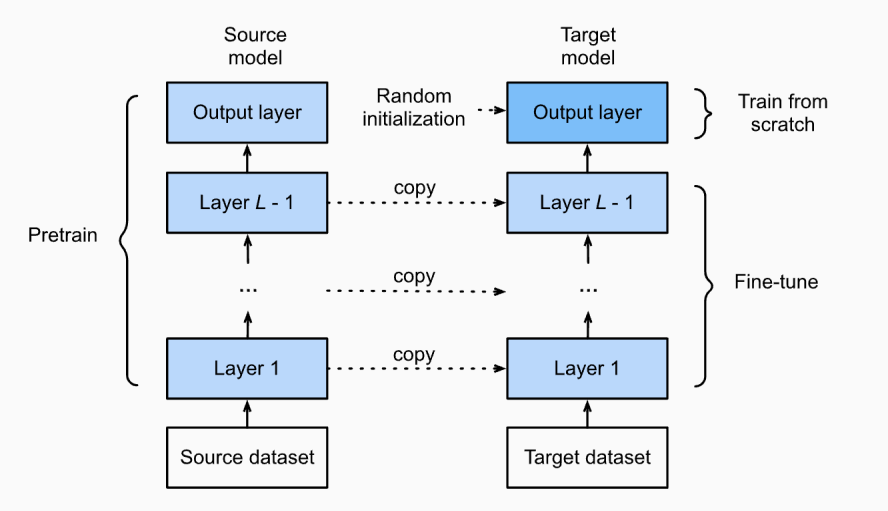

Figure 3.7: Fine-tuning from source model (e.g. PubMedBERT) to target model

3.2.3 Evaluation metrics to assess text classification models

To evaluate performances of classification models, common evaluation metrics have been used include: Precision, Recall, F-measure (F1). Essentially, Precision and Recall are calculated based on the number of true positives (TP), false positives (FP) and false negatives (FN), which can be described as:

- True Positive (TP): is the number of the correctly predicted labels.
- False Positive (FP): is the number of wrongly predicted labels.
- False Negative (FN): is the number of actual labels that are missed by the models.

F1 is the harmonic mean score of Precision and Recall. The formulae of these metrics are provided below:

$$Precision = \frac{TP}{TP+FP} \tag{3.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.2}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \tag{3.3}$$

The Area Under the Receiver Operating Characteristic curve (AU-ROC) is also another popular evaluation metric. The AU-ROC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AU-ROC, the better the performance of the model at distinguishing between the positive and negative classes. AU-ROC is calculated as the Area Under the *Sensitivity* (TPR)- *Specificity* (FPR) Curve:

$$Sensitivity = Recall = \frac{TP}{TP+FN} \tag{3.4}$$

$$Specificity = \frac{TN}{FP+TN} \tag{3.5}$$

3.3 Information extraction (Named entity recognition)

3.3.1 Task definition

Named entity recognition, NER (also known as entity identification, entity chunking and entity extraction) is a sub-task of information extraction that seeks to locate and classify named entities in text into predefined categories. The exact entity types that are of interest vary across different settings of the problem. When using supervised learning, NER is generally cast as either a token classification task or a sequence labeling task. In a sense, sequence labeling is also token classification, in which the goal is to classify each token (word) in a sequence of words (e.g. a sentence) with a categorical label [163]. However, different from the standard classification problem which assumes individual inputs (tokens) are independent, in sequence labeling tasks, labels of tokens are dependent on the labels of other tokens in the sequence, particularly their neighbors [163]. In this sense, the sequence labeling model takes into consideration the start and end of every relevant phrase according to the classification categories the model is trained for and tries to optimize the label sequence. For example, considering a disease entity extraction task the following sentence: *"The novel coronavirus disease (COVID-19) pandemic, caused by severe* ***acute respiratory syndrome*** *coronavirus 2 (SARS-CoV-2), remains a global challenge."*. To extract "disease" entity from this sentence, the sequence labeling model will not only predicts that the word "acute" is belong to "disease", but also considers this prediction into predicting that the next word "respiratory" and "syndrome" are belong to "disease" as well.

3.3.2 Methods for NER

Data encoding

In order to represent this as a sequence labeling problem, we need to convert the entity annotations into a sequence of labels. This labelling needs to handle the fact that multi-word phrases (such as "cluster randomized control trial") can refer to a single entity (study design). This process is called "data encoding". Different ways of encoding information in a set of labels make different chunk representations. The two most popular schemes are BIO and BILOU [163]. BIO stands for Beginning, Inside and Outside of a text segment. An NER system tries to predict the boundary of an entity which is encoded into the beginning (B) and inside (I) labels. Similar but more detailed than BIO, BILOU encodes the Beginning, the Inside and Last tokens of multi-token chunks while differentiating them from unit-length chunks.

For example, consider the sentence:

"The study design was a double-blind, parallel, randomized, controlled superiority trial.".

Then, a BIO scheme for study design entity could be represented as the following:

"The[O] study[O] design[O] was[O] a[O] double [B-blinding] blind[I-blinding] parallel[B-study design] randomized [I-study design] controlled[I-study design] superiority [B- Comparative Intent] trial[O]."

Deep learning algorithms for NER

A typical deep learning model for NER contains two components as shown in Figure 3.8: (1) a context encoder to obtain vector representations of the input which often are word level embeddings. Additional features (such as POS tags) can be added by obtaining their embeddings and concatenate with the word embeddings to get new representation of the words; and(2) a tag decoder to predict tags for each token in the input sequence.

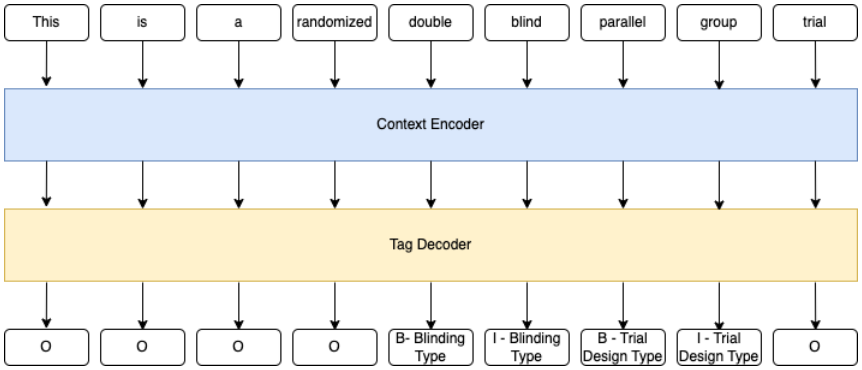

Figure 3.8: General Architecture of DL-based NER model

Context Encoder

For the NER task, there are two common encoder architectures:

(1) Bidirectional long short-term memory networks (BiLSTM): which is a part of the recurrent neural network (RNN) family that operates on sequential data [164]. Unlike standard LSTM that we discussed before, the input flows in both directions, and it is capable of utilizing information from both sides. BiLSTM adds one more LSTM layer, which reverses the direction of information flow. Therefore, it is a powerful tool for modeling the sequential dependencies between words and phrases in both directions of the sequence.

(2) Transformer-based pretrained language models: Transformer-based pretrained language models such as BERT and its variants, PubMedBERT and BioBERT have been used as contextual embeddings resources and have achieved good results for many NER tasks. One major advantage of the Transformer-based pretrained embedding models over others (e.g. versus BiLSTM), is that the original Transformers do not rely on past hidden states to capture dependencies with previous words. They instead process a sentence as a whole. That is why there is less risk to lose (or "forget") past information. Moreover, multi-head attention and positional embeddings both provide information about the relationship between different tokens [155].

Tag Decoder

There are two common tag decoders that can be used in a deep learning NER model: (1) Token Classification and (2) CRF as shown in Figure 3.9.

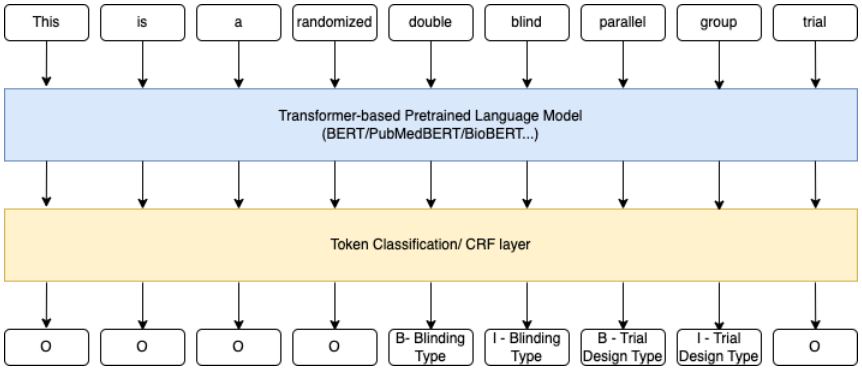

Figure 3.9: NER model with Token classification layer or CRF layer as decoder and BERT as encoder

(1) Token classification: This tag decoder is essentially a classification model. The decoder is a linear fully-connected classifier layer in the end of the whole NER model to get predictions for each of the tokens in the input independently.

(2) Conditional Random Fields (CRF): This tag decoder puts a CRF layer instead of a token classification layer in the end of the NER model to get token prediction.

Conditional Random Fields (CRF) is a class of discriminative models best suited to prediction tasks where contextual information or state of the neighbors affect the current prediction [165]. Essentially, CRF is a probabilistic model, in which, given a sentence (also called an "observation" or "sequence of observations"), the model tries to predict the best sequence of labels that corresponds to the input sequence of observations. CRFs maximize a conditional probability of labels given an observation sequence, consider all possible sequences of labels and choose the label sequence which is most probable given the observation sequence

input [165]. CRFs find their applications especially successfully in NER and part of speech tagging tasks, and has been proved to achieve high performances in various NER tasks with biomedical text [166], [167].

3.3.3 Evaluation metrics to assess NER models

Similar to the text classification task, the most common metrics to evaluate performances of NER models are: Precision, Recall and F1 score. In information extraction context, these metrics can be calculated at different levels: at the token level and at the full named-entity level. At the token level, the metrics are determined by comparing predicted labels vs. actual labels of each token in the text using the formula as provided in section 3.2.3. At the entity level, these metrics need to be calculated in a consideration if a model is able to identify the exact span of an information item and if it is able to assign the correct entity type. Message Understanding Conference (MUC) in 1993 introduced new metrics for NER task, which consider different categories of errors in which these metrics can be defined as in terms of comparing the response of a system against the ground-truth annotation [168]:

- **Correct (COR)**: both are the same;
- **Incorrect (INC)**: the output of a system and the ground-truth annotation do not match;
- **Partial (PAR)**: system and the ground-truth annotation are somewhat "similar" but not the same;
- **Missing (MIS)**: a ground-truth annotation is not captured by a system;
- **Spurious (SPU)**: system produces a response which does not exist in the ground-truth annotation;

In 2013, The International Workshop on Semantic Evaluation (SemEval) introduced four different ways to measure precision/recall/f1-score results based on the metrics defined by MUC:

- **Strict**: the system predicts an entity correctly and the predicted text span is exactly matched with the ground-truth text span.
- **Type**: the system predicts an entity correctly and the predicted text span is partially matched with the ground-truth text span.

Table 3.1 shows examples of how each of the metrics defined by MUC falls into each of the scenarios: For Strict, Precision and Recall are calculated by the following formula:

$$Precision = \frac{COR}{COR+INC+PAR+SPU} \tag{3.6}$$

$$Recall = \frac{COR}{COR+INC+PAR+MISS} \tag{3.7}$$

Table 3.1: Examples of NER evaluation metrics

Golden Standard		System Prediction		Evaluation Schema	
Entity Type	**Surface String**	**Entity Type**	**Surface String**	**Type**	**Strict**
brand	warfarin			MIS	MIS
		brand	healthy	SPU	SPU
drug	warfarin	drug	of warfarin	COR	INC
drug	propranolol	brand	propranolol	INC	INC
drug	phenytoin	drug	phenytoin	COR	COR
group	contraceptives	drug	oral contraceptives	INC	INC

For Type, Precision and Recall are calculated by the following formula:

$$Precision = \frac{COR+0.5*PAR}{COR+INC+PAR+SPU} \tag{3.8}$$

$$Recall = \frac{COR+0.5*PAR}{COR+INC+PAR+MISS} \tag{3.9}$$

Chapter 4

Automatic extraction of study design to support evidence quality assessment

In chapter 1 and 2, I have discussed how to use **study design** as the most basic information to assess methodological quality of clinical research. This idea mainly derives from the concept of evidence hierarchy (Figure 1.3) that is well applied in EBM, in which studies are assigned to different levels of evidence quality based on their designs. Motivated by this possible use of **study design**, in this chapter, I present an approach of using NLP and ML to develop a classification model to automatically extract study design from clinical research and use that information to support clinical study quality assessment. The high-level implementation of such a model is shown in the diagram below (Figure 4.1), in which the model takes clinical papers as input, and classifies them into different study designs as outputs. This approach considers study design as the most coarse-grained level of information for EQA because of two reasons: (1) it can be determined at document level, which is the highest granularity of information can be determined per paper, and (2) it also helps to identify more fine-grained criteria applicable for a specific study design. In this chapter, we will present a development of such study design classification system and its application in the use case of drug-drug interaction (DDI) literature. More specifically, two classification models were developed and will be presented in this chapter: a model using Support Vector Machine (SVM) algorithm which was reported in a paper published in AMIA Annual Symposium 2020 [169] and a poster at Automatic Knowledge Base Construction conference 2019 [170]; and a deep learning model which is newly developed and only presented in this thesis.

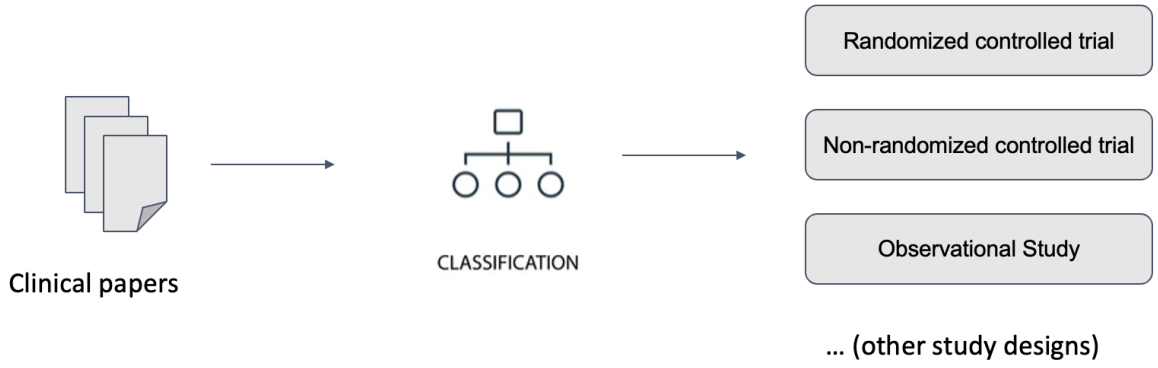

Figure 4.1: Classification of evidence quality types based on clinical study designs

4.1 Study design to assess evidence quality in drug-drug interaction literature

Identifying drug combinations that could result in a clinically meaningful alteration to patient safety or therapeutic efficacy is an important patient care activity, especially given that clinicians are known to have incomplete knowledge about drug-drug interactions (DDIs) [171], [172]. Computerized alerting systems can help clinicians by providing relevant reference information and suggestions, intelligently filtered and presented at appropriate times [173]. Unfortunately, the knowledge bases underlying these systems have long been known to be incomplete and inconsistent with one another. For example, a recent study by Fung et al. found that only 5% of 8.6 million unique interacting drug pairs were present in all 3 of the knowledge bases they included in the study [174].

We refer to individuals who maintain knowledge bases used by clinicians for DDI clinical decision support as compendium editors. Romagnoli et al. established the workflow of compendium editors which generally involves 4 main steps: topic identification, evidence search, evidence synthesis, and generating recommendations [175]. Focusing on the evidence synthesis step, compendium editors tend to evaluate evidence informally, with no dedicated support from information tools such as reference management software or databases. Although there exist systematic approaches to evaluate a collection of evidence relevant to establishing DDIs [176], [177], compendium editors generally reported using heuristic and subjective approaches to determine when sufficient evidence had been gathered to make a recommendation. Variation in evidence assessment suggests a potentially important factor underlying the lack of agreement that exists among different DDI knowledge bases.

Grizzle et al. introduced a guideline, named DRIVE (Drug Interaction eVidence Evaluation), to aid compendia editors as they assess DDI evidence. The guideline provided six evidence categories considered

sufficient for establishing the existence of a DDI, in which three of the categories are determined based on study design of the studies where the evidence comes from. Those include evidence from: ***well-designed and executed prospective controlled studies***; ***well-designed and executed observational studies***; and ***case reports*** that demonstrate probable or highly probable causality of an interaction [49]. The evidence categories then can be divided into sub-categories based on more fine-grained study design types. For example, prospective pharmacokinetic (PK) DDI study types include fixed-order studies, parallel group studies, and randomized crossover studies; while observational studies use claims or electronic medical records data to evaluate DDIs using cross-sectional, cohort, case-crossover, or case-control study designs.

Evaluating study design of research in clinical domain in general, and in DDI literature in particular, can present challenges because the nuances of flawed studies are not always obvious [178]. The research from Grizzle et al. found that inter-rater agreement on evidence sufficiency among compendia editors is poor [49]. On the other hand, the editors considered many studies insufficient because of flaws in study design. Small sample size, lack of transparency in methods, insufficient references to support extrapolations, and no inclusion of risk factors were a few of the examples cited for problematic study designs. Not all study flaws were as clearly delineated, as demonstrated by the following comment, *"This study design does not allow you to draw conclusions on causality of DDI"* [49]. Among the possible explanations for the finding from Grizzle's research is that experts tend to be subjective when assessing an evidence item's type and study design. For this reason, we think that a particularly promising future research direction would be using NLP and machine learning to automatically identify study design of DDI research, which will help experts be more efficient and objective in assessing DDI evidence from clinical literature.

While bibliographic databases such as MEDLINE include an article publication type, such as RCT publication type, the annotation is not applied with 100% accuracy or coverage. Studies have found that only about 85% of articles in MEDLINE considered RCTs for the purpose of systematic reviews are actually annotated with the publication type "Randomized Controlled Trial" [179], [180]. Because of that, automatic classification of study design has been a focus from computer scientists who develop tools to support evidence synthesis. However, those automatic classification models are often developed for biomedical literature retrieval purposes (e.g. searching for relevant studies for a systematic review), thus mostly are designed to recognize RCT studies only since this is the most common study design used in systematic reviews and meta analysis. Recently, Cohen et al. and his team developed a multi-tag machine learning model that recognizes 50 different publication types, including not only the core study designs of experimental and observational studies (RCT, cohort studies, case-controlled studies), but also the less common study designs such as follow up studies, longitudinal studies, etc. [181]. Nevertheless, none of the existing tools are designed specifically for evidence quality assessment. Therefore, in most cases, the study designs that they covered are not specific enough for this purpose. For example, most DDI studies in humans compare drug substrate (D) concentrations with and without the interacting drug (I), thus focusing on the pharmacokinetic type of

interaction. Therefore, the study designs that are often used for this purpose include: a randomized parallel group design (D in one group of subjects and D + I in another), a randomized crossover with a washout period (e.g., D followed by D + I, or D + I followed by D), or a randomized one-sequence crossover (e.g., D always followed by D + I or the reverse) [182]. These are the more fine-grained design types of RCTs that typical classification systems designed for retrieval purposes do not capture.

With this in mind, we developed and evaluated a classification system that identifies study designs that can be used for quality assessment purpose in the DDI literature. Our classification system was developed in conjunction with an ontology called DIDEO—the Potential Drug-drug Interaction and Potential Drug-drug Interaction Evidence Ontology, which defines 44 study types used in in vitro and in vivo pharmacokinetic DDI research [183]. Different from the existing models, by incorporating the study designs defined by the DIDEO ontology, our model captures not only high level study designs (e.g, RCT) but also low level study designs (RCT parallel-group) that are applicable for quality assessment purpose.

4.2 Methods

In this section, we will describe the methods to apply NLP models and techniques into the development of a classification system that identifies study designs and use it for quality assessment purpose in the DDI literature. Readers might want to refer back to chapter 3 for definitions of the NLP and ML concepts used in this method section.

We first will introduce the DIDEO ontology which is used as the backbone to develop our classification system, then we will discuss in details the steps of the development.

The DIDEO ontology is a foundational domain representation that specifies the necessary and sufficient conditions for 44 study design types that can be used to assess evidence from DDI research using terms either defined in DIDEO or imported from other formal ontologies [184]. We set out to test machine learning classifiers that predict 7 of the 44 specific types being reported in a DDI paper (blue boxes in Figure 4.2). These 7 types were chosen because they are commonly used in in vivo study designs that we thought would be useful for showing proof-of-concept and identifying requirements for a larger scale study. During the development of the training corpus for this study, we found a need to create additional types. For example, for those papers that were annotated as **Pharmacokinetic (PK) Trial** at the second level but were neither **Genotype PK trial** nor **Phenotype PK trial** at the third level, we labeled them as **non-polymorphic enzyme/transport PK Trial** which provides a novel alternative type at the second level. Figure 4.2 also shows the novel study types that we added (in orange).

The machine learning approach tested in this study was an ensemble of hierarchical classifiers (which means we combine multiple sub-classifiers into one single hierarchical classifier). This was chosen based on the observation that there exist multiple logical distinctions between the study designs. For example, **participant**

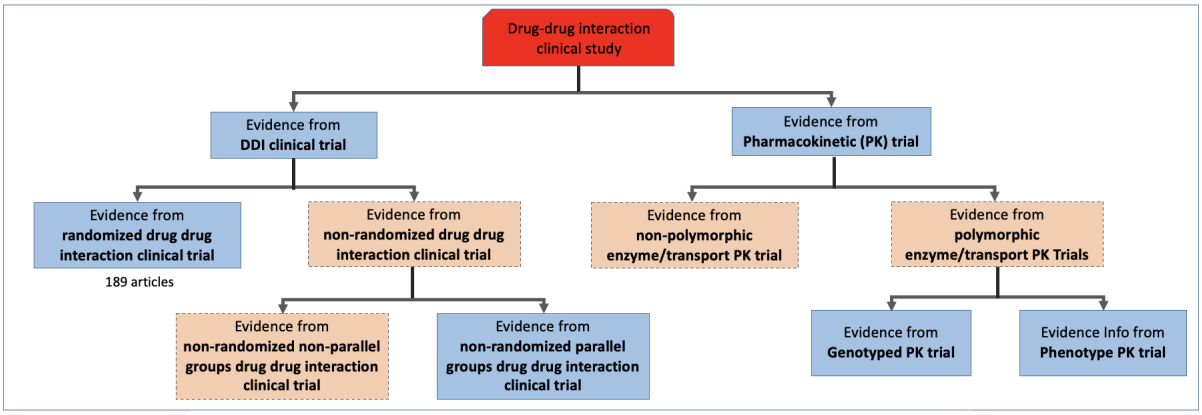

Figure 4.2: Study types hierarchy that was used in the classification system. The blue boxes represent the portion of DIDEO's study design type hierarchy that we used in this study. The orange boxes were added to cover the full range of study types identified in the training corpus. Numbers of papers in the training set for each category are shown below each box.

randomization is the major distinction between a *randomized DDI clinical trial* and a *non-randomized DDI clinical trial*. Similarly, **genetic genotyping** is the factor that distinguishes *the two polymorphic enzyme/transport PK trial* types. The hierarchical approach is a combination of multiple sub-classifiers– each is developed to classify a specific pair of study designs, therefore will be able to pick up on the primary distinguishing features better than a single model that forms the non-hierarchical approach. The development of the classification system consisted of four main steps:

- **Prepare data:** This step consists of two sub-steps:

  Annotate data: Since our task is the first of it kind, there was no available dataset to train the ML models that we were developing. Therefore, we manually annotated data from scratch.

  Preprocess data: We then preprocessed the annotated data to be ready for ML training and testing.

- **Design and develop baseline models:** in which we implemented two approaches to develop classification models: a support vector machine (SVM) classifier and a neural network classifier using PubMedBERT.

- **Evaluation of the baseline models:** in which we evaluated performances of the models (sub-classifiers individually, and hierarchical classifier as the whole) using Precision, Recall and F1; and discussed the comparison between the two approaches.

- **Error analysis** in which, we analyzed the wrong predictions of the two models on the testing data set; and also looked closer to the features used by the SVM model to understand the weaknesses and strengths of that model.

4.2.1 Prepare data

The training data set for this study contained 214 papers about DDIs that were systematically collected by DDI experts during previous knowledge representation research involving 65 drugs [185], [186]. The papers were manually annotated to assign study design labels from the DIDEO ontology using an annotation guideline by two experts of DDI literature. The annotation process was repeated to add two additional data sets that were used for "hold out" testing of the classifiers. One data set contained 32 DDI studies involving the same 65 drugs as for the training set. The other data set contained 94 papers identified using the same search strategy used for the other two data sets but focusing on drugs other than the 65 that were the focus of prior research.

In the pre-processing step, we automatically collected the papers' metadata, including titles, abstracts and PubMed publication types through the PubMed API [187]. We also manually collected full-text PDFs of these papers and converted them to plain text using the PDFMiner Python library [188]. For PDFs that were scanned images, we manually copied text from the PDFs and saved them as plain text. We then standardized the plain text files by converting the text into lowercase and removing English stop words from the Natural Language Toolkit (NLTK) library version 3.4.5 run on Python version 3.7 [189]. We also used MetaMap, a tool for recognizing biomedical concepts in text [190], to remove all of the drug and enzyme names from the text as well as regular expression to eliminate numeric strings, including numbers tied with measurement units. The reason for this is because we want to develop the models to be not limited to specific drugs, so as to generalizable as much as possible.

4.2.2 Design and development of classification system

To develop the classification system, we implemented two models, each corresponding traditional ML approach and deep learning approach that we discussed: (1) a model using Support Vector Machine (SVM); and (2) a model using neural network using pretrained PubMedBERT model. As discussed in chapter 3, we described how we implemented three components of the classification model including: feature extractor, machine learning algorithm to train model, and design of the classification model in the following.

Feature Engineering

For the SVM model, the features extracted from the papers included stemmed unigrams taken from the titles, abstracts, and the Methods sections of the papers. The Methods sections were included based on our observation during the annotation process that the section often contained information needed to determine the DDI study design that was not present in the title or abstract. Stemming was applied based on our observation that many words that experts use to distinguish study designs have the same roots (e.g. genotyped, genotyping and genotype) and should not be treated by the classification system as different

features. After the final feature engineering process, our feature space contained 11325 features for the 214 instances, corresponding to the 214 papers in our training set.

For the PubMedBERT model, feature engineering is not required. However, to use the pre-trained BERT model, we needed to pre-process the data in the same way it was trained. Figure 4.3 shows the input representation in PubMedBERT. To get the input ready for the PubMedBERT model, we transformed the original text through the following steps: (1) tokenize the input sequences; (2) insert [CLS] at the beginning; (3) insert [SEP] between each pair of sentences; (4) generate segment ids to indicate whether a token belongs to the first sequence or the second sequence; and (5) generate valid length depending on the max sequence length that our model accepts. We used the pre-trained PubMedBERT *base-uncased-abstract-fulltext* version to obtain initial embeddings for the inputs.

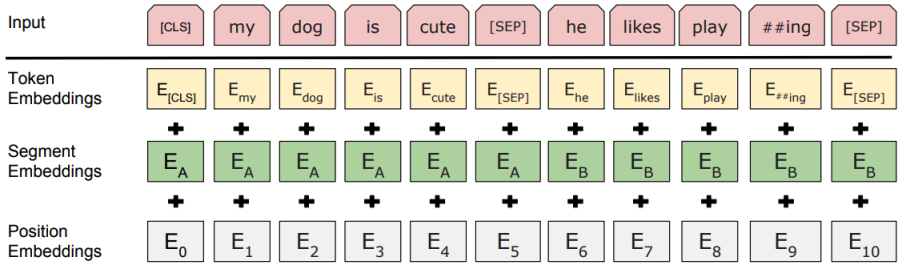

Figure 4.3: The input representation of BERT-variant models, including PubMedBERT [7]

Machine learning algorithm

We built a SVM model with linear kernel and applied class weights for balanced training data. The class weights was applied because we had an imbalanced data sets and some labels had very low number of examples (e.g. the number of articles that belong to "DDI clinical trial" group is much higher than the number of articles in the "Pharmacokinetic Trial" group). The class weights helps to penalize the mis-classification made by the minority class by setting a higher class weight and at the same time reducing the weight of the majority class. We used scikit-learn library[1] to implement this model.

The neural network model was essentially a fine-tuned model. Specially, we kept all layers of the original PubMedBERT model (as discussed in chapter 3), and only added a fully-connected linear layer with softmax to perform the classification task as the last layer of the model. During the training process, the weights from all the layers we kept from our original model will stay the same, and only the weights in our new layer was being updated. The model was trained with different set of hyper-parameters. However, here we only reported the one that achieved best performances: batch size (4), learning rate (1e-5), number of epochs (20), optimizer (Adam), and max sequence length (512). We used Huggingface library[2] to implement this model.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[2] https://huggingface.co/docs/transformers/model_doc/bert

Classification system design

The design of our classification system is shown in Figure 4.4. Corresponding to the hierarchy of study designs from the DIDEO ontology, our classification system is a combination of five sub-classifiers, each designed as a binary classifier that distinguished a specific pair of study design, hierarchically divided into three levels. Input of the whole hierarchical classification system will be DDI clinical papers. As the papers go through different sub-classifiers of the hierarchy at different levels, they will be given predicted study designs that belong to the corresponding level where they are at. The prediction results of the higher level sub-classifiers will decide which branch in the hierarchy a particular paper should go next. Subsequently, when the paper reaches to the lowest level of the hierarchy, the paper will be predicted with a final study design that belong to the lowest level.

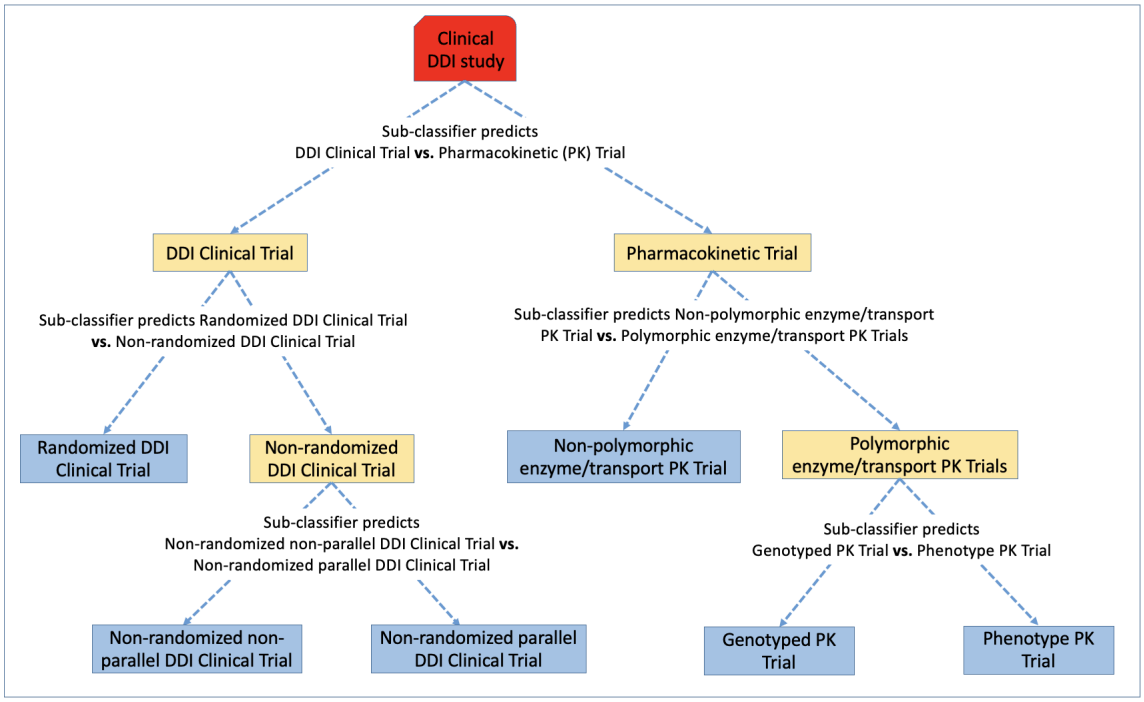

Figure 4.4: Design of the hierarchical classifier.

In both models, we use the same implementation setup, in which we use 5-fold cross-validation to randomly split the train and test sets to develop the sub-classifiers in order to prevent over-estimation of the classification system's accuracy. More specifically, in each cross-validation iteration, data was randomly split into a training set and a testing set. All papers in the training dataset were used to train and test the top-level sub-classifier. A subset of the training set from the top-level classifier was passed down to and used to train the next lower level sub-classifiers following a particular path in the hierarchy. Similarly, a subset of the testing set from the top classifier was used to test the next level sub-classifiers. This process was repeated until the sub-classifiers at the lowest level were trained and tested.

4.2.3 Evaluation of the classification system

We used Precision, Recall, F1 and AU-ROC metrics to evaluate performances of each sub-classifier and the hierarchical classifier as a whole. These metrics were calculated based on the predictions of each sub-classifier at each level against the actual labels of the papers at the same level. The four metrics were calculated in each sub-classifier for each cross-validation iteration. We then calculated the average of each metric, by dividing their sum by the number of cross-validation iterations (five). To take into account the hierarchical classifier structure, we supplemented these metrics with hierarchical precision, hierarchical recall, and F1 metrics for hierarchical classification systems [191]. We describe this in more detail in the next paragraph. The weighted average of the hierarchical classifier was calculated by dividing the sum of the hierarchical metrics by the number of cross validation (5 of them). After training the hierarchical classifier, we evaluated its performance on the two held-out data sets mentioned above. The hierarchical metrics aggregate the predictions of all sub-classifiers for every single data point into their formula [191]. For example, suppose that an instance is classified into the label *"Non RCT non parallel DDI Clinical trial"* while its actual label is *"Non RCT parallel DDI Clinical trial"* (Figure 3.3). To calculate our hierarchical measure, we extend the set of real labels:

Actual Label = "Non RCT parallel DDI Clinical Trial" with all its ancestors: ***Actual Labels' = "Non RCT parallel DDI Clinical Trial", "Non RCT DDI Clinical Trial", "DDI Clinical Trial"***.

We also extend the set of predicted labels:

Predicted Label = "Non-RCT non-parallel DDI Clinical Trial" with all its ancestors: ***Predicted Label' = "Non-RCT non-parallel DDI Clinical Trial", "Non RCT DDI Clinical Trial, "DDI Clinical Trial"***.

Then, the hierarchical precision (hP), recall (hR) and F1 (hF) score were calculated based on the extended label sets as following:

$$hP = \frac{\{ActualLabels\}' \cap \{PredictedLabels\}'}{\{PredictedLabels\}'} \tag{4.1}$$

$$hR = \frac{\{ActualLabels\}' \cap \{PredictedLabels\}'}{\{ActualLabels\}'} \tag{4.2}$$

$$hF = \frac{(\beta^2+1).hP.hR}{(\beta^2.hP+hR)} \tag{4.3}$$

In which, $\beta = 1$ giving precision and recall equal weights.

According to these formulas, the number of correctly assigned labels for this instance from the extended set would be the union of the actual labels and the predicted labels, which is 2, instead of 0. This approach reduces the penalty for mis-classification when the predicted label is "near" the actual label in the hierarchy.

Error Analysis

To better to understand how the models worked and what are the potential improvements that we can make in the future, we further analyzed the SVM model because it allows us to look into the features used by the model to make predictions. Specifically, we obtained insight into the SVM model's behavior by examining the most informative features (unigrams) that were associated with each study design as ranked by a Pearson's Chi-squared statistical test. We also further ran the SVM model on the set of 32 papers what are about different drugs, and compared the results between the hierarchical approach with non-hierarchical approach (a single classification that classifies all labels at the same level). We then looked at the papers that were given wrong predictions by the hierarchical classification system on the held-out data sets and analyzed the papers' titles, abstracts and Method sections in order to identify the possible reasons for the wrong predictions.

4.3 Results of automatic classification

4.3.1 Classification Performance

SVM model

Table 4.1 reports the sub-classifiers' prediction performance on the training data set using SVM. According to the results, all sub-classifiers in the hierarchy achieved average F1 scores, ranging from 0.77 to 0.87. The two sub-classifiers that have the highest performances are Randomized vs. Non-randomized DDI clinical trial classifier, and Genotyped vs. Phenotype PK trial. The classifier that has the lowest performance is the one that distinguishes polymorphic vs. non-polymorphic enzyme/transport PK trial. Randomized vs. non-randomized clinical trial classifier has a relative balanced performance between Precision and Recall, while other sub-classifiers yields larger differences between the two metrics.

PubMedBERT model

Table 4.2 reports the sub-classifiers' prediction performance on the training data set using PubMedBERT. According to the results, all sub-classifiers in the hierarchy achieved average F1 scores range from 0.87 to 0.96 which are higher than the performances of the SVM model. Comparing with the SVM model, the PubMedBERT model especially achieved much higher performances in the cases that have limited number of training data. Those are: the sub-classifier that distinguishes Polymorphic vs. Non-polymorphic enzyme/transport PK Trial (F1 scores are 0.77 vs. 0.88 for SVM vs. PubMedBERT respectively); and the sub-classifier that distinguishes Genotyped PK Trial vs. Phenotype PK Trial (F1 scores are 0.87 vs. 0.96 for SVM vs. PubMedBERT respectively).

Table 4.3 shows the performance of the SVM hierarchical classifier and the PubMedBERT classifier as a whole with the two held-out testing data sets. Overall, the PubMedBERT model achieved better results than

Table 4.1: Hierarchical classification system performance with 5-fold cross validation using SVM. Shown is the performance of each individual classifier at each level in the evidence hierarchy

Level	Sub-classifier	Average AUC ROC	Average Precision	Average Recall	Average F1
1	DDI Clinical Trial vs. Pharmacokinetic Trial	0.79	0.87	0.87	0.86
2	Randomized vs. Non-randomized DDI Clinical Trial	0.87	0.89	0.88	0.87
2	Polymorphic vs. Non-polymorphic enzyme/transport PK Trial	0.78	0.79	0.77	0.77
3	Non-randomized parallel vs. Non-randomized non-parallel DDI Clinical Trial	0.78	0.87	0.85	0.85
3	Genotyped PK Trial vs. Phenotype PK Trial	0.78	0.92	0.88	0.87

Table 4.2: Hierarchical classification system performance with 5-fold cross validation using PubMedBERT. Shown is the performance of each individual classifier at each level in the evidence hierarchy

Level	Sub-classifier	Average AUC ROC	Average Precision	Average Recall	Average F1
1	DDI Clinical Trial vs. Pharmacokinetic Trial	0.79	0.88	0.88	0.88
2	Randomized vs. Non-randomized DDI Clinical Trial	0.88	0.90	0.90	0.89
2	Polymorphic vs. Non-polymorphic enzyme/transport PK Trial	0.89	0.92	0.88	0.88
3	Non-randomized parallel vs. Non-randomized non-parallel DDI Clinical Trial	0.80	0.87	0.88	0.87
3	Genotyped PK Trial vs. Phenotype PK Trial	0.80	0.97	0.96	0.96

the SVM classifier in both testing sets. Individually, the SVM classifier achieved much better results with the data set about the same drugs as training data, compared with the data set about different drugs, especially in Recall. Similarly, the PubMedBERT model achieved significantly better results with the data set about the same drugs.

4.3.2 Error Analysis

To better to understand how the models worked, we further analyzed the SVM model because it allows us to look into the features used by the model to make predictions. More specifically, we printed out the most informative features (unigrams) used by the SVM model associated with each study design, ranked by Chi-square scores. We found that study designs that have terms strongly associated with them are

Table 4.3: Classification performance of the hierarchical classifiers, SVM and PubMertBERT, on the two held-out datasets

Dataset	Hierarchical SVM			Hierarchical PubMedBERT		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Performance on the held-out 32 papers about the same drugs as in the training set	0.80	0.86	0.83	0.92	0.92	0.92
Performance on the held-out 94 papers about entirely different drugs than in the training set	0.81	0.51	0.63	0.71	0.71	0.71

Table 4.4: Examples of the common unigrams used by SVM model between different study designs and their rankings

Ancestor study type	Study Type	Examples of features and their ranks (unigrams in stemmed format)
DDI Clinical Trial	Randomized	Placebo (rank 7), random (rank 26), crossov (rank 24), doubl (rank 61), blind (rank 64), Pharmacokinetic (rank 1149)
	Non-randomized non-parallel	Placebo (rank 26), random (rank 84), crossov (rank 807), doubl (rank 317), blind (153), Pharmacokinetic (rank 3807)
	Non-randomized parallel	Placebo (rank 579), random (rank 3898), crossov (rank 166), doubl (rank 1670), blind (847), Pharmacokinetic (rank 7518)
PK Trial	Non-polymorphic enzyme/Transport	Pharmacokinetic (rank 580), phenotyp (rank 389), genotype (rank 84)
	Genotyped	Pharmacokinetic (rank 398), phenotyp (rank 1218), genotype (rank 1)
	Phenotype	Pharmacokinetic (rank 182), phenotyp (rank 49), genotype (rank 1394)

easier to predict. For example, "Randomized DDI clinical trial" has terms related to this study design, including "random", "double", "blind", while "Genotyped PK Trial" has terms related to drug metabolism and excretion, including "genotype" and "polymorphism". Another observation is that certain unigrams are highly correlated with some of the study designs but not others, based on their Chi-square score ranking. For example, the "random" unigram and its variants (e.g. randomis, nonrandom) were given higher Chi-square scores and thus is more highly relevance ranked for RCT DDI Clinical Trial than any other types. In contrast, while the "genotyp" unigram was ranked as the most important unigram for Genotype evidence, it was ranked as less important for the others. Similarly, the "phenotyp" unigram is one of the top relevant unigrams for Phenotype , but not for Genotype . Table 4.4 shows the list of study design-associated features and their corresponding ranks in each study design.

We continued the error analysis by looking at the incorrect predictions of the SVM on the held-out 32

papers. Two examples are shown in Table 4.5. We found that the unigram features are not sufficient to assist the hierarchical classifier in making correct predictions in some cases where the study designs should be determined based on the whole context rather than single words. For example, in the first example of Table 4.5, the actual label is "Non RCT parallel group DDI trial", however, there are no mentions of **"parallel"** in the text. Instead, the authors described the parallel design differently by using phrases such as *"simultaneous and time-separated administration"* and *"in two treatment groups"*. In the second example in Table 3.4, the classifier's incorrect prediction of "RCT DDI Trial" was likely caused by the **"doubl", "blind" and "random"** unigrams which are among the most informative features for the RCT DDI Trial study design. However, in this case, they occur in the context of describing a population pharmacokinetics study rather than a DDI study.

Table 4.5: Examples of incorrect predictions of the hierarchical classifier on the held-out 32 papers

Actual study design	**Predicted study design**	**Sample text from the paper**
Non RCT parallel DDI Trial	Non RCT non parallel DDI Trial	Title: "Almorexant effects on CYP3A4 activity studied by its **simultaneous** and **time-separated** administration with simvastatin and atorvastatin." Abstract: "...To characterise further the previously observed cytochrome P450 3A4 (CYP3A4) interaction of the dual orexin receptor antagonist almorexant. Pharmacokinetic interactions were investigated **(n = 14 healthy male subjects in two treatment groups)** between almorexant at steady-state when administered either concomitantly..."
Non polymorphic enzyme transport PK Trial	RCT DDI Trial	Title: "Population pharmacokinetics and pharmacodynamics of rivaroxaban–an oral, direct factor Xa inhibitor in patients undergoing major orthopaedic surgery." Abstract: "...This analysis was performed to characterize the population pharmacokinetics and pharmacodynamics of rivaroxaban in patients participating in **two phase II**, **double-blind, randomized, active-comparator studies** of twice-daily rivaroxaban for the prevention of venous thromboembolism after total hip- or knee-replacement surgery..."

We also compared the prediction results of the SVM model vs. the PubMedBERT model on the data set of different drugs. We found that the PubMedBERT model performed much better than the SVM model in the cases that textual distinction is clear as shown in an sample in Table 4.6. For example, "Genotyped PK Trial" vs. "Phenotyped PK Trial" is the sub-classifier that the PubMedBERT model achieved highest performance during cross-validation training process ($F_1 = 0.96$). This trend continues with the testing set of different drugs, in which, out of 19 "Genotyped PK Trial" in the testing set, the PubMedBERT model predicted correctly 9; while the SVM model only predicted correctly 3. Looking at the text of these 19 trials, we found that popular keywords of genotyped studies such as "pharmacokinetic" or "genotype" appear in

Table 4.6: Example of SVM prediction vs. PubMedBERT prediction

Actual Study Design	Genotyped PK Trial
SVM Predicted Study Design	DDI Randomized clinical trials
PubMedBERT Predicted Study Design	Genotyped PK Trial
Title	Comparison of lansoprazole and famotidine for gastric acid inhibition during the daytime and night-time in different CYP2C19 **genotype** groups
Abstract	Fifteen healthy volunteers were given 20 mg famotidine twice a day
Methods Section	All subjects were first given a **placebo** dose in order to obtain the control or baseline 24-h intragastric pH data .... in a **randomized, crossover** manner for two separate 8-day periods...

all 19 trial (this aligned with the unigrams analysis that we performed above). However, at the same time, common keywords of "DDI clinical trials" such as "random", "healthy volunteers", or "crossover" also appear in 9 trials as shown in the example below. The SVM model made incorrect predictions for these articles. This can be interpreted as PubMedBERT taking the context surrounding keywords into account more successfully.

4.3.3 Discussion on the classification results

The results of both SVM and PubMedBERT models suggest that it is feasible to accurately automate the classification of a subset of DDI study designs. They also suggest that the hierarchical ensemble approach we tested based on the DIDEO evidence is a promising approach to build upon in future work. To our knowledge, this is the first study to test a hierarchical approach for classifying DDI clinical studies into highly specific evidence types based on study designs. A 2020 study on extracting evidence of drug repurposing classified studies into more basic evidence types such as "Pre-clinical" (F1 = 0.96), "Clinical observational study" (F1 = 0.84), and "Clinical trial" (F1 = 0.80) [192]. Another study in the same topic was published by Cohen et al. in 2021, in which the authors present a machine learning model that classifies 55 different publication types and study designs [181]. While both studies target broader sets of articles, our study focused on a more fine-grained detailed set of evidence types at lower levels of the evidence type ontology. While further work will be necessary to expand the classifiers to in vitro evidence types, this study is an important step towards more sophisticated computer support for DDI evidence synthesis.

Our study has several potential limitations. The training and held out datasets were relatively small. To overcome this problem, we need to obtain and have experts annotate more data. Alternatively, a computational approach to increasing the annotated data would be to semi-automatically collect and label the data using techniques such as rule-based distant supervision [193]. Each modeling approach has its own limitation and can be improved. For example, the SVM model used all features (unigrams) to train all of the sub-classifiers regardless of which labels they were predicting. In the future, developing different feature

selection strategies tailored to different sub-classifiers could be helpful, because in the error analysis, we found that some specific words and phrases (especially the ones indicating study design) are more important for some labels than others. As for the PubMedBERT model, we only fixed the learning rates. Using warmup learning rate schedule and trying out adaptive learning rate might help to improve the performances.

4.4 Summary of the chapter

In this work, drawing on an existing ontology of evidence types, DIDEO, we built a hierarchical classifier, which combined a series of sub-classifiers to categorize evidence types of DDI studies based on their study designs. In the future, such an automatic classification system could be a key component of a computerized system to help experts be more efficiently assess DDI evidence. Similar NLP approach could be developed to distinguish different study designs, beyond DDI literature, as the most coarse criteria to assess evidence quality of clinical studies in general. For example, a hierarchical classification model could be developed to classify study designs beyond the ones included in this study. And the output of such a model can help to determine what fine-grained information needs to be extracted for evidence quality assessment.

Chapter 5

Automatic reporting quality assessment of randomized clinical trials using CONSORT guidelines

In Chapter 1 and 2, we discussed the two aspects to assess quality of a clinical study: methodological quality and reporting quality. In Chapter 4, we presented an NLP approach to extract study designs as the most coarse-grained information to support EQA, mainly from methodological quality standpoint. In this chapter, we will focus on reporting quality and seek to classify randomized controlled trial result publications at finer granularity to allow their assessment for compliance with the CONSORT guidelines, a checklist recommended by many clinical journals. In particular, we used NLP and ML to develop a classification model that automatically maps sentences in full-text RCT publications to the information items recommended in the CONSORT guidelines. The high level implementation of such a model is shown in the diagram below (Figure 5.1), in which the model takes sentences from RCT publications as input, and classifies them into different information items recommended from reporting guidelines as outputs. Outputs of such a model will be helpful for quality assessment purpose in several ways: (1) it helps to assess compliance of authors with the guidelines, which is the most efficient way to assess reporting quality of a research, (2) by highlighting relevant methodological information at sentence level (since many CONSORT items are about methodology), this paves the way for more fine-grained automated analysis of methodological quality assessment in Chapter 6. The content of this chapter is based on a journal article in Journal of Biomedical Informatics [194], a conference publication in AMIA Informatics Summit 2022 [195], and a poster presented at the AMIA Annual Symposium 2020 [196].

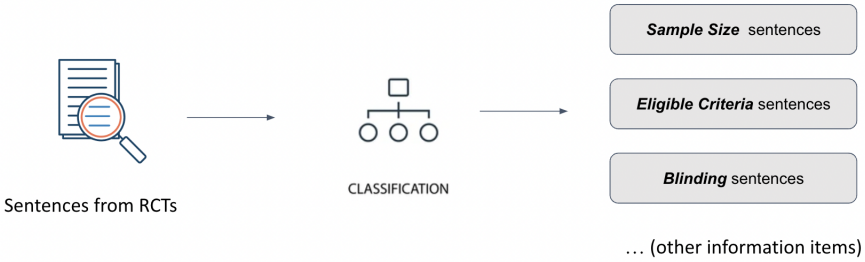

Figure 5.1: Classification of CONSORT Items from RCTs

5.1 Current practices of reporting quality assessment and NLP support

In Chapter 1, we briefly introduced the CONSORT Statement, which is an evidence-based, minimum set of recommendations for reporting RCTs. The statement comprises two components: a 25-item checklist (called the CONSORT checklist), which focuses on reporting how the trial was designed, analyzed, and interpreted; and a flow diagram, which is designed to display the progress of all participants through the trial. In chapter 2, we also discussed how the CONSORT checklist has been used as a guideline to facilitate the process of assessing RCT reporting quality. In this section, we will look deeper into that practice, to understand how researchers and other stakeholders leverage the existing reporting checklists such as CONSORT to assess reporting quality of clinical studies, as well as the existing NLP and ML research that have been developed to automate such process.

Using reporting guidelines to assess reporting quality of clinical studies is a well established practice and being applied in many quality assessment studies for a couple of decades [197]–[200]. Since RCTs are considered as gold standard for evaluating the effectiveness of interventions due to their potential to limit all sorts of bias, the CONSORT checklist is the most commonly used and recommended by journal editors. For example, in a survey of the editors from 165 high impact journals, Hopewell et al. in 2008 found that 88% of journals recommended authors comply with the CONSORT Statement, in which 62% said they would require authors using the statement as reporting mandate [201]. The use of CONSORT checklist has been shown to enhance the reporting quality of RCTs [202]. In an analysis of reporting quality in RCTs after adoption of the CONSORT statement, Kane et al. found that the CONSORT guideline helped to improve in all aspects of RCT reporting significantly and consistently when they were implemented [202]. Nevertheless, despite the importance and extensive applications of the CONSORT checklist into the current quality assessment practices, the process of cross-checking information reported in clinical publications against information items recommended in CONSORT checklist is still done manually. For example, in one of the most recent meta-research studies to examine the reporting quality and transparent research practices, Schulz et al.

2022 used checklist items from CONSORT as criteria to assess methods and results reporting quality of sports medicine and orthopaedic clinical trials [203]. In this study, assessors/reviewers needed to manually extract methodology-related information from full-text articles and compared with criteria recommended in CONSORT [203], which makes the process time-consuming and hard to scale.

From technical standpoint, NLP and ML can be applied to develop a system that mimic the current assessment practices by automatically mapping information items recommended from the checklist with sentences from full text clinical papers. In fact, in the same meta-research study, Schulz suggested automated screening tools may efficiently flag missing information for assessors. However, the authors also pointed out that most of the existing tools are available to screen for risk of bias (such as RobotReviewer), and there is much less focus on tools that leverage reporting checklists to extract reporting quality criteria that can assist assessors in quality assessment task [203]. In a another screening study to assess reporting quality of COVID-19 preprints, Weissgerber et al. used variety of tools that help to automate the screening process [204]. Those are SciScore, an automated tool that evaluates research articles based on their adherence to key rigor criteria [205]; ODDPub, a text mining model to detect data sharing in biomedical publications [206]; JetFighter, a screening tool for preprints which use color maps to improve data presentation [207], and an automatic recognition model of self-acknowledged limitations from clinical trials [208]. However, none of these tools considers the use of reporting checklists as a comprehensive guideline for automatic quality assessment (except the limitation recognition model, which covers one information item –"limitation", recommended in the CONSORT checklist [208]). In a study to assess quality of 176,620 RCTs (including methodological quality via risk of bias assessment and reporting quality via checking reporting compliance), the authors used a very simple approach– searching for the "CONSORT" keyword in the full-text publications [209] and used that as an indication for reporting quality, e.g. article that mentioned CONSORT keyword in the full text was assumed to be complied with the guideline. Motivated by the opportunity to leverage CONSORT checking to automate quality assessment process, we developed a classification model to recognize information items recommended in the CONSORT checklist from full-text clinical articles as described in the following sections.

5.2 Methods

From technical standpoint, we cast the problem of associating sentences with CONSORT items as a sentence classification task. Readers might need to refer to the NLP and ML preliminaries in chapter 3 for details of models, techniques, as well as evaluation metrics that we used for the classification model development in this chapter.

We limit the experiment to the text that was taken from Methods sections of RCT papers and try to map them with methodology-related CONSORT items which cover the key methodological details most relevant to trial methodological quality. The reason is because, in general, Methods are the major focus of studies of

reporting quality as well as rigor and replication/reproducibiliy. Also, almost half of the CONSORT items are methods-related. In addition, Output of this work could be used to support the development of later research in chapter 5, in which we look at fine-grained methodological characteristics of RCTs that can be used for methodological quality assessment. As a result, in this work, we focus on 17 specific information items that are associated with the trial methodology in the CONSORT checklist as shown in Table 5.1.

Table 5.1: List of information items from CONSORT checklist that are used as label outputs of our classification system

Section/Topic	Item Number	Item Description
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons
Participants	4a	Eligibility criteria for participants
	4b	Settings and locations where the data were collected
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
	6b	Any changes to trial outcomes after the trial commenced, with reasons
Sample size	7a	How sample size was determined
	7b	Explanation of any interim analyses & stopping guidelines
Randomization: Sequence generation	8a	Method used to generate the random allocation sequence
	8b	Type of randomisation; details of any restriction
Randomization: Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned
Randomization: Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how
	11b	If relevant, description of the similarity of interventions
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses

Similar to the work in chapter 3, the development of the classification model contains 4 main steps:

- **Prepare data:** since our task is the first of it kind, there was no available dataset to train ML models. We built models based on a dataset that had already been annotated by experts involved in the study.
- **Design and develop baseline models:** in which we implemented two baseline classification models

corresponding to each of the machine learning approaches: a traditional ML model using support vector machine (SVM); and a deep learning model using BioBERT.

- **Evaluation of the baseline models:** in which we evaluated the performances of baseline models using Precision, Recall and F1; and provide a comparison between the two models.
- **Improvement of baseline models using automatic labeled data:** in which we tried to address a limitation of the baseline models (their small training data set). In particular, we attempted to improve the baseline models by applying weak supervision to automatically generate labeled data that can be used for training models.

5.2.1 Prepare Data

To develop the classification model, we used the data set which was manually annotated by the co-authors in an annotation study conducted prior to the development of the classification system. We referred to this data set as CONSORT-TM. In the scope of this thesis, we will only briefly introduce the data set, and will focus more on model development.

The data set contained 50 articles from 11 journals annotated with the full-list of 37 items from the original CONSORT guidelines. The 50 RCT papers were annotated by an annotation team of six. A subset (31 articles) was double-annotated and adjudicated, while 19 were annotated by a single annotator and reconciled by another. The annotators annotated each sentence in every article with the relevant CONSORT items (or none). Inter-annotator agreement was measured at article level using the Measuring Agreement on Set-Valued Items (MASI), which is a distance metric to quantify the degree of similarity across annotations [210], and at checklist item level using Krippendorff's alpha, which is a reliability coefficient to measure the agreement among observers [211]. The inter-annotator agreement at both levels was moderate (mean MASI = 0.60, median = 0.63; and average Krippendorff alpha = 0.47). After the annotation study, the corpus contains over 10K sentences, 45% of which were annotated with CONSORT checklist items (all were considered). In the end, the CONSORT-TM data set was created with a total of 10,709 sentences, 4,845 (45%) of which were annotated with 5,246 labels. About 6.5% of the annotated sentences were annotated with multiple CONSORT items. An example of a sentence labeled with multiple CONSORT items is *"After screening, patients were randomised to bosentan or placebo (1:1 ratio) by* ***sequential allocation of randomization numbers*** *distributed to each center in* ***blocks*** *of four"*. This item was annotated with items 8a and 8b since it indicates both information: method used to generate the random allocation sequence and type of randomisation (highlighted in bold). Since we only focus on 17 methodology-related information items, only sentences that belong to the Methods sections were used for development. This resulted into a corpus with total of 2,564 sentences.

5.2.2 Develop baseline models

Since there are sentences which were annotated with more than one CONSORT item in our data set, the classification model was built as a multi-label classifier. Two ML models were developed: a SVM model and a BioBERT model. Similar to the classification model development in Chapter 4, the process included 3 main steps: feature engineering, ML algorithm and classification model.

Feature Engineering

The SVM model used the following features: tf-idf representations of unigrams of the sentence and the enclosing subsection header. More specifically, the section header was prepended to the sentence and included in TF-IDF calculation. The reason we included section headers as features is because the Methods sections of RCT papers often contain sub-sections, each often describing different aspects of the RCT methodology such as inclusion and exclusion criteria, randomization or statistical methods. Therefore, the section headers can be strong indicators that help to classify sentences (e.g. sentences which appear in "randomization" sub-section of the "methods" section are likely to map to CONSORT items related to randomization such as 8a, 8b, 9 or 10). We excluded common English words using the NLTK stopword list.

As for the deep learning model, we used BioBERT, another variant of the original BERT language model as a pre-trained model of our classifier. BioBERT has the same architecture as the original BERT model, and was also trained on PubMed abstracts and PubMed Central full-text articles as PubmedBERT. The main difference between BioBERT vs. PubmedBERT is that PubMedBERT was trained on a larger set of biomedical corpora, which contains total 16.8 billion words and it was trained from scratch; while BioBERT was trained on a slightly smaller corpora which contains total 13.5 billion words and used the original BERT weight initialization. Even though, it is known that PubMedBERT is able to obtain consistent gains over BioBERT in most tasks [158], by the time this project was developing, PubMedBERT was not released yet, therefore, we used BioBERT as the state-of-the-art language model at that point. To use pre-trained BioBERT, we also transformed original text into the input representation of BERT model (which contained three components: token embeddings, section embeddings and position embeddings) as described in Chapter 4.

Machine learning algorithm and classification model

Similar to SVM model in Chapter 4, the SVM model was linear kernel. We also used the LIBLINEAR SVM in the scikit-learn package to implement this. C regularization parameter was set to 10 after a grid search. The classifier was embedded into a one-vs-rest classifier to enable prediction of multiple labels for each sentence.

As for the BioBERT model, sentence text and its subsection header were fed as input to the BioBERT encoder, whose output was then used to train the final sigmoid layer for multi-label classification. We used

the simpletransformers package[1] to implement multi-label text classification. The following hyperparameters were used for model training and evaluation: batch size (4), learning rate (3e-5), number of epochs (30), optimizer (Adam), dropout (0.1).

5.2.3 Results of baseline models

The results of the experiments with two methods (SVM and BioBERT) are shown in Table 5.2. Overall, the BioBERT-based model performed better in the majority of categories (average 0.82 precision, 0.63 recall, 0.72 F1 score). This model performed particularly well for items with larger numbers of annotations, such as Interventions (5), Primary and Secondary Outcomes Measures (6a), Statistical Methods (12a). However, it performed poorly for rare items yielding no correct predictions for several items, such as Changes to Trial Outcomes (6b), Interim Analyses and Stopping Guidelines (7b), Allocation Concealment (9) which led to its lower macro-averaged performance, compared to macro-averaged performance of the linear SVM classifier. The results based on model combinations are provided in Table 5.3. We only provide the micro- and macro-averaged results for these combinations. The SVM and BioBERT model combination (SVM + BERT) improves upon the best base model by about 2 F1 points (0.72 to 0.74).

5.2.4 Discussion on the baseline results

Both models, SVM and BioBERT, achieved better performances for common checklist items such as Intervention (5), Primary and Secondary Outcomes (6a), and Statistical Methods (12a) comparing with the rare items. This suggests the need for large quantities of labeled data for training effective supervised machine learning models. We found that approaches that combine predictions from different models can also improve classification performance. We used standard settings for supervised learning, and it may be possible to achieve better performance with more advanced features or modeling approaches [155]. In the case of SVM classification, we experimented with semantic features derived from MetaMap [212] (entities and their semantic types extracted from sentences). Semantic types feature slightly improved results, although not statistically significantly. However, this approach could be explored further using more sophisticated methods. In another approach, we can cast the problem as a sequence labeling task, leveraging the fact that discussion of items often follows a predictable sequence (e.g., Methods sections generally begin with Study Design sentences). In training the BioBERT-based model, a simple sigmoid layer was used on top of the BERT encoder for classification, which can be substituted by more layers or a more complex neural architecture, such as convolutional or recurrent neural network (CNN or RNN) for higher classification performance. Note, however, that the BioBERT model is already much more complex than and takes orders of magnitude longer to train than the SVM classifier (hours vs. seconds). Improvements due to additional layers or architectural features may not be sufficiently large to justify the added complexity. Overall, our preliminary results were

[1] https://simpletransformers.ai/

Table 5.2: Classification results per CONSORT Items from SVM model vs. BioBERT model. 3a: Trial Design; 3b: Changes to Trial Design; 4a: Eligibility Criteria; 4b: Data Collection Setting; 5: Interventions; 6a: Outcomes; 6b: Changes to Outcomes; 7a: Sample Size Determination; 7b: Interim Analyses/Stopping Guidelines; 8a: Sequence Generation; 8b: Randomization Type; 9: Allocation Concealment; 10: Randomization Implementation; 11a: Blinding Procedure; 11b: Similarity of Interventions; 12a: Statistical Methods for Outcomes; 12b: Statistical Methods for Other Analyses.

CONSORT Item	**Number of instances**	**SVM**			**BioBERT**		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
3a	67	0.70	0.58	0.62	0.93	0.49	0.63
3b	10	0.00	0.00	0.00	0.00	0.00	0.00
4a	160	0.87	0.60	0.70	0.90	0.82	0.85
4b	39	0.88	0.46	0.59	0.8	0.24	0.36
5	269	0.66	0.5	0.56	0.76	0.69	0.72
6a	655	0.74	0.64	0.69	0.84	0.78	0.81
6b	6	0.00	0.00	0.00	0.00	0.00	0.00
7a	113	0.93	0.70	0.79	0.88	0.80	0.84
7b	16	0.80	0.64	0.70	0.00	0.00	0.00
8a	43	0.92	0.64	0.74	0.86	0.26	0.38
8b	49	0.67	0.46	0.54	0.71	0.29	0.38
9	22	0.28	0.19	0.22	0.00	0.00	0.00
10	57	0.68	0.25	0.36	0.72	0.15	0.24
11a	57	0.84	0.45	0.58	0.77	0.29	0.42
11b	18	0.20	0.13	0.16	0.00	0.00	0.00
12a	269	0.72	0.64	0.67	0.75	0.76	0.75
12b	72	0.32	0.13	0.17	0.05	0.03	0.04
Micro	**-**	**0.74**	**0.56**	**0.64**	**0.82**	**0.63**	**0.72**
Macro	**-**	**0.60**	**0.41**	**0.48**	**0.52**	**0.33**	**0.38**

Table 5.3: Results of combining SVM and BioBERT models

Model	**Micro-average**			**Macro-average**		
	Precision	Recall	F1	Precision	Recall	F1
SVM	0.74	0.56	0.64	0.60	0.41	0.48
BioBERT	0.82	0.63	0.72	0.52	0.33	0.38
SVM + BioBERT	0.73	0.74	0.74	0.67	0.50	0.54

encouraging, although it is clear that there remains much room for improvement. Performance for several items may be acceptable for practical use (e.g., Eligibility Criteria, Sample Size Determination), whereas more work is needed for others.

A limitation of this study is that our data set consists of a small number of publications from 11 journals, which may not be representative of all RCT articles and also prevent us from improving the neural network approach which often requires a large amount of data to enhance the performance. However, manual annotation at sentence level takes a tremendous amount of time and human effort. In addition, our results show that the current text mining methods can recognize CONSORT items that are commonly reported with relative success, whereas they struggle with those items that are not commonly reported. To address this limitation, in the next section, I will present an extension of this work, in which, we explored weak

supervision to automatically annotate a larger number of clinical trial publications using simple heuristic rules and then use the resulting (somewhat noisy) data with the goal of training more effective classifiers, reported in the next section [193].

5.3 Improving baseline models with automatic labeled data generated by weak supervision

Considering the high cost and time demands of annotation and the need for large amounts of annotated data for training effective machine learning models, methods to automatically assign (somewhat noisy) labels to unlabeled data have been proposed. One well-known technique of weak supervision is distant supervision, originally proposed for relation extraction [193]. It is based on the assumption that any sentence that contains a pair of entities that participates in a known relation in a knowledge-base is likely to express that relation in some way. Similar approaches have been applied in biomedical text mining, as well. For example, Marshall et al. used risk-of-bias judgements as well as related text snippets that those judgements are based on in the Cochrane database of systematic reviews to automatically label sentences in RCT publications and used the noisy labels to train models for assessing risk of bias in the publications [121]. In this subsequent work, we investigated whether weak supervision techniques can be used to effectively label additional data and improve the baseline models that we had developed. More specifically, we focused on weak supervision using the Snorkel– a machine learning tool that can automatically generate labels for data based on human-defined labeling functions derived from a small number of available human labeled training examples [213] and used the labels that it generated as additional data for the baseline BioBERT-based model.

5.3.1 Snorkel

Snorkel has been proposed as a general weak supervision framework [213]. Snorkel is a ML model that learns the quality and correlations of multiple labeling functions using statistical modeling techniques. Labeling functions (LFs) are heuristic rules that assign weak labels to unlabeled instances. In practice, rules based on keywords, syntactic structures, or derived from external knowledge bases are commonly used. Generally, it is desirable for LFs to have high coverage and low overlap. In other words, we would like them to apply to as many instances as possible in our dataset, while remaining "unique" enough to distinguish instances with different labels. Given a set of LFs, Snorkel applies each to all instances to generate a label matrix. Next, it pools noisy signals from the label matrix into a generative model to learn the agreements and disagreements of the LFs, to assess the weights of accuracy for each LF. The model then takes into account these accuracies to make the final label prediction for each sentence. The predictions from the previous step can be used as probabilistic training labels for a noise-aware discriminative model which is intended to generalize beyond

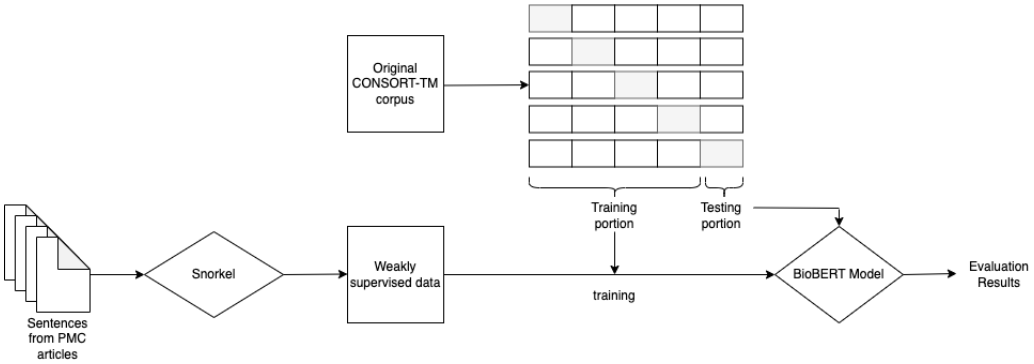

Figure 5.2: Training and evaluation with weakly supervised data

the information expressed in the labeling functions. Snorkel has been applied to several biomedical text mining tasks, outperforming distant supervision baselines and approaching manual supervision [213]. Other weak supervision approaches have also been developed for biomedical NLP tasks, including smoking status classification from clinical notes [214], semantic indexing [215], and clinical entity classification [216].

5.3.2 Materials and methods

We explored how to use Snorkel to automatically generate labeled data that can be used for training, thus improving the performance of the baseline BioBERT-based model. In this section, we first described the collection and pre-processing of unlabeled RCT data from PubMed Central (PMC). Then, we discussed the methodology to use Snorkel framework to automatically generate labelled data. Lastly, we used the automatically labeled data to train the BioBERT model and compared the results with baseline model to determine if Snorkel data help. The overall procedure is illustrated in Figure 5.2.

Data collection and pre-processing

We followed the data collection strategy used in the development baseline models to obtain a large set of RCT articles. Cochrane precision-maximizing search query[2] was used on 1/15/2021 to search PMC Open Access subset (PMC-OA) for RCT articles published between 1/1/2011 and 12/31/2020[3]. The results were further limited to articles that have full-text XML in PMC-OA. To get a more reliable RCT subset (since publication types in PubMed can be inaccurate), we filtered the results through RCT Tagger [217], a machine learning model that determines whether a publication is a RCT or not. Its accuracy was found to be 99.7% in predicting RCT studies included in Cochrane systematic reviews [218]. Lastly, we eliminated publications with the word *protocol* in their titles (which indicates that the article is a protocol, not a result RCT publication) for which CONSORT checklist items are not applicable.

[2] https://work.cochrane.org/pubmed
[3] The start date is chosen based on the most recent publication of CONSORT guidelines (2010) [29]

We used NCBI e-utilities API[4] to retrieve publications in XML format, and split them into sentences using our custom sentence splitter [219]. Only sentences that belong to Methods section of the publications were taken into account, again following the baseline model setup. Stanford CoreNLP package was used for tokenization and part-of-speech tagging [220]. We eliminated the sentences meeting the following criteria from further consideration, since they are unlikely to indicate CONSORT methodology items: (1) Contains fewer than five tokens; (2) Contains numbers only; (3) Is a section header or a table/figure caption. The reason we filtered out sentences that are empty and have less than 5 words is because we think short sentences might not convey a full meaning, thus are not good examples for the automatic labeling process.

Models setup

Our best-performing baseline model was the BioBERT-based, which was implemented using `simpletransformers`[5] package. We refer to this model as BASELINE below. In this extension experiment, we used the `huggingface`[6] BioBERT implementation. While mostly using the same hyperparameters as BASELINE (batch size: 4, number of epochs: 30, optimizer: Adam, dropout: 0.1), we modified two hyperparameters. First, we used adaptive learning rate instead of a fixed learning rate to optimize the algorithm with different rates based the model performance during training. Second, we set the gradient accumulation steps to 1 (16 for BASELINE), which increases the frequency of model parameter updates. We refer to this optimized model as BASELINE_OPT below.

Generating weak labels using Snorkel

Snorkel generates weak labels in three steps: a) Labeling functions construction; b) Creation of a generative model to capture label agreements/disagreements; and c) Generation of probabilistic labels for sentences. Input for Snorkel pipeline are unlabeled sentences from RCT publications from PMC-OA.

For labeling functions construction, we used three approaches to label CONSORT items: keyword-based, section header-based, and sentence similarity-based. 17 individual LFs were created for each approach (one corresponding to each label).

Keyword-based LFs: Each CONSORT item is associated with a set of keywords or phrases (e.g., *power to detect* with Sample Size Determination (7a)). A total of 232 phrases are used. Each LF checks whether an input sentence contains one of its key phrases, and if so, returns the corresponding label as a weak label (or NO-LABEL, if the sentence does not contain any relevant keyword/phrase). An example of keyword-based labeling function for Blinding information item is provided below. In this example, a keyword-based LB is defined by a list of keywords that are most common for Blinding (11a) item in the CONSORT checklist. When applying this LB, Snorkel model will detect the appearances of these keywords in the text and assign

[4] https://https://www.ncbi.nlm.nih.gov/books/NBK25501/
[5] https://simpletransformers.ai/
[6] https://huggingface.co/

label 11a to the text accordingly.

```
keyword_label_CONSORT_11a = make_keyword_lf(keywords=["to treatment allocation",
                        "masked to treatment", "masked", "blinded to",
                        "blind to","blinding",  "double-blinded to",
                        "masking","not have access"],
                label = CONSORT_11a)
```

Section header-based LFs. These LFs also mimic a baseline method from the previous work. In this case, common subsection headers in Methods sections are associated with CONSORT labels. 48 section header keywords/phrases are mapped to CONSORT items (e.g., the word *concealment* to the item Allocation Concealment (9)). These LFs check whether the header of the section to which the sentence belongs matches one of the relevant key phrases. An example of section header-based labeling function for Blinding information item is provided below. In this example, a section header-based LB is defined by searching for "masking" and "blinding" in the section headers using regular expression. When applying this LB, Snorkel model will detect the appearances of these keywords in the headers and assign label 11a to the instance accordingly.

```
@labeling_function()
def regex_11a(x):
    if ((re.search(r"blinding", x["section_hierarchy"], flags=re.I))
    or (re.search(r"masking", x["section_hierarchy"], flags=re.I))):
        return CONSORT_11a
    else:
        return CONSORT_0
```

Sentence similarity-based LFs. These LFs assign weak labels to unlabeled sentences based on their similarity to a set of "ground truth" sentences (95 sentences provided as examples for checklist items in the CONSORT Explanation and Elaboration document [75] and the CONSORT website[7]). We used BioBERT to generate vector representations of these sentences. Given an unlabeled sentence, we calculate its cosine similarity with every ground truth sentence and consider two labels based on similarity scores: the label of the sentence with the highest similarity and the label that appears most frequently for the top 10 most similar ground truth sentences. If two labels are the same, we use it as the sentence label. Manual checks showed this combination to be more accurate than the most similar sentence label only. An example of sentence similarity-based labeling function is provided below:

```
@labeling_function()
def similarity_lookup_CONSORT_11a(x):
```

[7] http://www.consort-statement.org/examples/sample

```
if ((x["highest_frequency_label_in_top10"] == x["highest_similarity_label"]) and
(x["highest_frequency_label_in_top10"] == "11a")):
    return CONSORT_11a
else:
    return CONSORT_0
```

Snorkel applies all LFs to generate a LF matrix that shows the coverage, overlaps, and conflicts between the LFs. *Coverage* information indicates the fraction of the dataset to which a particular LF is applied. *Overlap* shows the fraction of dataset where a particular LF and at least one other LF agree. *Conflict* indicates the fraction of dataset where a particular LF and at least one other LF disagree. Snorkel pools noisy signals from the these three features into a generative model to learn the agreements and disagreements of the LFs, thus assessing the weights of accuracy for each LF. The model then takes into account these accuracies to make a final label prediction for each sentence.

Evaluation

To evaluate whether weak supervision generated labels useful for improving sentence classification performance, we compared the results obtained with BASELINE model using 5-fold cross validation to results obtained when weakly labeled examples from different strategies are added to the training portion of the folds in cross validation. In this setup, data used for validation and testing in each fold remain the same for all the models. We used precision, recall, and their harmonic mean, F_1 score, and calculated 95% confidence intervals. In addition to calculating these measures per CONSORT item, we also report micro- and macro-averaged results and the area under ROC curve (AUC).

5.3.3 Results

Automatic labeled data from Snorkel results

Our search strategy retrieved 608K RCTs from PubMed, 155,183 of which have XML full text in PMC. RCT Tagger predicted 71,948 of these as RCTs. Considering only those predicted with a confidence score over 0.95 reduced the dataset to 14,534 publications. Further eliminating publications with *protocol* in the title, we obtained a set of 11,988 papers. A total of 721,948 sentences from these publications was reduced to 551,936 sentences after filtering.

We processed 551,936 unlabeled sentences using the Snorkel model, which generated 17 probabilities for each sentence. We empirically set a probability threshold of 0.8 to predict the final weak labels for the unlabeled sentences. If no label was predicted with a probability higher than 0.8, no label was assigned. The distribution of weak labels generated by Snorkel are shown in Table 5.4. Most weak labels corresponded to items that are already relatively well-represented in the dataset; thus, we limited the number of weakly

Table 5.4: The frequency of each methodology item in the original human annotated data set and the augmented data generated by Snorkel

CONSORT Item Snorkel	Number of instances Snorkel	Number of instances Original Data
Trial Design (3a)	3,932	67
Changes to Trial Design (3b)	0	10
Eligibility Criteria (4a)	17,182	160
Data Collection Setting (4b)	740	39
Interventions (5)	11,415	269
Outcomes (6a)	24,104	655
Changes to Outcomes (6b)	0	6
Sample Size Determination (7a)	6,674	113
Interim Analyses / Stopping Guidelines (7b)	124	16
Sequence Generation (8a)	7	43
Randomization Type (8b)	2,915	49
Allocation Concealment (9)	274	22
Randomization Implementation (10)	1,785	57
Blinding (11a)	525	57
Similarity of Interventions (11b)	3	18
Statistical Methods for Outcomes (12a)	45,353	269
Statistical Methods for Other Analyses (12b)	49	72
NO LABEL	436,854	

labeled examples for each CONSORT item to a pre-determined threshold in our classification experiments and randomly sampled these examples. We report the results with the threshold that performed best in our experiments (500).

5.3.4 Classification results

We evaluated BASELINE and BASELINE_OPT models using 5-fold cross-validation. For brevity, we only report the weak supervision results for the best-performing model-data size combinations. This is BASELINE_OPT model augmented with maximum 500 examples per label from Snorkel. The results are provided in Table 5.5. The results show that hyperparameter tuning (BASELINE_OPT) makes a significant difference in performance (5% increase in micro-F_1 and 36% in macro-F_1); while Snorkel data does improve the original BASELINE and it also leads to a slight performance degradation comparing with the (BASELINE_OPT).

5.3.5 Discussion

Approximately 21% of unlabeled sentences were weakly labeled by Snorkel. The number of weak labels reflected to some extent the distribution of labels in the original dataset. Many sentences were weakly labeled with common labels (e.g., Outcomes (6a)). On the other hand, Snorkel failed to weakly label any sentences with the two least frequent labels (Table 5.4). The quality of Snorkel labels depends largely on the quality of LFs. We used two LFs based on heuristics explored in previous work. Micro-F_1 for both methods were

Table 5.5: Classification results using the original human annotated data set and weakly supervised data. SNORKEL(500) uses BASE-LINE OPT with additional 500 instances per label from Snorkel data. 3a: Trial Design; 3b: Changes to Trial Design; 4a: Eligibility Criteria; 4b: Data Collection Setting; 5: Interventions; 6a: Outcomes; 6b: Changes to Outcomes; 7a: Sample Size Determination; 7b: Interim Analyses/Stopping Guidelines; 8a: Sequence Generation; 8b: Randomization Type; 9: Allocation Concealment; 10: Randomization Implementation; 11a: Blinding Procedure; 11b: Similarity of Interventions; 12a: Statistical Methods for Outcomes; 12b: Statistical Methods for Other Analyses. P: precision; R: recall; F: F1 score; CI: confidence interval; AUC: Area Under ROC Curve.

CONSORT Item	**BASELINE F1 [CI]**	**BASELINE-OPT F1 [CI]**	**Snorkel (500) F1 [CI]**
3a	0.63 [0.46, 0.80]	0.82 [0.69, 0.95]	0.75 [0.63, 0.88]
3b	0.00 [0.00, 0 00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
4a	0.85 [0.76, 0.95]	0.89 [0.82, 0.96]	0.88. [0.82, 0.94]
4b	0.36 [0.06, 0.65]	0.87 [0.74, 1.00]	0.79 [0.61, 0.97]
5	0.72 [0.66, 0.78]	0.75 [0.68, 0.81]	0.73 [0.66, 0.81]
6a	0.81 [0.74, 0.88]	0.82 [0.75, 0.89]	0.83 [0.72, 0.87]
6b	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
7a	0.84 [0.76, 0.92]	0.88 [0.87, 0.90]	0.90 [0.86, 0.94]
7b	0.00 [0.00, 0.00]	0.70 [0.47, 0.94]	0.70 [0.17, 1.22]
8a	0.38 [0.15, 0.60]	0.88 [0.77, 1.00]	0.86 [0.60, 0.91]
8b	0.38 [0.10, 0.67]	0.73 [0.53, 0.93]	0.67 [0.51, 0.83]
9	0.00 [0.00, 0.00]	0.45 [0.35, 0.54]	0.40 [0.03, 0.76]
10	0.24 [0.05, 0.43]	0.53 [0.36, 0.71]	0.50 [0.22, 0.77]
11a	0.42 [0.12, 0.71]	0.66 [0.59, 0.74]	0.59 [0.46, 0.72]
11b	0.00 [0.00, 0.00]	0.45 [0.06, 0.85]	0.41 [0.04, 0.77]
12a	0.75 [0.69, 0.81]	0.77 [0.69, 0.85]	0.78 [0.69, 0.87]
12b	0.04 [-0.06, 0.14]	0.32 [0.27, 0.38]	0.24 [0.07, 0.40]
Micro-average	**0.72 [0.66, 0.76]**	**0.77 [0.71, 0.84]**	**0.76 [0.69, 0.82]**
Macro-average	**0.38 [0.34, 0.41]**	**0.62 [0.55, 0.69]**	**0.58 [0.48, 0.68]**
AUC	**0.812**	**0.876**	**0.875**

found to be around 0.50 in previous work (0.50 for keyword-based and 0.45 for section header-based). More accurate LFs could improve Snorkel results.

To better understand the quality of Snorkel-generated weak labels, we sampled 318 sentences, I and another researcher independently labeled the sentences, without access to Snorkel labels. We calculated the agreement of these annotations with Snorkel-generated labels, using Krippendorff's α with the distance metric MASI which accounts for partial agreement in the case of multiple labels. Krippendorff's α agreements between Snorkel and each annotator were found to be 0.46 and 0.61, respectively. While the inter-annotator agreement between the annotators was 0.59. Interestingly, agreement between Snorkel and simple majority vote was 0.93. These results suggest that Snorkel may converge to this simple heuristic in some cases, and that it behaves more or less like another annotator in the process.

We found that a large percentage of annotator disagreement with Snorkel came from randomization-related labels (items 8a, 8b, 9, and 10). These items often appear in the same sentence and the clues for them can be overlapping, making it a challenge to label them accurately for both humans and automated methods. Snorkel tends to pick a single label for sentences, and this was especially problematic for randomization-related sentences.

We did not observe significant improvements in classification performance due to weakly supervised data, which did not led to any correct predictions for the two least frequent labels (3a, 6a).

In addition, somewhat to our surprise, we found that model hyperparameters made a much more significant difference in model performance. BASELINE_OPT model yielded about 5% improvement in micro-F_1 and 36% improvement in macro-F_1 over the BASELINE model, with improvements in almost all labels. To assess how hyperparameters interacted with weak supervision, we also measured performance when BASELINE model (instead of BASELINE_OPT) was trained with weakly supervised data. Using Snorkel for weak supervision in this scenario improved micro-F_1 from 0.72 to 0.75, suggesting that hyperparameter optimization may, in some cases, obviate the need for additional (noisy) data.

Our investigation was limited to one relatively small corpus. The findings regarding weak supervision may not be generalizable to other corpora. We used few heuristics with modest performance as LFs and Snorkel label quality is likely to be improved with with additional more accurate LFs; however, this requires significant domain expertise. While we performed some hyperparameter tuning, we did not do an exhaustive search, and it is possible that more optimal hyperparameters can improve results further.

In summary, we demonstrated weak supervision approach to automatically label data that could be used for machine learning classification models. Even though the approach does not show significant improvements in term of model performances, we believe that certain further enhancements can be done. For example, position of a sentence in a particular section could be used as a potential labeling function. We found that the first sentence of Methods section often time is the sentence that describes overall study design of the trial, which is item 3a in the CONSORT checklist. Or sentences that describe sample size calculation often appear

by the end of the Methods section.

5.4 Summary of the chapter

In this chapter, we present a preliminary study to use ML and NLP to develop classification models that map sentences from full-text RCT papers to information items recommended in the CONSORT checklist. This work demonstrates a potential direction to apply NLP in supporting researchers to assess reporting quality of clinical research. Instead of reading and checking full text clinical publications manually, researchers (including journal editors, reviewers, meta researchers, and other stake holders) can use such tool to automatically check the appearances of certain information items recommended by reporting guidelines in the full text, so as to assess reporting quality of the research accordingly. Even though, our preliminary classification model has limitations and has not yet achieved practically useful results (especially for those items that have rare data), our experiment shows that it is possible to develop and improve the models with more data and perhaps with other NLP and ML techniques such as data augmentation, few-shot learning, prompt learning. The output of this work could be useful for our next development, which is an information extraction system that extracts relevant mention-level information. We can use this classification model to pre-filter sentences that contain fine-grained methodological information to be extracted, especially sentences belonging to those CONSORT items that achieved high accuracy (e.g. Eligibility Criteria, Sample Size Determination). In the long term, the principles learned with CONSORT can also be applied to annotating corpora targeting other reporting guidelines, such as STROBE for observational studies.

Chapter 6

Automatic extraction of methodological characteristics from RCT publications

In Chapter 5, we presented an NLP approach to extract information items at sentence level from full-text randomized controlled trials (RCTs) that can be used for reporting quality assessment. In this chapter, we will focus on methodological quality and present an NLP approach to extract methodological characteristics at the most fine-grained level– term level, to support EQA. In particular, we used NLP and ML to develop a named entity recognition (NER) model that automatically extracts methodological characteristics from full-text RCT publications. The high level implementation of such a model shown in the diagram below (Figure 6.1), in which, the model takes sentences from RCT publications as input, and extracts mentions that correspond to different methodological characteristics of the studies as outputs. Outputs of such a model will be helpful since methodological details at fine-grained level can be stored in a structured representation, in which information are organized in key-value pairs so as can be retrieved, queried and reasoned with. Some parts of the content of this chapter are based on a conference publication at AMIA Annual Symposium 2022 [221].

6.1 Why is fine-grained information needed?

The developments described in Chapter 4 and Chapter 5 look at information that is used for EQA at a coarse granularity, document level and sentence level respectively. NLP models such as the one that we developed in Chapter 5 help end users to check whether an information item is reported or not. However, it is not

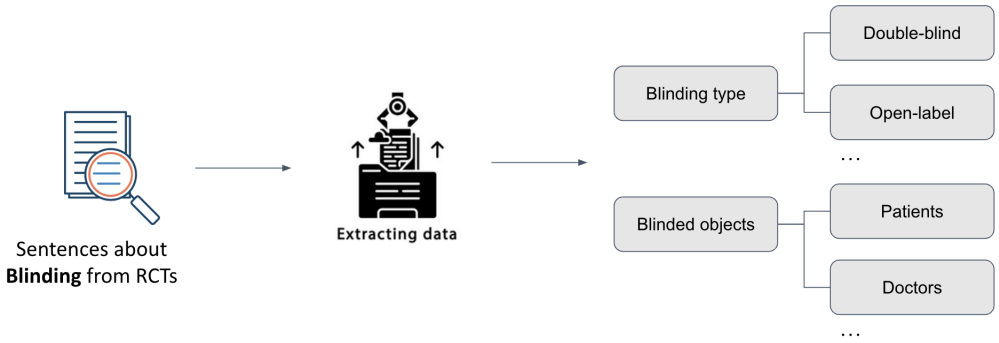

Figure 6.1: Information Extraction model from RCTs

sufficient to determine the rigor and robustness (i.e. methodological quality) of the study. Sometimes authors acknowledge their methodological issues as limitations. However, methodological quality issues can be implicit and hard to identify as they require domain knowledge and further analysis. For this, the sentences need to be analyzed at finer granularity, and different information items need to be cross-checked to identify potential pitfalls, weaknesses, and inconsistencies.

One of the most common methodological pitfalls in clinical research is the use of ***small sample size***. The purpose of estimating the appropriate sample size is to produce studies capable of detecting clinically relevant differences. Power is the probability of rejecting the null hypothesis when the null hypothesis is false. A Type II error can occur if there is not enough power in statistical tests, often resulting from sample sizes that are too small. In an empirical analysis of 48 neuroscience studies, Button et al. estimated the median statistical power of these studies is between 8% and 31%. Such low statistical power numbers shows serious consequences in overestimates of effect size and low reproducibility of results [222]. For this reason, for reviewers to examine the quality of a clinical study, looking at the sample size reported in the clinical publication is not sufficient. More importantly, reviewers need to be able to examine the validity of the sample size calculation. This task requires them to look at fine-grained information including power and other relevant statistical factors, such as statistical significance value, dropout rate, to replicate the calculation, similar to what Button et al. did in their studies. In such use case, automatic extraction of fine-grained information items from Sample Size sentences (as outputs of CONSORT classifier in Chapter 5) would be helpful.

To illustrate how methodological weaknesses can be identified by capturing fine-grained information, take the following example. In the process of conducting a clinical trial, ***Blinding*** is the act of masking the nature of the treatment from not only participants. Insufficient blinding of persons involved in RCTs is associated with performance bias [223]. Often time, blinding method is reported in a clinical publication. However, to assess the quality of a clinical study, further information might be needed. Considering an open-label RCT which reported in its full-text publication that *"Participants, investigators, or other medical or nursing staffs*

was aware of study group assignments during the trial": in terms of reporting quality, sentence classification models such as the CONSORT classifier in chapter 4 could label this sentence as item 11a, and indicate that reporting quality is good since the blinding strategy was explicitly provided. However, in terms of methodological quality, it is problematic because there is no blinding of patients or care providers which indicates potential risk of bias. In this case, extracting specific blinding type (open-label/no blinding) would allow reviewers to more easily assess methodological quality.

In the same ***Blinding*** topic, let's look at another example of how extracting fine-grained information could help to examine the methodological consistency. Considering a RCT study which claimed to be "double blind" in its title; however, in its Methods section, it said *"Only participants were blinded from their treatment allocation. Clinicians and assessors were informed of the allocated treatment."*. In this sentence, only one blinded object was mentioned (Participants). Thus, readers can interpret this as "single blind" method, and point out the inconsistency between what was reported in the study title vs. Methods section. Without capturing fine-grained information, in this case are Blinding Type (double blind) and Blinded Objects (participant), reviewers will not be able to flag the inconsistency accordingly. In fact, Saltaji et al. in a study that quantified the extent of bias associated with different blinding methods in RCTs, suggested that detailed information of blinding strategy, not only the blinding type (e.g. double-blind or single-blind) but also the corresponding attributes such as blinded objects (patients, doctors, and/or data analysts, etc.) should be all considered when assessing methodological quality of a study [224]. By capturing fine-grained information of blinding types and blinding objects, and represent it in a structure format, reviewers will not only be able to retrieve the information faster (e.g. find clinical publications that look at drug A for disease B where care provider is blinded); but also reason to detect inconsistencies by cross-checking the information, and also make this information machine-readable (allowing computational reasoning down the line).

Extracting fine-grained methodological information in this manner also offers benefits for researchers to compare methodological quality between clinical studies. For example, considering different methods of ***Allocation Concealment***, which is a technique of ensuring that implementation of the random allocation sequence occurs without knowledge of which patient will receive which treatment during trial conduction [72]. Good methods of generating a random allocation sequence include using a "central randomization", in which the individual recruiting the patient contacts a center by phone or secure computer after the patient is enrolled; or a "Sequentially numbered, opaque, sealed envelopes" in which trial operators use sequentially numbered, opaque sealed envelopes to perform randomization concealment [225]. Even though both methods are widely used, it is also acknowledged that the latter method has a higher potential risk of bias due to its vulnerability to manipulation. Therefore, identifying exactly which method is used for allocation concealment (instead of just identifying that there is an allocation concealment process in place), would be informative for quality assessment.

Extraction of such fine-grained information would also enable semantic searching of the literature based

on methodological quality of the articles. Retrieval of relevant biomedical scientific publications is essential directly to researchers in search of specific information, as well as to a range of downstream tasks, including technologically assisted reviews and question answering. Semantic indexing and searching thus has been one of the main focuses of researchers in the biomedical domain especially due to the tremendous amount of research published every year. While search engines used for biomedical literature incorporate some semantic features (e.g., MESH expansion of terms, Publication type search), semantic search based specifically on methodological characteristics remains under-explored. In a review of web tools for searching biomedical literature, only five systems aim to analyze search results and present summarized knowledge of semantics (biomedical concepts and their relationships) based on information extraction techniques [226]. The idea of using fine-grained semantic indexing of biomedical literature, beyond the descriptors of MeSH, at the semantic level of corresponding concepts just has been newly explored [215]. At a fine-grained level, researchers will be able to perform semantic search with complex queries based on methodological characteristics of the studies. For example, one could perform a semantic search query to retrieve diabetes studies in which the participants have been masked to the treatment.

To my best knowledge, currently there are no automated approaches to comprehensively capture methodological information at a granular level to support quality assessment and information retrieval. In this chapter, I will present our development of such a model, which can not only support systematic reviewers, journal editors, meta researchers to speed up the process of assessing quality of clinical studies, but also support semantic search of biomedical literature.

6.2 Overview of methods

Continuing the work that has been developed in Chapter 5, the work in this chapter focus on RCT studies only. There are four major stages of the work:

- **Design an methodological quality assessment data model**: In this stage, we proposed and developed a data model that captures the relevant RCT methodological characteristics that can be used for EQA. In particular, first we reviewed existing data models that represent RCT studies in the literature. Then we consolidated information items from these existing models that can be used for quality assessment. Lastly, we characterized each of the information with fine-grained attributes and sub-categories (if there are any) and represented them into a structured semantic representation. This representation is eventually used as the guideline for our annotation study and also the backbone of information extraction model later on.

- **Preliminary Annotation Study**: In this stage, we conducted an annotation study in which a set of RCT publications are manually annotated based on the information items in the data model that we developed. We designed annotation guidelines, conducted training for annotators, measured and

analyzed the inter-agreement between the annotators accordingly.

- **Information extraction model development**: In this stage, we implemented several Information Extraction models using different NLP and ML approaches, including: token-based classification, sequence labeling and rule-based.
- **User Study**: In this stage, we conducted an user study with a potential end user. The goal is for the user to use the model output and to evaluate if the prediction results from the system is correct or incorrect, and whether the model output can be helpful for the end user in downstream quality assessment tasks.

6.3 Data model development

6.3.1 Existing representations of Randomized Controlled Trials

Representations to describe components of a RCT and guidelines of how to report one have been developed [227]–[229]. In this section, we will review the existing data models that represent RCT study design and discuss how we leverage them to design our data model that can be used specifically for EQA.

The most popular representation of RCTs is the PICO model which is widely used by the systematic review community as well as evidence-based medicine, more broadly [227]. The PICO model defines the four most important information that defines an RCT study, including:

- Patients (also called Participants or Population) refers to the description of the study's patients;
- Intervention refers to the main treatment that is being used in the study;
- Comparison refers to the alternative treatment that is compared to the main treatment;
- Outcomes refers to the characteristics that are measured to determine the effect of interventions.

The PICO model provides a minimum structure that describes basic information about an RCT. Therefore, reviewers often start with identifying information related to PICO in order to understand the topic of a clinical study (what is it about). Corresponding to this model, Cochrane Collaboration has created PICO ontology that defines more fine-grained elements belonging to the high level PICO components and relations between them [227]. In the PICO ontology, each information item is defined as an entity, and relations between entities are defined as well. For example, "Population" entity has the following attributes: age, gender, condition, inclusion criteria and exclusion criteria; "Intervention" entity has relations with the following attributes: dose, schedule, duration, settings. Nevertheless, it is important to keep in mind that the PICO ontology was originally designed to model the questions asked and answered in Cochrane's systematic reviews [227]. While PICO-related information is helpful for researchers/systematic reviewers to understand the topics

and characteristics of a clinical study, they are less informative for quality assessment. The class from the PICO ontology that we found potentially relevant to quality assessment is Settings of an intervention (e.g. multicenter or single center, and locations). A study suggested that Setting information should be considered when the results of RCTs and meta-analyses are interpreted since it might have a certain effect on the overall quality of the studies [14].

Different from the PICO ontology, RCT schema developed by Sim et al. in 2004 with the same goal of representing RCT, consists of a hierarchy and is not limited to the PICO framework [228]. In particular, the RCT schema identifies RCT methodology components based on the actual tasks that need to be executed, from the top-level target tasks to the decomposed sub-tasks and the methods by which each sub-task is to be accomplished. For example a top-level target task could be assessing RCT validity, and the corresponding first level sub-tasks are judging internal validity and external validity. Depending on the tasks, a list of questions to fulfill the tasks were then identified, such as to judge the internal validity, questions could be *"Was the statistical design of the trial appropriate? Were the intervention groups comparable? Was there any intervention assignment bias?"*. Finally, based on the list of tasks and questions, Sim et al. developed RCT Schema that captures details of a RCT in regards of administration, design, execution, and results needed to answer the questions and fulfill the tasks. According to the authors, the RCT schema contains 147 unique information items organized in hierarchical order corresponding to the tasks and questions from high level information items (e.g. Participants, Interventions, Outcomes) to some of more complex information items (e.g. description of the outcome measurements, rate outcomes need a denominator, cost outcomes need a discount rate).

In a subsequent research, Sim et al. developed a more comprehensive ontology that represents clinical research in general called The Ontology of Clinical Research (OCRe) [229]. In this work, instead of focusing on RCT only, the OCRe ontology contains other study designs such as Observational study design. Each study design is a class in the data model, as shown in Figure 6.2. Under the *Interventional* study design, fine-grained study designs are also defined, depending on the patient-treatment allocation methods, which include: parallel group, cross over, single group, and N-1 crossover.

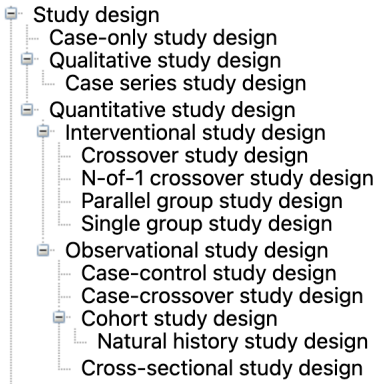

Figure 6.2: List of study designs captured in OCRe ontology

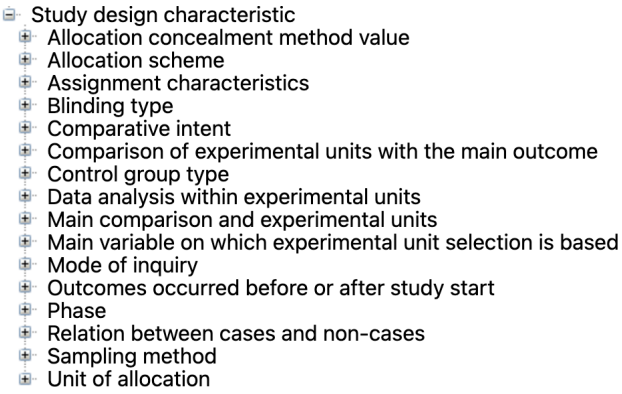

Figure 6.3: Methodological characteristics of RCT captured in OCRe ontology

Corresponding to these study designs, the ontology also defines a list of study characteristics as shown in Figure 6.3. Many of these characteristics are relevant for methodological quality assessment. For example: "Allocation concealment method value" helps us to understand the method used to allocate patients to treatments, yet at the same time to preserve the randomness of the allocation in the study; "blinding type" helps us to understand the method used to masking the treatment to the study personnel and also the objects who are blinded; and "allocation scheme" (which includes random allocation) provides information about different types of randomization.

On the other hand, based on the work presented in Chapter 5, we know that CONSORT reporting guideline has been successfully used to improve reporting quality and transparency of RCT publications. The checklist contains some methodological information items (e.g., study design, blinding, randomization, statistical methods). Therefore, conceptually, we also can use the CONSORT checklist as another resource to get the list of characteristics that can be used for EQA.

6.3.2 Data Model Development

The process of designing and developing data model is a iterative exercise of (1) reviewing existing data models that represent RCTs (such as PICO, OCRe, CONSORT), (2) getting a list of initial methodological characteristics that can be used for EQA purposes, (3) checking in the literature for additional information, and (4) piloting an annotation study to confirm the data model components. In particular, we largely used OCRe Ontology and the CONSORT checklist as the two main resources to collect the initial list of information items[1]. After that, we also conducted a preliminary annotation study to examine the initial components that are identified in the preliminary model and discover more information items that are relevant to be added into the data model. After the preliminary data model was developed, we annotated several articles to assess the feasibility of annotating the items in the model. As a result, the data model was also refined.

[1]PICO also provides some methodological characteristics that can be used for EQA such as Settings and Location. However, since both of these information are also covered in the CONSORT checklist, we do not use any information items from PICO directly

In OCRe, we primarily focused on sub-classes of Interventional Study Design and Study Design Characteristic classes, including Blinding Type and Randomization Type. Some relevant characteristics were derived from data properties (e.g., Planned Sample Size and Actual Sample Size). Additional characteristics relevant for methodological quality were drawn from the CONSORT methodology checklist. For example, fine-grained information related to Sample Size Calculation (CONSORT item 7a) and Trial Settings (4b) were included, such as Power and Alpha values, and Multicenter vs. Single-center distinction. The main criterion for inclusion was whether the characteristic provides any information about methodological quality, which we ensured through literature review, and whether they can be identified in RCT publications. For example, through literature review, we affirmed the reason for the additional Settings characteristics. Single-center trials are available only at the study creator's hospital, while multi-center trials are conducted at various locations and offer diversity advantages and lower risk of bias over single-center trials [230]. For sample size calculation characteristics, knowing the sample size (number of patients) is not sufficient for quality assessment purpose. More importantly, understanding how the sample size estimation was calculated allows reviewers to examine statistical quality of trials (e.g. if targeted sample size is large enough to statistically detect different effects between treatments)[231]. We also noted that some characteristics have properties whose values can be important in interpreting them (e.g., Block Size for Block Randomization). These properties were included in the data model, as well. In the end, overall, each information item is identified based on the following conceptual questions:

- What is the definition of the information item?
- What is the corresponding information item in the OCRe Ontology?
- What is the corresponding information item in the CONSORT checklist?
- What attributes does the information item have?
- What are the subcategories of the information item?

6.3.3 Data Model

Our final data model contains main seven domains: Trial Design, Blinding, Randomization, Allocation Concealment, Settings, Sample Size and Sample Size Calculation. In each domain, there are in total 19 top-level characteristics (note that the Allocation Concealment domain has no top-level characteristics). For each of the characteristics, we defined their sub-types and properties relevant to the sub-types accordingly. The resulting data model is provided in Figure 6.4.

We provided the descriptions of the seven domain captured in our data model below. Detail descriptions the 19 top level characteristics, their subcategories and attributes are provided in the Appendix A.

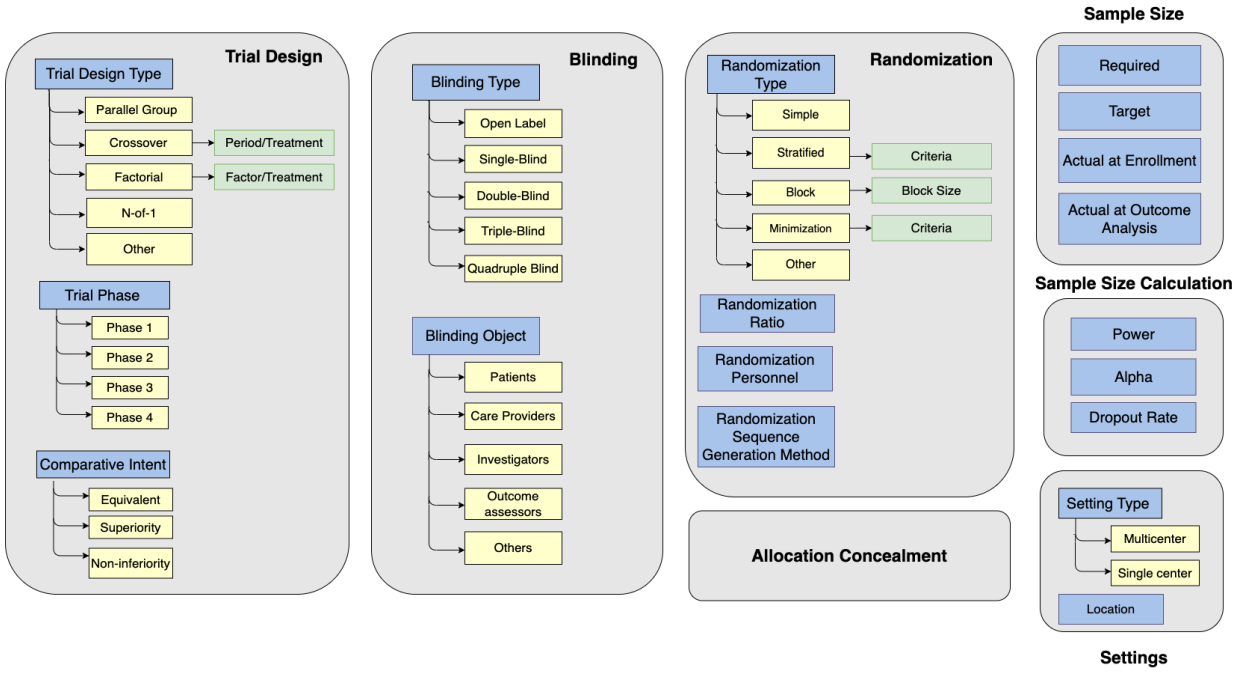

Figure 6.4: Our proposed data model to capture methodological characteristics from RCT publications. Domains (Trial Design, Blinding, etc.) are shown in gray boxes. Top-level characteristics are shown in blue rectangles. Their subtypes are shown in yellow, and properties relevant to the subtypes are shown in green.

Table 6.1: Methodological items in Trial Design domain

Top level domain	**Trial Design**
Definition	Contains methodological characteristics that are related to the design of the trial
Corresponding CONSORT item	Study Design (3a) - Description of trial design (such as parallel, factorial) including allocation ratio
Corresponding OCRe characteristic	Study Design (http://purl.org/net/OCRe/study design.owl#OCRE100056)
Top level methodological characteristics	Trial Design Type: a categorical information item which refers to how participants are assigned into different treatment groups (e.g. parallel-group, cross-over, factorial). Trial Phase: describes the level of a trial required of drugs before (and after) they are routinely used in clinical practice (e.g. Phase 1, Phase 2). Comparative Intent: refers to the intent of comparison made in a study with two or more interventions (e.g. non-inferiority, superiority).

Table 6.2: Methodological items in Blinding domain

Top level domain	**Blinding**
Definition	Blinding or masking is the process of keeping the study group assignment hidden after allocation, which is commonly used to reduce the risk of bias in clinical trials with two or more study groups.
Corresponding CONSORT item	Blinding (11a) - If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how.
Corresponding OCRe characteristic	Blinding Type (http://purl.org/net/OCRe/OCRe.owl#OCRE574000)
Top level methodological characteristics	Blinding Method : refers to the act of masking the nature of the treatment from parties involved such as patients, doctors, statisticians, etc. Blinding Objects: refers to who are the people that were blinded.

Table 6.3: Methodological items in Randomization domain

Top level domain	**Randomization**
Definition	Refers to the sequence by which participants will be allocated to the study groups. This practice is meant to keep researchers and participants unaware of the sequence, with the goal of preventing the researchers from (unconsciously or consciously) influencing the group assignment of study participants.
Corresponding CONSORT item	Randomization Sequence Generation (8a) - Method used to generate the random allocation sequence. Randomization Sequence Generation (8b) - Type of randomization; details of any restriction (such as blocking and block size). Randomization Implementation (10) -Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions.
Corresponding OCRe characteristic	Random Allocation (http://purl.org/net/OCRe/study_design.owl#:OCRE100033)
Top level methodological characteristics	Randomization Type: what type of randomization that describes how patients are assigned into different treatment groups (e.g. block randomization, stratification randomization). Randomization Ratio: Ratio of randomization into treatment groups. This attribute is not tied to any particular type of randomization. Randomization Sequence Generation Method: How the randomized sequence is generated (e.g. using a computer random number generator; random number table; coin tossing; shuffling cards or envelopes; throwing dice). Randomization Personnel: refers to the person, people, organization, who is involved in creating/generating the randomization sequence.

Table 6.4: Methodological items in Allocation Concealment domain

Top level domain	**Allocation Concealment**
Definition	Allocation concealment is performed when the treatment allocation system is set up so that the person enrolling participants does not know in advance which treatment the next person will get. Different from blinding, allocation concealment ensures that the treatment to be allocated is not known before the patient is entered into the study. Blinding ensures that the patient/physician is blinded to the treatment allocation after enrollment into the study.
Corresponding CONSORT item	Allocation Concealment (9) - Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned.
Corresponding OCRe characteristic	Allocation Concealment Method Value (http://purl.org/net/OCRe/OCRe.owlOCRE825000)
Top level methodological characteristics	Allocation Concealment Methods

Table 6.5: Methodological items in Sample Size domain

Top level domain	**Sample Size**
Definition	Refers to the number of patients based on the required sample size calculation, or actual sample size of the trial.
Corresponding CONSORT item	Sample Size Calculation (7a) - How sample size was determined (e.g. what is the number of required sample size). Participants Flow (13a) - For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome.
Corresponding OCRe characteristic	has required sample size (http://purl.org/net/OCRe/OCRe.owlOCRE855955) has planned sample size (http://purl.org/net/OCRe/OCRe.owlOCRE900203) has actual sample size (http://purl.org/net/OCRe/OCRe.owlOCRE900200)
Top level methodological characteristics	Required Sample Size: The number of patients based on the required sample size calculation. Target Sample Size: The target number of patients based on the required sample size. Actual Sample Size at Enrollment: The number of patients who actually enrolled in the study at the beginning of the study. Actual Sample Size at Outcome Analysis: The number of patients who actually completed the study and collected data for analysis.

Table 6.6: Methodological items in Sample Size Calculation domain

High level domain	**Sample Size Calculation**
Definition	Refers to the statistical values used to calculate required sample size of a trial.
Corresponding CONSORT item	Sample Size Calculation (7a) - How sample size was determined
Corresponding OCRe characteristic	has power calculation (http://purl.org/net/OCRe/OCRe.owlOCRE900204)
Top level methodological characteristics	Power Value: What is the power value used to calculate required sample size. Alpha Value: What is the alpha value used to calculate required sample size. Sometimes, this value is also provided as a significance level or p-value. Drop Out Rate Value: The sample size estimation formula will provide a number of evaluated subjects required for achieving desired statistical significance for a given hypothesis. However in practice we may need to enroll more subjects to account for potential dropouts.

Table 6.7: Methodological items in Settings domain

High level domain	**Settings**
Definition	Settings and location of the study
Corresponding CONSORT item	Participants (4b) - Settings and locations where the data were collected.
Corresponding OCRe characteristic	N/A
Top level methodological characteristics	Settings Type: what is the setting of the study? Choose between two values: Single center or Multi center Location: Describes the location of the study. This information should be city, country, area names.

6.4 Annotation Study

6.4.1 Annotation process and guideline

The annotation study was conducted through two phases. Phase one is a preliminary annotation study as mentioned above to piloting the annotation process as well as refining the data model. In phase two, we annotated 150 articles based on the refined model. In this section, I describe the annotation study from data collection, annotator training and inter-annnotator agreement calculation to annotation reconciliation and ground truth generation.

Collect data

For annotation, we collected a set of RCT publications from PubMed Central Open Access Subset. 25 publications came from the CONSORT-TM corpus [194]. We collected another set of 125 articles by issuing a search query that limited by the publication type "Randomized Controlled Trials" and full-text availability[2]. We eliminated publications reporting study protocols or multiple RCT studies from the search results. We then also filtered and collected only the RCTs that have ClinicalTrials.gov identification numbers. ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world maintained by the U.S. National Library of Medicine. Before conducting trials, clinical teams are required to register their protocols which include details of the study designs and methodology into the database. By limiting to the RCTs that have ClinicalTrials.gov registrations, we can potentially use methodological information registered in the database to evaluate our information extraction results. From the remaining articles, we randomly selected 125 articles, for a total of 150 articles.

Annotation environment setup

Three annotators conducted the annotation, two PhD students and a faculty member with experience in biomedical literature and annotation. As for annotation tool, we used the same annotation tool, Brat[3] [232]. The full-text of RCT publications, including their titles and abstracts, were imported into Brat. Figure 6.5 shows a fragment of a RCT publication annotated using Brat. The system was set up on a web server, data was uploaded for each user separately and information items defined within the system as shown in Figure 6.6.

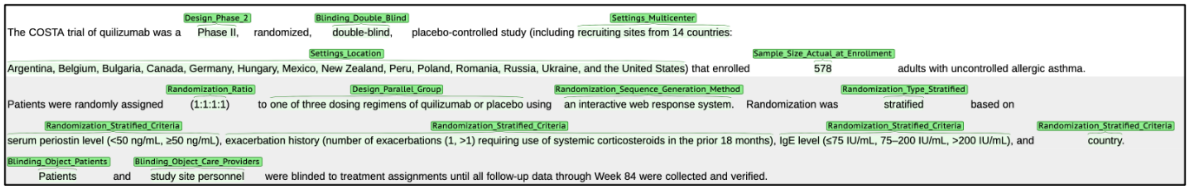

Figure 6.5: Annotation example on brat interface.

[2]PubMed search query: "randomized controlled trial"[Publication Type]) AND (fft[Filter])
[3]https://brat.nlplab.org/

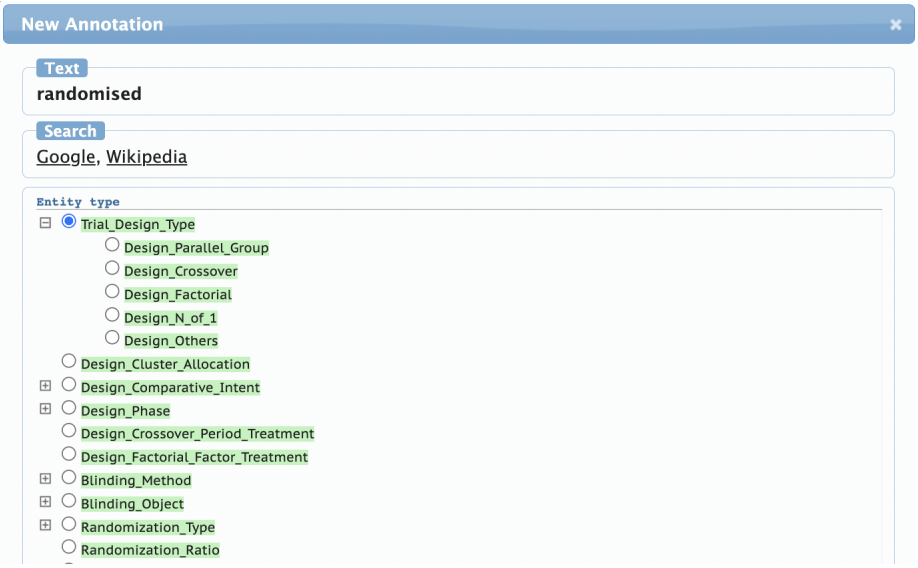

Figure 6.6: Example of information items in our data model for annotation on Brat.

Annotation guidelines

The annotation guidelines were created in an iterative process. A initial draft was created, containing definition of information items, general rules that show how to annotate them, examples of typical annotations. In the guideline, information items are divided into two types: categorical items or free-text items. For categorical items, based on the text in the article, the annotator needs to decide which sub-category the information item belongs to and annotate the text span that supports the judgments. For free-text items, there are no predefined categorical values to choose, the annotators highlight the text span that describes the information item in the text. All three annotators annotated the first 10 articles. Disagreements and inconsistencies were discussed and resolved. The annotation guideline was refined accordingly. We also added examples of edge cases and solutions how to annotate them. The final annotation guideline is provided in the Appendix A. After the annotation guideline was finalized and all annotators agreed on the annotation rules and process, three annotators continued to annotate the next 20 articles, which were used to calculate inter-annotator agreements. In the final annotation round, two annotators with the highest agreement in the previous round individually labeled 70 and 50 articles each.

Annotation process and reconciliation

We adopted a minimal annotation approach, focusing on annotating the shortest meaningful text spans for a given item, often a clause. The categories that have sub-classes (e.g., Patient, Investigator, etc. for Blinding Object) were annotated at the finest granularity justified by the text span. We focused primarily on abstracts and Methods sections, as they were most likely to contain methodological information. Some categories, particularly those related to Sample Size, were also annotated in the Results sections, where they were often reported. During the annotation process, the annotators were instructed to annotate a mention

only once for a particular characteristic in a given section, although different mentions corresponding to the same concept (e.g., *no blinding* and *open label* for the Open Label category) were expected to be annotated. This helped reduce annotation burden while generating a diverse set of examples.

For example, consider the following sentence:

"The study is a multicenter, randomized, open, parallel group trial conducted at 33 centers in four European countries (France, Germany, Italy, and Spain) with a target enrollment of 153 patients accessible for the primary end point analysis in each group."

Below is how the sentence was annotated.

"The study is a ***multicenter [multicenter]****, randomized,* ***open [blinding type]****,* ***parallel-group [trial design type]*** *trial conducted at* ***33 centers [multicenter]*** *in four European countries (****France, Germany, Italy, and Spain [locations]****) with a target enrollment of* ***153 [targeted sample size]*** *patients accessible for the primary end point analysis in each group."*

We then reconciled annotations from the annotators into one single final data set that can be used for model development through several different approaches. For the first set of 10 articles which were annotated by all three annotators, one annotator checked and finalized them into a single set. For the set of 50 articles which were annotated by an individual annotator, the other annotator checked and finalized the annotations. For the whole data set, as methodological information that we model appears in different places throughout a RCT publication (e.g. "double-blind" not only appears in sentences that describe blinding methods, but also ones that describe study design), we generated additional annotations by automatically labeling all instances of the mentions that already appear in the same section of the document in the manually annotated set. Some of these automatic annotations were incorrect, we manually removed them (e.g., not all instances of the mention *blind* are about the blinding type of the study). Automatic annotation helped us increase the number of examples in the data set without significantly increasing annotator burden.

Inter-annotator agreement

We calculated inter-annotator agreement at the span and document levels. For span level agreement, we used exact match for all categories considered. We used F_1 score for span level agreement, considering annotations from one annotator to be the ground truth and those from the other as predictions [233]. Document level agreement was calculated for items with subcategories: Trial Design Type, Phase, Comparative Intent, Blinding Type, Randomization Type, and Setting. In this case, we examined whether two annotators agreed on whether the publication reported a particular study characteristic (e.g., Double-Blind as the Blinding Type). We used both Cohen's κ and F_1 score for inter-annotator agreement at the document level.

6.4.2 Annotation Study Results

We annotated a total of 150 RCT articles in this pilot study. Table 6.8 shows the descriptive statistics of the annotated corpus. Among the top level categories, Sample Size had the highest number of annotations (637) followed by Randomization (557), Trial Design (481), Blinding (442), and Settings (210). Allocation Concealment Method was rarely discussed (19 instances). At the fine-grained level, Parallel Group (258), Actual Sample Size (227), Double-Blind (191), and Multicenter Settings (188) were annotated most frequently. Although we represented some characteristics in the data model to maintain consistency with OCRe, we did not find any instances of these in the corpus: N-of-1, Factorial Factor/Treatment, Triple-Blind.

Table 6.8: Statistical information of the annotated corpus.

Statistic	Completed Corpus	Train Set	Test Set
Total number of articles	150	135	15
Total number of sentences	22,000	20,490	1,510
Total number of sentences with annotations	1417	1238	179
Total number of tokens	674,277	624,563	49,714
Total number of annotated tokens	9,199	7758	1,441
Total number of annotations	2724	2346	378

Table 6.9 shows pair-wise inter-annotator agreement results obtained on 20 articles at span and document levels using Cohen's κ and F_1 score. The results show overall high agreement. Cohen's κ scores indicate substantial to perfect agreement between the annotators (0.74-0.83). F_1 score agreement is over 0.9 in all cases. Overall, annotators 1 and 2 achieved higher agreement at both span and document levels. These two annotators annotated the last 120 articles.

Table 6.9: Pair-wise agreement at span and document levels. Document level agreement is calculated for categories with sub-classes only.

	Ann1 vs. Ann2		**Ann2 vs. Ann 3**		**Ann1 vs. Ann3**	
	Cohen's κ	**F_1**	**Cohen's κ**	**F_1**	**Cohen's κ**	**F_1**
Span level		0.94		0.90		0.90
Document level	0.83	0.95	0.74	0.92	0.79	0.93

Our annotation study showed that annotating RCT methodological items at the span level was feasible. We obtained high inter-annotator agreement, indicating that these characteristics can be more or less reliably annotated. Several items were challenging. For example, Parallel Group is easy to annotate when it is explicit (e.g., *parallel-group*). However, it is often implicit and can only be determined from the description of the intervention (e.g., *intravenous rhEPO 40 000 IU or placebo fortnightly*). While annotators were instructed to annotate such implicit cases in the annotation guidelines, their annotations were less consistent for these cases.

6.5 NER Model Development

We approach the task of identifying mentions of methodological items and their characteristics as a named entity recognition (NER) task. In our case, entities are the methodological characteristics that we defined in our data model as described above.

The NLP approaches that we applied to develop our NER models are: (1) supervised learning approaches: the model is built using traditional ML and deep learning algorithms; and (2) Rule-based approaches: the model is built based on hand-crafted rules. Since majority of our implementation is supervised learning approach using deep learning. Readers might need to refer back to Chapter 3 for the technical details of the methods used in this section.

6.5.1 Methods

Model setup

Machine learning-based models

As reviewed the most common methods for NER models development, in this work, we experimented with NER models based on current baseline neural network architectures using pretrained language model as context encoder, and Token classification and CRF as tag encoder. More specifically, we applied BIO tag scheme to represent token labels in sentences. Similar to the work in Chapter 4, we used PubMedBERT (*base-uncased-abstract-fulltext*) model [234] as the sentence encoder and experimented with two different classification layers: (1) a fully-connected token classification layer; and (2) a classification layer based on CRF. We used the prebuilt TokenClassification model from the huggingface library[4] to implement the token classification model; and a public BERT-CRF implementation[5] to implement the CRF-based model. Both models were developed with the same settings of hyper parameters: batch size of 4, Adam optimizer, learning rates of 1e-5, 2e-5, 3e-5, and 5e-5, and number of epochs of 10, 20, 30. For final training, we used the learning rate of 5e-5 for the token classification model and 3e-5 for the CRF-based model and 20 epochs for both models, which yielded the best performances.

Methodological information that we model generally occurs over a handful of sentences in a RCT publication. Including all sentences of the publication in training leads to a very imbalanced data set. To address this problem, we adopted four strategies to sample sentences for inclusion in training:

- *Positive sentences only*: Only sentences that include at least one annotated span are included.
- *Random sampling:* Positive sentences + a random sentence with no annotations (i.e., negative sentence) for each positive sentence

[4] https://huggingface.co/docs/transformers/tasks/token$_c lassification$
[5] https://github.com/Louis-udm/NER-BERT-CRF

- *Similarity sampling:* Positive sentences + negative sentence with the highest cosine similarity with positive sentence
- *Random+Similarity sampling:* Positive sentences + random sampling for half of the positive sentences + similarity sampling for the other half.

For cosine similarity calculation, we generated vector representations of the sentences using pretrained PubMedBERT embeddings. The dataset was split into training and test sets, 135 articles and 15 articles respectively[6].

Rule-based model

Along side with machine learning-based NER models, we also developed a rule-based model with human-defined hand-crafted rules. Rules were learnt from the training set which are essentially lexical patterns for each of the entities in our data model. An example of a lexical rule for Blinding information item is provided below:

```
# Blinding Method information item
# No examples of triple and quadruple blind methods
blinding_double_blind_patterns = ["double-blind", "double-masked", "double-blinded",
                            "double-mask", "double blind"]
blinding_single_blind_patterns = ["single-blind", "single-blinded", "assessor-blind",
                          "examiner-blind", "patient-blind"]
blinding_open_label_patterns = ["open-label", "were not masked", "were not blinded",
                                "aware of treatment"]
```

Besides lexical rules (based on key-words), we also defined syntactical rules for some information items if they are applicable. For example, "parallel group trial design" can be identified by not only the keyword "parallel group", but often time, via a text span that describes the treatments of the study. A syntactic pattern from many parallel-group design examples could be represented as "patients received treatment A vs. placebo". Based on this syntactic structure, we defined a syntactical rule accordingly as following:

```
# Parallel group
patterns = ["received", "were randomized to", "were located", "were randomised to"]
for pattern in patterns:
    results = re.search(pattern+'(.*)or placebo',sentence)
    if results:
        print (results.group(1))
```

[6]The set of 15 articles for testing was fixed from the very beginning of the project when we only had 35 articles for training at the first point we developed our model. We did not change the test set even the training set was expanded to keep the implementation consistent.

Model evaluation

We evaluated the two machine learning-based models: token classification and CRF-based at two levels:

- Span-level evaluation: which was done for each information item. It means B-tagged tokens and I-tagged tokens of the same information item should be considered together for a full predicted entity. We particularly focus on two metrics that take into account the entity type predictions: **Strict** and **Type** (discussed in Section 3.3.3). For each metric, we calculated their Precision, Recall and F1 following the formula that we provided in the NLP and ML background review.

- Document-level evaluation: some categorical information items in our data model could be evaluated at document level. For example, given an RCT, we may want to know whether or not a model is able to predict correctly type of the blinding method in which the trial used. For document level, we only calculated regular Precision, Recall and F1 score for categorical information items, including: Trial Design Type, Randomization Type, Blinding Type, Settings (multicenter vs. single center) and Phase.

As for the rule-based model, we evaluated it at document level only and compared the results with the best performing machine learning-based model.

6.5.2 Results for NER models

For each machine learning-based NER approach (token classification vs. CRF-based), we developed four models each corresponding to a sampling strategy for training: Positive Sentences, Random Sampling, Similarity Sampling, and Random+Similarity Sampling.

Table 6.10 shows the performances of the NER models at the span level. In both strict and partial evaluation, CRF-based classification using Similarity Sampling achieved the best F_1 scores. The results with token classification are consistently lower than CRF-based results. While using Positive Sentences only for training yields lowest F_1 results, its recall is among the highest. Sampling strategy has a more significant effect on precision than on recall.

Table 6.10: Model performances at the span level with four sampling strategies for training.

Sampling Strategy	**Token classification**						**CRF-based**					
	Strict			**Type**			**Strict**			**Type**		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Positive Sentences	0.24	0.58	0.34	0.30	0.71	0.42	0.28	0.62	0.39	0.35	0.77	0.48
Random Sampling	0.51	0.53	0.52	0.63	0.66	0.65	0.47	0.61	0.53	0.60	0.76	0.67
Similarity Sampling	0.51	0.53	0.52	0.64	0.66	0.65	0.52	0.77	0.54	0.66	0.73	**0.69**
Random + Similarity Sampling	0.49	0.47	0.48	0.63	0.61	0.62	0.49	0.58	0.53	0.61	0.72	0.66

We analyzed the results of the best-performing model (PubMedBERT with CRF layer trained with Similarity Sampling strategy) in more detail. These results, obtained with span level evaluation, are shown in

Table 6.11. The results show that the results vary widely among different characteristics. Some characteristics were recognized by the model relatively well, such as Power for sample size calculation and randomization Ratio (0.95 F_1 and 0.94 F_1 respectively), Stratification Criteria for stratified randomization cases (0.86 F_1), and sample size Alpha Value (0.78 F_1). Except Stratification Criteria, these characteristics are generally expressed in a small number of ways in publications, which may explain the higher performance. Another factor is that these characteristics are relatively frequent in the dataset. The model fails on several characteristics, such as Allocation Concealment Method and Period/Treatment for crossover design, which only had a few examples in the data set. In contrast, the model had more success with some other characteristics, which also had few examples, such as Comparative Intent. This can also be attributed to the fact that the expressions for these characteristics are less diverse than those for, say, Randomization Personnel, which include a wide range of expressions such as *individuals not associated with study conduct* or *separate unblinded statistical team.* For some items, strict vs. partial evaluation results are the same (e.g., Comparative Intent), while there is a significant different for others that involve numbers, which deserves further investigation (e.g., Block Size, randomization Ratio).

Table 6.11: Performances of the best model (CRF-based model trained with Similarity Sampling) at the span level. Characteristics with * next to their name are fine-grained items, while others have subtypes.For characteristics with subtypes, the results are aggregated for brevity. For example, Trial Design:Type results include predictions for Parallel Group, Factorial, etc. Similarly, Sample Size:Type aggregates the results for different sample size calculations: Required, Targeted, Actual at Enrollment, and Actual at Outcome Analysis.

Domain	Characteristics	Strict			Type		
		P	**R**	**F_1**	**P**	**R**	**F_1**
Trial Design	Type	0.35	0.41	0.38	0.63	0.72	0.67
	Phase	0.91	0.72	0.80	0.91	0.72	0.80
	Comparative Intent	0.50	0.87	0.63	0.50	0.87	0.63
	*Crossover Period/Treatment	0.67	1.00	0.80	0.67	1.00	0.80
	*Factorial Factor/Treatment	0.43	0.60	0.50	0.43	0.60	0.50
Blinding	Type	0.82	0.73	0.77	0.82	0.73	0.77
	Objects	0.40	0.50	0.44	0.43	0.54	0.48
Randomization	Type	0.75	0.32	0.45	0.75	0.32	0.45
	*Block Size	0.20	0.33	0.25	0.60	1.00	0.75
	*Minimization Criteria	0.44	0.50	0.47	0.56	0.63	0.59
	*Stratification Criteria	0.64	0.67	0.65	0.84	0.88	0.86
	*Personnel	0.07	0.20	0.10	0.27	0.80	0.40
	*Ratio	0.88	0.88	0.88	0.94	0.94	0.94
	*Sequence Generation	0.15	0.21	0.17	0.56	0.79	0.65
Sample Size	Type	0.54	0.72	0.62	0.63	0.85	0.72
	*Alpha	0.52	0.55	0.54	0.68	0.87	0.76
	*Dropout Rate	0.25	0.43	0.32	0.42	0.71	0.53
	*Power	0.91	1.00	0.95	0.91	1.00	0.95
Settings	Type	0.75	0.66	0.70	0.82	0.72	0.77
	*Location	0.50	0.50	0.50	0.77	0.77	0.77
Allocation Concealment	*Allocation Concealment Methods	0.00	0.00	0.00	0.00	0.00	0.00
OVERALL		0.52	0.57	0.54	0.66	0.73	0.69

We also evaluated the models at document level.

Table 6.12 shows the performances of the two machine learning-based NER models at the document level, which are largely consistent with the results at the span level. CRF-based models consistently outperform the token classification counterparts. Similarity Sampling yields highest F_1 score and precision performance, while its recall is lower than of the Positive Sentence sampling. Since we are ultimately interested in summarizing methodological characteristics of a study at the document level, we consider document-level evaluation results as the main results for this study.

Table 6.12: Document-level performances of four models using two different classification layers

Sampling Strategy	**Token classification**			**CRF-based**		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Positive Sentences	0.71	**0.83**	0.77	0.81	0.82	0.81
Random Sampling	0.82	0.76	0.78	0.87	0.71	0.78
Similarity Sampling	0.84	0.76	0.80	0.91	0.74	**0.82**
Random + Similarity Sampling	0.86	0.76	0.81	**0.92**	0.73	**0.82**

The document-level results with the best sampling strategy (Similarity Sampling) were also compared to the results of the rule-based method. This comparison is limited to the methodological characteristics in the data model with subtypes. We looked closer at the best performing models, the TokenClassification-based and CRF-based models using Similarity sampling method, and compared them to a rule-based model. The CRF-based model, once again, achieved the best performances. Table 6.13 shows a comparison between the three models, and breakdown of their performances on five categorical information items that we evaluated at document-level: Clinical Trial Design Type, Randomization Type, Blinding Type, Phase and Settings.

Table 6.13: Document-level performances of TokenClassification-based model vs. CRF-based model using Similarity sampling method, vs. rule-based model.

Categorical Information Item	**Token Classification**			**CRF**			**Rule-based**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Trial Design Type	0.83	0.71	0.77	1.00	0.79	0.88	0.69	0.73	0.71
Randomization Type	1.00	0.67	0.8	0.85	0.65	0.77	0.72	0.72	0.72
Blinding Type	0.93	1.00	0.97	0.93	0.93	0.93	0.86	0.86	0.86
Phase	0.5	0.90	0.64	0.5	0.91	0.65	0.50	1.00	0.64
Settings	0.90	0.82	0.86	1.00	0.79	0.88	0.77	0.77	0.77
OVERALL	0.84	0.76	0.80	0.91	0.74	0.82	0.73	0.77	0.75

6.5.3 Discussion of NER results

CRF-based models performed comparatively better than token classification models in NER, indicating that capturing label sequences is important for methodological IE. This is not surprising, since sentences where many methodological characteristics of the RCT are mentioned together are common (e.g., *This phase 2b, double-blind, placebo-controlled, parallel-group, dose-ranging randomized clinical trial...*) and capturing such

patterns may benefit the models. Since only a small number of sentences in each article was annotated, we sampled un-annotated sentences to increase the training set size. Similarity-based sampling yielded the best results overall, indicating that providing more difficult negative examples to the training procedure is beneficial.

We printed out predictions of the best performing PubMedBERT-CRF model and also rule-based model for data of the test set to compare differences between the two approaches. We found that machine learning-based model predicted exceptionally well for those information items that are more descriptive. One of the most popular information items that encounter this phenomenon is "Parallel Group Trial Design". According to our annotation guideline, there are edge cases in which the article does not explicitly mention "parallel-group" keyword (which is used to identify this information item in our rule-based model). However, it lists several treatment groups, then we can assume that it is a parallel design type. In such cases, annotators should highlight the names of the treatments as an annotation of parallel trial design. Accordingly, while it is very hard to define rules to capture such cases, the PubMedBERT-CRF model with its design to aware of the context between word sequences, can capture these cases very well. For example, considering a sentence from the test set: *"26 patients presenting with acute myocardial infarction randomised to receive an intravenous infusion of etanercept (10 mg) or saline placebo."*. Comparing with human annotation, the PubMedBERT model predicted correctly both "Sample Size Actual at Enrollment" information item ("26 patients") and the "Design Parallel Group" information item ("an intravenous infusion of etanercept (10 mg) or saline placebo"). While, rule-based model missed out both information items. In contrast, the PubMedBERT model tends to generate more false negatives than the rule-based model for information items that can be easily captured by keyword based rules. One of the common items for which this phenomenon occurs is "Trial Phase". For example, considering the following sentence from the test set: *"Bosentan treatment of digital ulcers related to systemic sclerosis : results from the RAPIDS-2 randomised , double-blind , placebo-controlled trial"*, the PubMedBERT model gave a wrong prediction of "RAPIDS-2" as "Phase 2", most likely because of the number "2" in the text that the model learnt from training data. On the other hand, the rule-based methods accurately identifies this item. We also observed that, it appears easy to identify some characteristics with limited examples using simple lexical rules, such as Comparative Intent types such as Superiority, Non-inferiority. As the results, the rule-based model achieved a slightly better performance comparing to the deep learning models. For example, the rule-based model gave correct prediction of the Comparative Intent information item in this sentence *This was a* ***superiority*** *trial, with prednisolone as the control intervention.*. But without sufficient training data, the PubMedBERT model totally missed it out.

6.6 User Study

6.6.1 User Study Design

To extrinsically evaluate the performances as well as examine the utility of our information extraction model, we conducted a pilot user study with a methodology and meta-research expert who conducts research on manual and semi-automated methods to improve research and reporting quality. The goal of the study is to have the end user looking at outputs of our model and performing two evaluation tasks: (1) evaluate if the prediction results from the model is correct or incorrect; and (2) provide feedback/comments if the extraction result would be helpful for the end user in any downstream tasks (e.g quality assessment).

We randomly selected 10 RCTs and ran our best performing NER model (PubMedBERT-CRF) to extract information from their full text publications. We designed an evaluation template which contained 10 tabs, each corresponding to one RCT. In each tab, we listed down the list of information items extracted by the model, including: the information item name, the text span, the sentence where the predicted span located, and the section to which the sentence belongs. We asked the end user to go through the list and make judgements whether a prediction is correct or not. If not, the end user was recommended to make comments to explain for their judgements. In the end, we asked the end user put a narrative feedback that he/she might have for the tool overall. (e.g. does the evaluator find the tool helpful? If yes, in which use cases? If not, why and does evaluator have any suggestions for future improvements?). The user evaluation template and results are provided in the Appendix B.

6.6.2 User Study Results

In all 10 articles, the model yielded a total of 225 predictions. Based on evaluation results from the end user, 166 predictions were marked as correct (75%); 53 predictions were marked as incorrect (24%); 6 predictions were marked as partly correct (1%).

"Randomization", "Sample Size" and "Allocation Concealment Method" are the domains that got the highest number of incorrect predictions (21, 12 and 6 respectively). Within the "Randomization" domain, majority of them were predicted for Randomization Personnel and Randomization Sequence Generation Method; which are the two most descriptive information items in this group. The model often picked up some common keywords that connect to randomization sequence generation methods such as *"interactive web-based program"*, *"standard automated calling system"* or *"telephone"* and predicted those as the information items. However, as pointed out by the end user, the phrases appear in sentences that are not related to randomization. Similarly, the model picked up text spans that describe human subjects, such as *"research team"*, *"local study team"*, *"trial team"* and predicted those as Randomization Personnel. Nevertheless, our end user determined those appear in sentences in that related to Intervention, not randomization. It was no surprise that the model made a high number of wrong predictions for Allocation Concealment information since this item has

very limited examples in the training data set, and is also one of the most descriptive information items in our data model (e.g. the span of text that describe allocation concealment is often lengthy and also can vary substantially). As for Sample Size domain, the model struggled to differentiate different types of sample size (e.g. actual at enrollment, actual at outcome analysis, required vs. target). This is not a surprise given how similar the language describing these information is. The end user's overall feedback aligns with what we have discussed. In the narrative feedback, the end user stated that *"Overall, the algorithms performed better than I expected, and most errors seemed to occur when classifying details related to randomization and sample size."*

The end user also pointed out some limitations of the model and provided suggestions for future improvements. For example, according to the end user's comment, sometimes sentences with predictions were picked up in the background or discussion that describe other studies. Even though, these sentences contain methodological information, it does not belong to the current study. The end user suggested that *"restricting the screening to title, abstract, methods, results sections could improve this"*. Secondly, the end user noted *"in a few places, there are some opposing classifications that could likely be dealt with with conditional logic. For example, if a classification is made as "Settings/Multicenter" with high confidence, a subsequent classification of "Settings/Single center" could be excluded."* Another example is the "Setting/Location" item; the end user suggested some location logic could be used to determine what is a broader country-level setting vs. local setting. More importantly, the end user pointed out that the model seem to miss relevant sentences that contain important methodological details for evaluating the design and execution of each trial. Therefore, he suggested the important next step would be to evaluate what the model did not pick up at all.

Finally, the end user provided overall feedback about the tool and how it can be used to support researchers in different contexts. Direct quote is provided below:

"I could see this being useful in several contexts. First, when extracting information for systematic reviews or meta-research projects, it is possible that this would save time so the screener could see the information in context beside the paper. Second, for general evaluation/peer review of the literature, it could help to evaluate the rigor/transparency of trials. An interesting next step could be to further classify sentences beyond just whether they reported something to whether it was implemented. For example for blinding, if it was explicitly stated that investigators or analysts were not blinded. Finally, automated extractions could be used to build a large database of trials so one could filter by randomization type, ratio, blinding, and so on if they wished to study a specific corpus of articles with certain methodological characteristics, or to track trends in methodological characteristics over time and over disciplines."

Overall, the user study results show that our model outputs are promising and can be useful in different use cases. There are multiple directions for improvements as suggested by the end user. For example, instead of running the model on all sentences from the full text publications, maybe focusing on certain parts of the full-text (e.g. only Methods section) will improve the model accuracy. Combining logical rules to validate

prediction results (e.g. a study is predicted with multicenter settings cannot be single center) could be helpful as well. Our user study has some limitations. First, it was conducted with only a single person. The user study needs to be expanded to reach broader set of end users. Second, the design of the user study evaluation was still informal (we only asked the end user to go through the list and make judgements whether a prediction is correct or not; and asked for general comments/feedback in the end). A more formal evaluation approach can be used. For example, a formal user acceptance testing can be conducted with more informative evaluating mechanism such as rating-scale questions.

6.7 Chapter Summary

We presented an annotation study and baseline NER models for recognizing methodological characteristics in RCT publications. We focused on characteristics that may affect the methodological quality of and strength of evidence from a RCT study. To our knowledge, this is the first study to focus on representing, annotating, and extracting these methodological characteristics at a fine-grained level and in a comprehensive manner. Our work complements the PICO-based characterizations which, while very important, do not address methodological quality, and automated risk of bias assessment models, which focus on classification rather than IE and thus, do not provide granular information. Our data model was adopted from OCRe and CONSORT. OCRe, by formalizing various aspects of clinical studies, and CONSORT, by detailing the characteristics of a RCT study that needs to be reported in a publication for transparency, provide a solid foundation for methodological IE.

Our NER models have limitations. First, the annotated corpus needs to be expanded to cover rare information items, so as to be more broadly useful. We anticipate that NER models would benefit from additional training data, as well. This study showed the feasibility of reliably annotating methodological characteristics at the span level and we plan to expand our corpus in future work. It would be particularly important to capture a larger number of infrequently discussed characteristics, such as allocation concealment methods, since the current models fail at recognizing them.

We only experimented with baseline NER models. While they yield promising results, more advanced NER methods can be applied (e.g., BERT with BiLSTM+CRF layers). We experimented with the learning rate hyperparameter, but tuning other hyperparameters could also be beneficial. A simple rule-based approach seems adequate for some items and it may be worthwhile to use them, especially for the items that do not have sufficient training examples.

We created a set of 16 lexical rules based on regular expressions for characteristics that can be categorized into subtypes (e.g., Blinding or Randomization Types). Although it covers only a subset of the items, this method yielded comparable results to token classification model for the characteristics that it covered. This indicates that an expanded set of such rules may be effective in methodological IE, although this involves

some manual effort and requires expertise.

We conducted a pilot user study to examine the utility of the model in practice with a methodology and meta-research expert. Though our user study has limitations (a single user, informal evaluation approach), the results shows that that our model outputs are promising and can be useful in different use cases. All in all, the pilot user study was encouraging and showed that our models can be expanded and refined to support tools for methodology and meta-research experts, and other stakeholders.

Chapter 7

Conclusions and Future Directions

7.1 Revisiting thesis research questions

This thesis builds on a motivation rooted from the evidence synthesis process, a critical yet challenging step is assessing quality of clinical research. Existing issues of this process include the fast growing body of medical literature and the extensive domain knowledge needed to fulfill the task. These motivated us to explore different approaches of using artificial intelligence, specifically NLP and ML methods, to automate some steps, so as to assist the stakeholders in the quality assessment process. With that motivation, at the beginning of the thesis, we asked two research questions:

- Research question 1: What information do biomedical researchers and other stakeholders need to assess evidence quality?
- Research question 2: How can we use NLP and ML techniques to automatically extract them?

To answer these two questions, this thesis presented three research studies where we looked at different levels of information granularity that can be used for clinical research quality assessment, and developed NLP models to automatically extract the information from full-text clinical publications:

- At the highest level of information granularity–document level, we built a classification model to distinguish different clinical study designs. This work answers the two research questions as following:
  - RQ1: we proposed to use "study design" as document-level information to assess quality of clinical research.
  - RQ2: we built two different ML models (one based on SVM and the other on PubMedBERT architecture) to automatically identify "Study Design" information from full-text clinical publications.

- At the second highest level of information granularity–sentence level, we built a classification model to map information items from reporting guidelines to sentences of full-text RCT publications. This work answers the two research questions as following:
  - RQ1: we proposed to use information items from reporting checklist–CONSORT as sentence-level information to assess quality of clinical research.
  - RQ2: we built two different ML models (one based on SVM and the other on BioBERT architecture) to automatically identify information items in CONSORT checklist from full-text RCT publications.
- Finally, at the lowest level of information granularity–mention level (token level), we developed an information extraction model that automatically identifies and extracts fine-grained methodological characteristics from full-text RCT publications. This work answers the two research questions as following:
  - RQ1: we developed our own data model that captures fine-grained methodological characteristics as mention-level information to assess quality of clinical research.
  - RQ2: we demonstrated how to use three different ML models (Token classification, CRF and rule-based) to automatically identify these fine-grained characteristics from full-text RCT publications.

All three studies follow the same theme of implementation in which our approach is to combine existing knowledge representations with the NLP and ML methods: the first development used DIDEO ontology as the backbone of the hierarchical classification model of study designs; the second development used CONSORT checklist as the data model to capture reporting information items and used them as labels of the classification model; and finally, in the third work, we developed our own data model (drawn from existing clinical research ontology OCRe and CONSORT) and leveraged their classes as labels for our information extraction model. Even though, the three developments were built separately, they conceptually connect to each other by looking at information used for quality assessment from coarse-grained to fine-grained granularity. Identifying information at different granularity levels is particularly helpful because at each stage of the evidence quality assessment process, different stakeholders have different needs. For example, systematic reviewers at the screening step can use the Study Design classifier to filter a certain type of clinical study (e.g. RCT). Journal editors at manuscript reviewing step can use the CONSORT classifier to check how much a submitted manuscript in compliance with reporting guidelines. And finally, meta researchers, who want to examine the effective sample size of a clinical study, can use our methodological characteristics extractor to identify relevant information such as sample size, power, alpha, dropout rates for the assessment. In a single scenario in which models from all three studies could be used at once, a stakeholder (such as a systematic reviewer) can start with using the first development to retrieve clinical trials with randomized controlled study design, then using the second development to identify sentences from the full text that report CONSORT information items to assess reporting quality (e.g. sentences describe blinding), then using the

third development to extract fine-grained methodological characteristics from those sentences and use the information to appraise the studies accordingly.

As mentioned in Chapters 4, 5, and 6, for each study, we made specific data and methodological choices. Our selections have several limitations. Data in all three developments are relatively small. Several factors contribute to this drawback. First, domain expertise is much needed to annotate data but it is costly (in terms of time and effort) and hard to find. Second, our work mostly are preliminary studies, therefore, we do not have existing data sets to rely on and need to create new data sets from scratch. On the other hand, in the development of the CONSORT classification model, we explored methods to use weak supervision to expand the data set but results has not shown improvement over models trained on manually annotated corpus (compare with simple hyperparameter tuning). Or in the development of the methodological characteristics extraction model, we only experimented baseline machine learning models such as pure token classification or CRF layer with the fine-tuned PubMedBERT. Even though, this thesis more focuses on making methodological contribution, and is less concerned with ML innovation, future work is needed to expand the methodologies and frameworks proposed in this thesis to better explore and apply.

7.2 Future Directions and Research

This thesis opens up several research avenues that are worth further pursuit. First, I discuss a few possible avenues of investigation from a technical point of view. Then, I discuss future direction from a application point of view to combine the three models into a single pipeline of automation tools that can be used to assist researchers in EQA and other tasks.

7.2.1 Technical improvements

First and foremost, to overcome the limitation of small data, since obtaining human annotated data is time-consuming and costly, different automatic labeling data approaches need to be explored. In recent years, research on few-shot learning (FSL) [235] or zero-shot learning (ZSL) [236] have been established. These are types of machine learning methods that allow to train models on a very small number of samples. Using prior knowledge, FSL or ZSL have the potential to rapidly generalize to new tasks containing only a few samples with supervised information. Such methods could be applied to address the limitation of data for our developments. Some data augmentation (DA) techniques also can be used to generate additional, synthetic data based on the existing data that we have [237]. For example, "back translation" is a DA technique to translate the text data to some language and then translate it back to the original language. This can help to generate textual data with different words while preserving the context of the text data. Or "synonym replacement" is another technique to replace each of words in the text with one of its synonyms to generate new data. At the same time, as shown in the Chapter 5 study, fine-tuning hyper-parameters of

machine learning model could play an important role to improve the performances. In such cases, methods to experiment with different hyperparameters could be applied such as using adaptive learning rate.

Second, instead of using the original model architecture from the existing pretrained language models (e.g. PubMedBERT or BioBERT) and only fine-tuning last layer, more complex model architectures could be explored. For example, to develop the NER model, instead of using BERT and CRF only, we can add one more layer of BiLSTM model, which has been applied and seems to get promising results in some NER tasks [238]. Or recent years have seen the paradigm shift of NER systems from sequence labeling to span prediction [239]. And the combination of traditional sequence labeling model with span prediction model have shown improvements in some NER tasks [240]. Therefore, incorporating span prediction models into our work could be helpful to improve the model performances.

7.2.2 Application improvements

Each study in this thesis, individually, can be expanded in order to be applicable for more use cases. For example, the model in Chapter 4 was built based on 7 specific evidence types from DIDEO ontology (out of total 44 evidence types defined by the ontology). This indicates that there are a lot of potential to expand this model to a broader and more comprehensive classification system that covers all 44 evidence types. The model in Chapter 5 was built based on the CONSORT checklist which is specifically design for only RCTs and also only focus on methodology-related information items. Expending our current model to capture other items such as Outcomes, Results would be helpful for downstream tasks. The same classification model can also be developed using other similar checklists and guidelines for other study designs (such as STROBE for observational studies [87]). Or the model in Chapter 6 only extracts the information items defined by our data model. Some other methodological characteristics that meta-researchers are interested in to know more, such as statistical methods, are not covered in our data model. Therefore, expanding our information extraction model to broader set of information would be helpful. Also, our work so far look at information from the body of the full text publications, while some relevant information for EQA can be extracted from tables or figure captions. For example, authors often report baseline characteristics of the studies in tables, including information about number of patients for each treatment arm or statistical analysis methods and results [241]. Expanding our extraction models to capture those information from tables could be an useful enhancement.

More importantly, in the end, we want apply these models in practice to assist researchers in EQA tasks. Before that, for all three studies, we should conduct extensive user studies at a much larger scale and in a more formal way compared with the one we did for the work in Chapter 6, to make sure the tools that we develop meet requirements and expectations from potential users. After the tools being maturely developed and evaluated, we envision a broader transparency and rigor portal to access a database of clinical studies and have all the three models in the back end. That portal will assist multiple stakeholders (could be systematic

reviewers, journal editors, meta researchers, authors) through a streamlined process of quality assessment which include the following features:

- Search for and filter specific study design publications on different topics.
- Classify sentences from full text publications to information items recommended by reporting guideline, so as to assess compliance and reporting quality.
- Extract and view methodological characteristics of the publications a structured representation, so as to assess methodological quality and spot any issues.
- Further filter relevant publication by different methodological characteristics (such as randomization type, ratio, blinding), and so on if end users wish to study a specific corpus of articles with certain methodological characteristics.

Figure 7.1 shows the idea of how the models can be connected into one single pipeline to assist researchers in the EQA process. The idea of such as streamlined portal is complimentary to the existing tools such as TrialStreamer, a living annotated database of 803,727 RCTs [141]. While TrialStreamer focuses on curating and representing RCTs based on PICO framework, our models capture other characteristics that can be used alongside PICO elements, so as to serve different purposes, even beyond EQA.

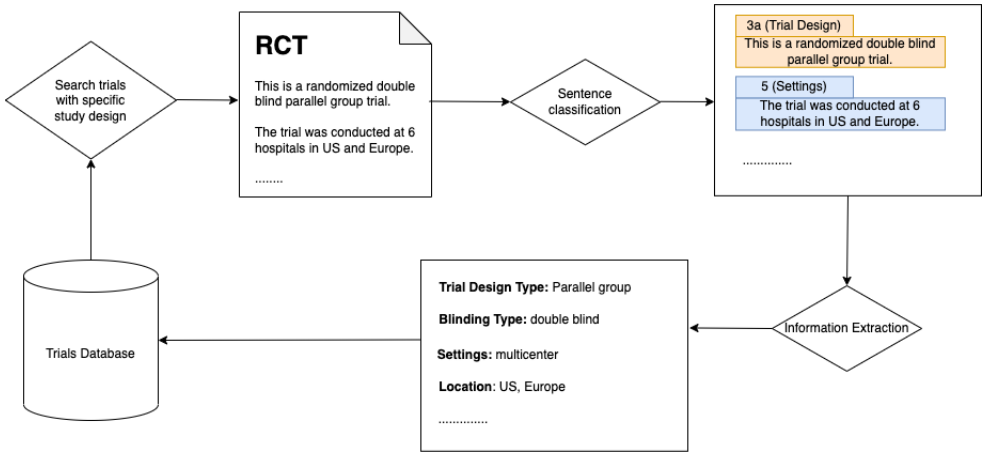

Figure 7.1: Connect three models into one single pipeline of automation tools

7.3 Final Statement

The COVID-19 pandemic in the last 3 years shows the importance of high-quality medical research for addressing global health challenges. However, research is being created in greater quantity, faster than ever before, posing challenges to identifying trustworthy scientific knowledge. In this context, quality assessment of clinical research comes into the picture as the crucial step for judging the overall strength of evidence on

given research topic and helping to answer the question whether or not a research can be applied. In this thesis, we have contributed to the growing research on using NLP techniques to automate parts of evidence synthesis, in particular evidence quality assessment. We believe that practical tools that build on such models can accelerate the evidence synthesis process and contribute to evidence-based medicine and better patient care.

References

[1] Global Alliance for Infections in Surgery, *Evidence-based medicine components*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://infectionsinsurgery.org/lets-support-evidence-based-medicine/.

[2] NT Health Library Services, *Steps to practice ebm in the form of systematic review*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://library.health.nt.gov.au/EBP/overview.

[3] Cochrane, *The evidence-based medicine pyramid*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://s4be.cochrane.org/blog/2014/04/29/the-evidence-based-medicine-pyramid/.

[4] C. Heffernan, *Cohort studies*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://www.drcath.net/toolkit/cohort.

[5] J. A. Sterne, J. Savović, M. J. Page, *et al.*, "Rob 2: A revised tool for assessing risk of bias in randomised trials," *Bmj*, vol. 366, 2019.

[6] S. Bansal, *Introduction to svm – support vector machine algorithm of machine learning*, [Online; accessed 10 01, 2022], 2021. [Online]. Available: https://www.analytixlabs.co.in/blog/introduction-support-vector-machine-algorithm.

[7] 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)*, Association for Computational Linguistics, 2019, pp. 4171–4186.

[8] D. L. Sackett, "Evidence-based medicine," *Seminars in perinatology*, vol. 21, no. 1, pp. 3–5, 1997.

[9] B. M. Melnyk and E. Fineout-Overholt, *Evidence-based practice in nursing & healthcare: A guide to best practice*. Lippincott Williams & Wilkins, 2011.

[10] R. E. Slavin, "Best evidence synthesis: An intelligent alternative to meta-analysis," *Journal of clinical epidemiology*, vol. 48, no. 1, pp. 9–18, 1995.

[11] P. C. Giannelli, E. J. Imwinkelried, A. Roth, J. C. Moriarty, and V. E. Beety, *Scientific evidence*. HeinOnline, 1986.

[12] K. Khan, R. Kunz, J. Kleijnen, and G. Antes, *Systematic reviews to support evidence-based medicine, 2nd edition*. CRC Press, 2011.

[13] R. B. Haynes, D. L. Sackett, W. S. Richardson, W. Rosenberg, and G. R. Langley, "Evidence-based medicine: How to practice & teach ebm," *Canadian Medical Association Journal*, vol. 157, no. 6, p. 788, 1997.

[14] C. M. d. C. Santos, C. A. d. M. Pimenta, and M. R. C. Nobre, "The pico strategy for the research question construction and evidence search," *Revista latino-americana de enfermagem*, vol. 15, pp. 508–511, 2007.

[15] A. Cortegiani, G. Ingoglia, M. Ippolito, A. Giarratano, and S. Einav, "A systematic review on the efficacy and safety of chloroquine for the treatment of covid-19," *Journal of critical care*, vol. 57, pp. 279–283, 2020.

[16] J. P. Ioannidis, "Why most published research findings are false," *PLoS medicine*, vol. 2, no. 8, e124, 2005.

[17] G. Guyatt, A. D. Oxman, E. A. Akl, *et al.*, "Grade guidelines: 1. introduction—grade evidence profiles and summary of findings tables," *Journal of clinical epidemiology*, vol. 64, no. 4, pp. 383–394, 2011.

[18] 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*, pp. 205–228, 2019.

[19] J. M. Young and M. J. Solomon, "How to critically appraise an article," *Nature Clinical Practice Gastroenterology & Hepatology*, vol. 6, no. 2, pp. 82–91, 2009.

[20] P. Jüni, D. G. Altman, and M. Egger, "Assessing the quality of controlled clinical trials," *Bmj*, vol. 323, no. 7303, pp. 42–46, 2001.

[21] A. Al-Jundi and S. Sakka, "Critical appraisal of clinical research," *Journal of clinical and diagnostic research: JCDR*, vol. 11, no. 5, JE01, 2017.

[22] K. Pussegoda, L. Turner, C. Garritty, *et al.*, "Identifying approaches for assessing methodological and reporting quality of systematic reviews: A descriptive study," *Systematic reviews*, vol. 6, no. 1, pp. 1–12, 2017.

[23] B. J. Shea, J. M. Grimshaw, G. A. Wells, *et al.*, "Development of amstar: A measurement tool to assess the methodological quality of systematic reviews," *BMC medical research methodology*, vol. 7, no. 1, pp. 1–7, 2007.

[24] A. R. Jadad, R. A. Moore, D. Carroll, *et al.*, "Assessing the quality of reports of randomized clinical trials: Is blinding necessary?" *Controlled clinical trials*, vol. 17, no. 1, pp. 1–12, 1996.

[25] D. A. Grimes and K. F. Schulz, "An overview of clinical research: The lay of the land," *The Lancet*, vol. 359, no. 9300, pp. 57–61, 2002.

[26] S. Sanderson, I. D. Tatt, and J. Higgins, "Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography," *International journal of epidemiology*, vol. 36, no. 3, pp. 666–676, 2007.

[27] X. Wang and X. Ji, "Sample size estimation in clinical research: From randomized controlled trials to observational studies," *Chest*, vol. 158, no. 1, S12–S20, 2020.

[28] I. Simera, D. Moher, J. Hoey, K. F. Schulz, and D. G. Altman, "A catalogue of reporting guidelines for health research," *European journal of clinical investigation*, vol. 40, no. 1, pp. 35–53, 2010.

[29] D. Moher, K. F. Schulz, D. G. Altman, C. Group, *et al.*, "The consort statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials," *The Lancet*, vol. 357, no. 9263, pp. 1191–1194, 2001.

[30] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, "The strengthening the reporting of observational studies in epidemiology (strobe) statement: Guidelines for reporting observational studies," *Bulletin of the World Health Organization*, vol. 85, pp. 867–872, 2007.

[31] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman, "Improving bioscience research reporting: The arrive guidelines for reporting animal research," *Osteoarthritis and cartilage*, vol. 20, no. 4, pp. 256–260, 2012.

[32] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *Systematic reviews*, vol. 10, no. 1, pp. 1–11, 2021.

[33] D. G. Altman, I. Simera, J. Hoey, D. Moher, and K. Schulz, "Equator: Reporting guidelines for health research," *Open Medicine*, vol. 2, no. 2, e49, 2008.

[34] CONSORT, *Impact of consort*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: http://www.consort-statement.org/about-consort/impact-of-consort.

[35] R. C. Group, "Effect of hydroxychloroquine in hospitalized patients with covid-19," *New England Journal of Medicine*, vol. 383, no. 21, pp. 2030–2040, 2020.

[36] A. B. Cavalcanti, F. G. Zampieri, R. G. Rosa, *et al.*, "Hydroxychloroquine with or without azithromycin in mild-to-moderate covid-19," *New England Journal of Medicine*, vol. 383, no. 21, pp. 2041–2052, 2020.

[37] S. Garba, A. Ahmed, A. Mai, G. Makama, and V. Odigie, "Proliferations of scientific medical journals: A burden or a blessing," *Oman medical journal*, vol. 25, no. 4, p. 311, 2010.

[38] A. Ross-White and C. Godfrey, "Is there an optimum number needed to retrieve to justify inclusion of a database in a systematic review search?" *Health Information & Libraries Journal*, vol. 34, no. 3, pp. 217–224, 2017.

[39] 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*, vol. 7, no. 2, e012545, 2017.

[40] R. Smith, "Peer review: A flawed process at the heart of science and journals," *Journal of the royal society of medicine*, vol. 99, no. 4, pp. 178–182, 2006.

[41] J. Huisman and J. Smits, "Duration and quality of the peer review process: The author's perspective," *Scientometrics*, vol. 113, no. 1, pp. 633–650, 2017.

[42] R. E. Gropp, S. Glisson, S. Gallo, and L. Thompson, "Peer review: A system under stress," *BioScience*, vol. 67, no. 5, pp. 407–410, 2017.

[43] B. Zhong, "How to calculate sample size in randomized controlled trial?" *Journal of thoracic disease*, vol. 1, no. 1, p. 51, 2009.

[44] D. Moher, C. S. Dulberg, and G. A. Wells, "Statistical power, sample size, and their reporting in randomized controlled trials," *Jama*, vol. 272, no. 2, pp. 122–124, 1994.

[45] A. W. Chan and D. G. Altman, "Epidemiology and reporting of randomised trials published in pubmed journals," *The Lancet*, vol. 365, no. 9465, pp. 1159–1162, 2005.

[46] H. J. Lamberink, W. M. Otte, M. R. Sinke, *et al.*, "Statistical power of clinical trials increased while effect size remained stable: An empirical analysis of 136,212 clinical trials between 1975 and 2014," *Journal of Clinical Epidemiology*, vol. 102, pp. 123–128, 2018.

[47] C. C. Serdar, M. Cihan, D. Yücel, and M. A. Serdar, "Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies," *Biochemia medica*, vol. 31, no. 1, pp. 27–53, 2021.

[48] A. Penić, D. Begić, K. Balajić, M. Kowalski, A. Marušić, and L. Puljak, "Definitions of blinding in randomised controlled trials of interventions published in high-impact anaesthesiology journals: A methodological study and survey of authors," *BMJ open*, vol. 10, no. 4, e035168, 2020.

[49] A. J. Grizzle, L. E. Hines, D. C. Malone, O. Kravchenko, H. Hochheiser, and R. D. Boyce, "Testing the face validity and inter-rater agreement of a simple approach to drug-drug interaction evidence assessment," *Journal of biomedical informatics*, vol. 101, p. 103 355, 2020.

[50] L. Turner, L. Shamseer, D. G. Altman, K. F. Schulz, and D. Moher, "Does use of the consort statement impact the completeness of reporting of randomised controlled trials published in medical journals? a cochrane reviewa," *Systematic reviews*, vol. 1, no. 1, pp. 1–7, 2012.

[51] J. P. Ioannidis, "Meta-research: Why research on research matters," *PLoS biology*, vol. 16, no. 3, e2005468, 2018.

[52] D. L. Lorenzetti and W. A. Ghali, "Reference management software for systematic reviews and meta-analyses: An exploration of usage and usability," *BMC medical research methodology*, vol. 13, no. 1, pp. 1–5, 2013.

[53] L. Hoang and J. Schneider, "Opportunities for computer support for systematic reviewing-a gap analysis," in *Transforming Digital Worlds. iConference Lecture Notes in Computer Science*, Springer, 2018, pp. 367–377.

[54] C. C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: Success, failure and the future," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 132–144, 2016.

[55] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera, "Systematic review automation technologies," *Systematic reviews*, vol. 3, no. 1, pp. 1–15, 2014.

[56] R. Van Dinter, B. Tekinerdogan, and C. Catal, "Automation of systematic literature reviews: A systematic literature review," *Information and Software Technology*, vol. 136, p. 106 589, 2021.

[57] H. Kilicoglu, "Biomedical text mining for research rigor and integrity: Tasks, challenges, directions," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1400–1414, 2018.

[58] I. J. Marshall and B. C. Wallace, "Toward systematic review automation: A practical guide to using machine learning tools in research synthesis," *Systematic reviews*, vol. 8, no. 1, pp. 1–10, 2019.

[59] 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*, vol. 23, no. 1, pp. 193–201, 2016.

[60] 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*, vol. 45, no. 1, pp. 266–277, 2016.

[61] S. R. Jonnalagadda, P. Goyal, and M. D. Huffman, "Automating data extraction in systematic reviews: A systematic review," *Systematic reviews*, vol. 4, no. 1, pp. 1–16, 2015.

[62] L. Schmidt, B. K. Olorisade, L. A. McGuinness, J. Thomas, and J. P. Higgins, "Data extraction methods for systematic review (semi) automation: A living review protocol," *F1000Research*, vol. 9, 2020.

[63] E. V. Villanueva, E. A. Burrows, P. A. Fennessy, M. Rajendran, and J. N. Anderson, "Improving question formulation for use in evidence appraisal in a tertiary care setting: A randomised controlled trial," *BMC medical informatics and decision making*, vol. 1, no. 1, pp. 1–9, 2001.

[64] D. G. Altman and C. Dore, "Randomisation and baseline comparisons in clinical trials," *The Lancet*, vol. 335, no. 8682, pp. 149–153, 1990.

[65] S. West, V. King, T. S. Carey, *et al.*, "Systems to rate the strength of scientific evidence," *Evidence report technology assessment (Summary)*, no. 47, pp. 1–11, 2002.

[66] D. Moher, A. R. Jadad, G. Nichol, M. Penman, P. Tugwell, and S. Walsh, "Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists," *Controlled clinical trials*, vol. 16, no. 1, pp. 62–73, 1995.

[67] T. C. Chalmers, H. Smith Jr, B. Blackburn, *et al.*, "A method for assessing the quality of a randomized control trial," *Controlled clinical trials*, vol. 2, no. 1, pp. 31–49, 1981.

[68] R. DerSimonian, L. J. Charette, B. McPeek, and F. Mosteller, "Reporting on methods in clinical trials," in *Medical uses of statistics*, CRC Press, 2019, pp. 333–347.

[69] J. Pell, *Clinical Trials Dictionary: Terminology and Usage Recommendations*. Emerald Group Publishing Limited, 2013.

[70] A. G. Chidambaram and M. Josephson, "Clinical research study designs: The essentials," *Pediatric investigation*, vol. 3, no. 04, pp. 245–252, 2019.

[71] E. Hariton and J. J. Locascio, "Randomised controlled trials—the gold standard for effectiveness research," *BJOG: an international journal of obstetrics and gynaecology*, vol. 125, no. 13, p. 1716, 2018.

[72] J. Dettori, "The random allocation process: Two things you need to know," *Evidence-based spine-care journal*, vol. 1, no. 03, pp. 7–9, 2010.

[73] 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*, vol. 7, no. 1, pp. 1–11, 2020.

[74] J. W. Song and K. C. Chung, "Observational studies: Cohort and case-control studies," *Plastic and reconstructive surgery*, vol. 126, no. 6, p. 2234, 2010.

[75] D. Moher, S. Hopewell, K. F. Schulz, *et al.*, "Consort 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials," *International journal of surgery*, vol. 10, no. 1, pp. 28–55, 2012.

[76] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman, "Improving bioscience research reporting: The arrive guidelines for reporting animal research," *Journal of Pharmacology and Pharmacotherapeutics*, vol. 1, no. 2, pp. 94–99, 2010.

[77] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement," *Annals of internal medicine*, vol. 151, no. 4, pp. 264–269, 2009.

[78] M. K. Campbell, G. Piaggio, D. R. Elbourne, and D. G. Altman, "Consort 2010 statement: Extension to cluster randomised trials," *Bmj*, vol. 345, 2012.

[79] R. Wang, V. DeGruttola, Q. Lei, *et al.*, "The vitamin d for covid-19 (vivid) trial: A pragmatic cluster-randomized design," *Contemporary clinical trials*, vol. 100, p. 106 176, 2021.

[80] P. J. Karanicolas, F. Farrokhyar, and M. Bhandari, "Blinding: Who, what, when, why, how?" *Canadian journal of surgery*, vol. 53, no. 5, p. 345, 2010.

[81] DistillerSR, *Distillersr: Literature review software smarter reviews: Trusted evidence*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://www.evidencepartners.com/products/distillersr-systematic-review-software.

[82] Covidence, *Covidence: Better systematic review management*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://www.covidence.org/.

[83] EPPI-Centre, *Eppi-centre*, [Online; accessed 10 01, 2022], 2019. [Online]. Available: https://eppi.ioe.ac.uk/cms/.

[84] Cochrane, *Epoc resources for review authors*, [Online; accessed 10 01, 2022], 2021. [Online]. Available: https://epoc.cochrane.org/resources/epoc-resources-review-authors.

[85] Institute of health economics Alberta Canada, *Ihe quality appraisal checklist for case series studies*, [Online; accessed 10 01, 2022], 2016. [Online]. Available: https://www.ihe.ca/publications/ihe-quality-appraisal-checklist-for-case-series-studies.

[86] National Heart, Lung, and Blood Institute, *Study quality assessment tools*, [Online; accessed 10 01, 2022], 2021. [Online]. Available: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.

[87] J. A. Sterne, M. A. Hernán, B. C. Reeves, *et al.*, "Robins-i: A tool for assessing risk of bias in non-randomised studies of interventions," *bmj*, vol. 355, 2016.

[88] C. Marshall, A. Sutton, H. O'Keefe, E. Johnson, *Systematic review toolbox*, [Online; accessed 10 01, 2022], 2019. [Online]. Available: http://www.systematicreviewtools.com/.

[89] D. Demner-Fushman and J. Lin, "Knowledge extraction for clinical question answering: Preliminary results," in *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*, American Association for Artificial Intelligence, 2005, pp. 9–13.

[90] D. Demner-Fushman, B. Few, S. E. Hauser, and G. Thoma, "Automatically identifying health outcome information in medline records," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 52–60, 2006.

[91] M. Fiszman, D. Demner-Fushman, F.-M. Lang, P. Goetz, and T. C. Rindflesch, "Interpreting comparative constructions in biomedical text," in *Biological, translational, and clinical language processing*, 2007, pp. 137–144.

[92] G. Chung and E. Coiera, "A study of structured clinical abstracts and the semantic classification of sentences," in *Biological, translational, and clinical language processing*, 2007, pp. 121–128.

[93] K. Hara and Y. Matsumoto, "Extracting clinical trial design information from medline abstracts," *New Generation Computing*, vol. 25, no. 3, pp. 263–275, 2007.

[94] 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," *Medinfo*, vol. 129, pp. 550–4, 2007.

[95] 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*, vol. 15, no. 1, pp. 9–16, 2007.

[96] 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*, American Medical Informatics Association, vol. 2008, 2008, p. 141.

[97] M. J. Hansen, N. Ø. Rasmussen, and G. Chung, "A method of extracting the number of trial participants from abstracts describing randomized controlled trials," *Journal of Telemedicine and Telecare*, vol. 14, no. 7, pp. 354–358, 2008.

[98] G. Y. Chung, "Sentence retrieval for abstracts of randomized controlled trials," *BMC medical informatics and decision making*, vol. 9, no. 1, pp. 1–13, 2009.

[99] G. Y.-C. Chung, "Towards identifying intervention arms in randomized controlled trials: Extracting coordinating constructions," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 790–800, 2009.

[100] R. Summerscales, S. Argamon, J. Hupert, and A. Schwartz, "Identifying treatments, groups, and outcomes in medical abstracts," in *The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009)*, Indiana University Bloomington, IN, USA, 2009.

[101] S. Kiritchenko, B. De Bruijn, S. Carini, J. Martin, and I. Sim, "Exact: Automatic extraction of clinical trial characteristics from journal publications," *BMC medical informatics and decision making*, vol. 10, no. 1, pp. 1–17, 2010.

[102] 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*, vol. 10, no. 1, pp. 1–6, 2010.

[103] H. Xu, Y. Lu, M. Jiang, *et al.*, "Mining biomedical literature for terms related to epidemiologic exposures," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2010, 2010, p. 897.

[104] F. Boudin, L. Shi, and J.-Y. Nie, "Improving medical information retrieval with pico element detection," in *European Conference on Information Retrieval*, Springer, 2010, pp. 50–61.

[105] S. Lin, J. P. Ng, S. Pradhan, J. Shah, R. Pietrobon, and M.-Y. Kan, "Extracting formulaic and free text clinical research articles metadata using conditional random fields," in *Proceedings of the NAACL HLT 2010 second Louhi workshop on text and data mining of health documents*, 2010, pp. 90–95.

[106] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, BioMed Central, vol. 12, 2011, pp. 1–10.

[107] R. L. Summerscales, S. Argamon, S. Bai, J. Hupert, and A. Schwartz, "Automatic summarization of results from clinical trials," in *2011 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE, 2011, pp. 372–377.

[108] K.-C. Huang, C. C.-H. Liu, S.-S. Yang, *et al.*, "Classification of pico elements by text features systematically extracted from pubmed abstracts," in *2011 IEEE International Conference on Granular Computing*, IEEE, 2011, pp. 279–283.

[109] 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*, 2012, pp. 579–589.

[110] J. Zhao, P. Bysani, and M.-Y. Kan, "Exploiting classification correlations for the extraction of evidence-based practice information," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2012, 2012, p. 1070.

[111] I. Amini, D. Martínez, and D. M. Aliod, "Overview of the alta 2012 shared task," in *Australian Language Technology Association*, 2012.

[112] H. Zhu, Y. Ni, P. Cai, Z. Qiu, and F. Cao, "Automatic extracting of patient-related attributes: Disease, age, gender and race," in *Quality of Life through Quality of Information*, IOS Press, 2012, pp. 589–593.

[113] W. Hsu, W. Speier, and R. K. Taira, "Automated extraction of reported statistical analyses: Towards a logical representation of clinical trial literature," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2012, 2012, p. 350.

[114] 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*, vol. 46, no. 5, pp. 940–946, 2013.

[115] C. Kelly and H. Yang, "A system for extracting study design parameters from nutritional genomics abstracts," *Journal of integrative bioinformatics*, vol. 10, no. 2, pp. 82–93, 2013.

[116] H. Hassanzadeh, T. Groza, and J. Hunter, "Identifying scientific artefacts in biomedical literature: The evidence based medicine use case," *Journal of biomedical informatics*, vol. 49, pp. 159–170, 2014.

[117] G. Karystianis, I. Buchan, and G. Nenadic, "Mining characteristics of epidemiological studies from medline: A case study in obesity," *Journal of biomedical semantics*, vol. 5, no. 1, pp. 1–11, 2014.

[118] 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)*, IEEE, 2015, pp. 1–4.

[119] C. Blake and A. Lucic, "Automatic endpoint detection to support the systematic review process," *Journal of biomedical informatics*, vol. 56, pp. 42–56, 2015.

[120] W. Suwarningsih, A. Purwarianti, and I. Supriana, "Indonesian medical question classification with pattern matching," in *2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT)*, IEEE, 2015, pp. 106–109.

[121] 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*, vol. 17, no. 1, pp. 4572–4596, 2016.

[122] K. Raja, N. Dasot, P. Goyal, and S. R. Jonnalagadda, "Towards evidence-based precision medicine: Extracting population information from biomedical text using binary classifiers and syntactic patterns," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 203, 2016.

[123] T. Basu, S. Kumar, A. Kalyan, *et al.*, "A novel framework to expedite systematic reviews by automatically building information extraction training corpora," *arXiv preprint arXiv:1606.06424*, 2016.

[124] D. D. A. Bui, G. Del Fiol, J. F. Hurdle, and S. Jonnalagadda, "Extractive text summarization system to aid data extraction from full text in systematic review development," *Journal of biomedical informatics*, vol. 64, pp. 265–272, 2016.

[125] G. Singh, I. J. Marshall, J. Thomas, J. Shawe-Taylor, and B. C. Wallace, "A neural candidate-selector architecture for automatic structured clinical text annotation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1519–1528.

[126] I. J. Marshall, J. Kuiper, E. Banner, and B. C. Wallace, "Automating biomedical evidence synthesis: Robotreviewer," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, NIH Public Access, vol. 2017, 2017, p. 7.

[127] G. Karystianis, K. Thayer, M. Wolfe, and G. Tsafnat, "Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews," *Journal of biomedical informatics*, vol. 70, pp. 27–34, 2017.

[128] A. Lucic and C. L. Blake, "Improving endpoint detection to support automated systematic reviews," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2016, 2016, p. 1900.

[129] D. Jin and P. Szolovits, "Advancing pico element detection in biomedical text via deep neural networks," *Bioinformatics*, vol. 36, no. 12, pp. 3856–3862, 2020.

[130] 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*, 2018, pp. 67–75.

[131] B. Nye, J. J. Li, R. Patel, *et al.*, "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*, NIH Public Access, vol. 2018, 2018, p. 197.

[132] D. Demner-Fushman, J. G. Mork, W. J. Rogers, S. E. Shooshan, L. Rodriguez, and A. R. Aronson, "Finding medication doses in the liteature," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2018, 2018, p. 368.

[133] 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*, vol. 18, no. 1, pp. 1–14, 2018.

[134] C. Baladrón, A. Santos-Lozano, J. M. Aguiar, A. Lucia, and J. Martın-Hernández, "Tool for filtering pubmed search results by sample size," *Journal of the American Medical Informatics Association*, vol. 25, no. 7, pp. 774–779, 2018.

[135] A. J. Brockmeier, M. Ju, P. Przybyła, and S. Ananiadou, "Improving reference prioritisation with pico recognition," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–14, 2019.

[136] T. Kang, S. Zou, and C. Weng, "Pretraining to recognize pico elements from randomized controlled trial literature," *Studies in health technology and informatics*, vol. 264, p. 188, 2019.

[137] X. Yuan, L. Xiaoli, L. Shilei, S. Qinwen, and L. Ke, "Extracting pico elements from rct abstracts using 1-2gram analysis and multitask classification," in *Proceedings of the third International Conference on Medical and Health Informatics 2019*, 2019, pp. 194–199.

[138] J. Guo, C. Blake, and Y. Guan, "Evaluating automated entity extraction with respect to drug and non-drug treatment strategies," *Journal of Biomedical Informatics*, vol. 94, pp. 103–112, 2019.

[139] C. Norman, M. Leeflang, R. Spijker, E. Kanoulas, and A. Névéol, "A distantly supervised dataset for automated data extraction from diagnostic studies," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 105–114.

[140] J. Brassey, C. Price, J. Edwards, M. Zlabinger, A. Bampoulidis, and A. Hanbury, "Developing a fully automated evidence synthesis tool for identifying, assessing and collating the evidence," *BMJ Evidence-Based Medicine*, vol. 26, no. 1, pp. 24–27, 2021.

[141] I. J. Marshall, B. Nye, J. Kuiper, *et al.*, "Trialstreamer: A living, automatically updated database of clinical trial reports," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1903–1912, 2020.

[142] V. R. Walker, C. P. Schmitt, M. S. Wolfe, *et al.*, "Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr," *Environment international*, vol. 159, p. 107 025, 2022.

[143] M. Edwards and C. Marshall, "Evaluating robotreviewer for automated risk of bias assessment in a systematic review: A case study," *Value in Health*, vol. 20, no. 9, A774, 2017.

[144] A. Gates, B. Vandermeer, and L. Hartling, "Technology-assisted risk of bias assessment in systematic reviews: A prospective cross-sectional evaluation of the robotreviewer machine learning tool," *Journal of clinical epidemiology*, vol. 96, pp. 54–62, 2018.

[145] 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*, vol. 53, no. 2, pp. 246–254, 2021.

[146] 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*, vol. 22, no. 1, pp. 1–12, 2022.

[147] I. J. Marshall, B. T. Johnson, Z. Wang, S. Rajasekaran, and B. C. Wallace, "Semi-automated evidence synthesis in health psychology: Current methods and future prospects," *Health Psychology Review*, vol. 14, no. 1, pp. 145–158, 2020.

[148] C. Friedman, P. Kra, and A. Rzhetsky, "Two biomedical sublanguages: A description based on the theories of zellig harris," *Journal of biomedical informatics*, vol. 35, no. 4, pp. 222–235, 2002.

[149] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches to educational data mining," *Complexity*, vol. 2019, 2019.

[150] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2016, pp. 260–270.

[151] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[152] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[153] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13, Curran Associates Inc., 2013, pp. 3111–3119.

[154] 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, pp. 1532–1543.

[155] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[156] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, *et al.*, "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Research*, vol. 304, p. 114 135, 2021.

[157] J. Lee, W. Yoon, S. Kim, *et al.*, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[158] Y. Gu, R. Tinn, H. Cheng, *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, 2021.

[159] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining text data*, Springer, 2012, pp. 163–222.

[160] J. Gu, Z. Wang, J. Kuen, *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[161] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[162] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[163] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[164] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, Makuhari, vol. 2, 2010, pp. 1045–1048.

[165] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[166] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, 2004, pp. 107–110.

[167] S. Liu, Y. Sun, B. Li, W. Wang, and X. Zhao, "Hamner: Headword amplified multi-span distantly supervised method for domain specific named entity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 8401–8408.

[168] N. Chinchor and B. Sundheim, "MUC-5 evaluation metrics," in *Proceedings of Fifth Message Understanding Conference (MUC-5)*, 1993.

[169] L. Hoang, R. D. Boyce, N. Bosch, B. Stottlemyer, M. Brochhausen, and J. Schneider, "Automatically classifying the evidence type of drug-drug interaction research papers as a step toward computer supported evidence curation," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2020, 2020, p. 554.

[170] L. Hoang, R. D. Boyce, N. Bosch, B. Stottlemyer, M. Brochhausen, and J. Schneider, "Automatically classifying study designs of biomedical literature relevant for inclusion in the drug interactions knowledge bases," in *Scientific literature knowledge bases at AKBC Proceedings*, Automatic Knowledge Base Construction, 2019.

[171] H. Van Der Sijs, J. Aarts, T. Van Gelder, M. Berg, and A. Vulto, "Turning off frequently overridden drug alerts: Limited opportunities for doing it safely," *Journal of the American Medical Informatics Association*, vol. 15, no. 4, pp. 439–448, 2008.

[172] R. T. Scheife, L. E. Hines, R. D. Boyce, *et al.*, "Consensus recommendations for systematic evaluation of drug–drug interaction evidence for clinical decision support," *Drug safety*, vol. 38, no. 2, pp. 197–206, 2015.

[173] A. M. Sirajuddin, J. A. Osheroff, D. F. Sittig, J. Chuo, F. Velasco, and D. A. Collins, "Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision support: Effective cds is essential for addressing healthcare performance improvement imperatives," *Journal of Healthcare Information Management*, vol. 23, no. 4, p. 38, 2009.

[174] K. W. Fung, J. Kapusnik-Uner, J. Cunningham, S. Higby-Baker, and O. Bodenreider, "Comparison of three commercial knowledge bases for detection of drug-drug interactions in clinical decision support," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 806–812, 2017.

[175] K. M. Romagnoli, S. D. Nelson, L. Hines, P. Empey, R. D. Boyce, and H. Hochheiser, "Information needs for making clinical recommendations about potential drug-drug interactions: A synthesis of literature review and interviews," *BMC medical informatics and decision making*, vol. 17, no. 1, pp. 1–9, 2017.

[176] Y. Böttiger, K. Laine, M. L. Andersson, *et al.*, "Sfinx—a drug-drug interaction database designed for clinical decision support systems," *European journal of clinical pharmacology*, vol. 65, no. 6, pp. 627–633, 2009.

[177] K. Seden, S. Gibbons, C. Marzolini, *et al.*, "Development of an evidence evaluation and synthesis system for drug-drug interactions, and its application to a systematic review of hiv and malaria co-infection," *PLoS One*, vol. 12, no. 3, e0173509, 2017.

[178] R. Boyce, C. Collins, J. Horn, and I. Kalet, "Computing with evidence: Part i: A drug-mechanism evidence taxonomy oriented toward confidence assignment," *Journal of biomedical informatics*, vol. 42, no. 6, pp. 979–989, 2009.

[179] L. S. Wieland, K. A. Robinson, and K. Dickersin, "Understanding why evidence from randomised clinical trials may not be retrieved from medline: Comparison of indexed and non-indexed records," *Bmj*, vol. 344, 2012.

[180] T. Edinger and A. M. Cohen, "A large-scale analysis of the reasons given for excluding articles that are retrieved by literature search during systematic review," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2013, 2013, p. 379.

[181] A. M. Cohen, J. Schneider, Y. Fu, *et al.*, "Fifty ways to tag your pubtypes: Multi-tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine," *medRxiv [Preprint under review]*, 2021.

[182] L. Lewis, "Drug–drug interactions: Is there an optimal way to study them?" *British journal of clinical pharmacology*, vol. 70, no. 6, p. 781, 2010.

[183] M. Brochhausen, *Drug-drug interaction and drug-drug interaction evidence ontology*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: http://purl.obolibrary.org/obo/dideo/release/2022-06-14/dideo.owl.

[184] M. Brochhausen, J. Schneider, D. Malone, P. E. Empey, W. R. Hogan, and R. D. Boyce, "Towards a foundational representation of potential drug-drug interaction knowledge," in *CEUR workshop proceedings*, NIH Public Access, vol. 1309, 2014, p. 16.

[185] R. D. Boyce, C. Collins, J. Horn, and I. Kalet, "Modeling drug mechanism knowledge using evidence and truth maintenance," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 4, pp. 386–397, 2007.

[186] J. Schneider, M. Brochhausen, S. Rosko, *et al.*, "Formalizing knowledge and evidence about potential drug-drug interactions.," in *BDM2I@ ISWC*, 2015.

[187] National Center for Biotechnology Information, National Library of Medicine, *Pubmed apis*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/home/develop/api/.

[188] pdfminer.six, *Pdfminer.six's documentation*, [Online; accessed 10 01, 2022], 2019. [Online]. Available: https://pdfminersix.readthedocs.io/en/latest/.

[189] NLTK Team, *Natural language toolkit*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://pdfminersix.readthedocs.io/en/latest/.

[190] National Library of Medicine. National Institutes of Health, *Metamap - a tool for recognizing umls concepts in text*, [Online; accessed 10 01, 2022], 2020. [Online]. Available: http://metamap.nlm.nih.gov/.

[191] S. Kiritchenko, S. Matwin, A. F. Famili, *et al.*, "Functional annotation of genes using hierarchical text categorization," in *Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 2005.

[192] S. Subramanian, I. Baldini, S. Ravichandran, *et al.*, "A natural language processing system for extracting evidence of drug repurposing from scientific publications," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 369–13 381.

[193] 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*, 2009, pp. 1003–1011.

[194] H. Kilicoglu, G. Rosemblat, L. Hoang, *et al.*, "Toward assessing clinical trial publications for reporting transparency," *Journal of biomedical informatics*, vol. 116, p. 103 717, 2021.

[195] L. Hoang, L. Jiang, and H. Kilicoglu, "Investigating the impact of weakly supervised data on text mining models of publication transparency: A case study on randomized controlled trials," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2022, 2022, p. 254.

[196] H. Kilicoglu, L. K. Hoang, and S. Wadhwa, "Identifying sample size characteristics in randomized controlled trial publications," in *AMIA Annual Symposium Proceedings*, 2020.

[197] J. C. Sánchez-Thorin, M. C. Cortés, M. Montenegro, and N. Villate, "The quality of reporting of randomized clinical trials published in ophthalmology," *Ophthalmology*, vol. 108, no. 2, pp. 410–415, 2001.

[198] R. Agha, D. Cooper, and G. Muir, "The reporting quality of randomised controlled trials in surgery: A systematic review," *International Journal of Surgery*, vol. 5, no. 6, pp. 413–422, 2007.

[199] N. Parsons, R. Hiskens, C. Price, J. Achten, and M. Costa, "A systematic survey of the quality of research reporting in general orthopaedic journals," *The Journal of Bone and Joint Surgery. British Volume*, vol. 93, no. 9, pp. 1154–1159, 2011.

[200] Y. Yin, F. Shi, Y. Zhang, X. Zhang, J. Ye, and J. Zhang, "Evaluation of reporting quality of randomized controlled trials in patients with covid-19 using the consort statement," *PloS one*, vol. 16, no. 9, e0257093, 2021.

[201] S. Hopewell, D. G. Altman, D. Moher, and K. F. Schulz, "Endorsement of the consort statement by high impact factor medical journals: A survey of journal editors and journal 'instructions to authors'," *Trials*, vol. 9, no. 1, pp. 1–7, 2008.

[202] R. L. Kane, J. Wang, and J. Garrard, "Reporting in randomized clinical trials improved after adoption of the consort statement," *Journal of clinical epidemiology*, vol. 60, no. 3, pp. 241–249, 2007.

[203] R. Schulz, G. Langen, R. Prill, M. Cassel, and T. L. Weissgerber, "Reporting and transparent research practices in sports medicine and orthopaedic clinical trials: A meta-research study," *BMJ open*, vol. 12, no. 8, e059347, 2022.

[204] T. Weissgerber, N. Riedel, H. Kilicoglu, *et al.*, "Automated screening of covid-19 preprints: Can we help authors to improve transparency and reproducibility?" *Nature medicine*, vol. 27, no. 1, pp. 6–7, 2021.

[205] J. Menke, M. Roelandse, B. Ozyurt, M. Martone, and A. Bandrowski, "The rigor and transparency index quality metric for assessing biological and medical science methods," *Iscience*, vol. 23, no. 11, p. 101 698, 2020.

[206] N. Riedel, M. Kip, and E. Bobrov, "Oddpub – a text-mining algorithm to detect data sharing in biomedical publications," *Data Science Journal*, vol. 19, p. 42, 2020.

[207] S. Saladi, "Jetfighter: Towards figure accuracy and accessibility," *Elife*, 2019.

[208] H. Kilicoglu, G. Rosemblat, M. Malički, and G. Ter Riet, "Automatic recognition of self-acknowledged limitations in clinical research literature," *Journal of the American Medical Informatics Association*, vol. 25, no. 7, pp. 855–861, 2018.

[209] C. H. Vinkers, H. J. Lamberink, J. K. Tijdink, *et al.*, "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*, vol. 19, no. 4, e3001162, 2021.

[210] A. G. Pandya, L. S. Hynan, R. Bhore, *et al.*, "Reliability assessment and validation of the melasma area and severity index (masi) and a new modified masi scoring method," *Journal of the American Academy of Dermatology*, vol. 64, no. 1, pp. 78–83, 2011.

[211] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.

[212] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: The metamap program.," in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, 2001, p. 17.

[213] 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*, NIH Public Access, vol. 11, 2017, p. 269.

[214] Y. Wang, S. Sohn, S. Liu, *et al.*, "A clinical text classification paradigm using weak supervision and deep representation," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–13, 2019.

[215] A. Nentidis, A. Krithara, G. Tsoumakas, and G. Paliouras, "Beyond mesh: Fine-grained semantic indexing of biomedical literature based on weak supervision," *Information Processing & Management*, vol. 57, no. 5, p. 102 282, 2020.

[216] J. A. Fries, E. Steinberg, S. Khattar, *et al.*, "Ontology-driven weak supervision for clinical entity classification in electronic health records," *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.

[217] A. M. Cohen, N. R. Smalheiser, M. S. McDonagh, *et al.*, "Automated confidence ranked classification of randomized controlled trial articles: An aid to evidence-based medicine," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 707–717, 2015.

[218] J. Schneider, L. Hoang, Y. Kansara, A. 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*, 2021.

[219] H. Kilicoglu and D. Demner-Fushman, "Bio-SCoRes: A Smorgasbord Architecture for Coreference Resolution in Biomedical Text," *PLOS One*, vol. 11, no. 3, pp. 1–38, 2016.

[220] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[221] L. Hoang, Y. Guan, and H. Kilicoglu, "Methodological information extraction from randomized controlled trial publications: A pilot study," in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, 2022.

[222] K. S. Button, J. Ioannidis, C. Mokrysz, *et al.*, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nature reviews neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[223] M. T. Haahr and A. Hróbjartsson, "Who is blinded in randomized clinical trials? a study of 200 trials and a survey of authors," *Clinical Trials*, vol. 3, no. 4, pp. 360–365, 2006.

[224] H. Saltaji, S. Armijo-Olivo, G. G. Cummings, M. Amin, B. R. da Costa, and C. Flores-Mir, "Influence of blinding on treatment effect size estimate in randomized controlled trials of oral health interventions," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–18, 2018.

[225] G. S. Doig and F. Simpson, "Randomization and allocation concealment: A practical guide for researchers," *Journal of Critical Care*, vol. 20, no. 2, pp. 187–191, 2005.

[226] Z. Lu, "Pubmed and beyond: A survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.

[227] Cochrane, *Pico ontology*, [Online; accessed 10 01, 2022], 2022. [Online]. Available: https://linkeddata.cochrane.org/pico-ontology.

[228] 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*, vol. 37, no. 2, pp. 108–119, 2004.

[229] I. Sim, S. W. Tu, S. Carini, *et al.*, "The ontology of clinical research (ocre): An informatics foundation for the science of clinical research," *Journal of biomedical informatics*, vol. 52, pp. 78–91, 2014.

[230] A. Dechartres, I. Boutron, L. Trinquart, P. Charles, and P. Ravaud, "Single-center trials show larger treatment effects than multicenter trials: Evidence from a meta-epidemiologic study," *Annals of internal medicine*, vol. 155, no. 1, pp. 39–51, 2011.

[231] E. J. Mascha and T. R. Vetter, "Significance, errors, power, and sample size: The blocking and tackling of statistics," *Anesthesia & Analgesia*, vol. 126, no. 2, pp. 691–698, 2018.

[232] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "Brat: A web-based tool for nlp-assisted text annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 102–107.

[233] G. Hripcsak and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American medical informatics association*, vol. 12, no. 3, pp. 296–298, 2005.

[234] Y. Gu, R. Tinn, H. Cheng, *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[235] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[236] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

[237] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[238] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, "Named entity recognition using bert bilstm crf for chinese electronic health records," in *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, IEEE, 2019, pp. 1–5.

[239] J. Fu, X. Huang, and P. Liu, "Spanner: Named entity re-/recognition as span prediction," in *Annual Meeting of the Association for Computational Linguistics*, 2021.

[240] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[241] N. Milosevic, C. Gregson, R. Hernandez, and G. Nenadic, "A framework for information extraction from tables in biomedical literature," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 1, pp. 55–78, 2019.

Appendix A

Annotation Guideline - Methodological characteristics of RCTs

A.1 Introduction of the project

Evidence Synthesis is the process of synthesizing information from clinical literature to translate the research findings into patient care and healthcare policy. Throughout the evidence synthesis process, a critical yet challenging step is the quality assessment of clinical studies. Quality in research can be considered through two aspects: methodological quality which concerns how rigorously a research is designed and conducted, and reporting quality which describes how transparently a piece of scientific work is reported as a publication. This research explores natural language processing (NLP) approaches to support evidence quality assessment of clinical studies. Specifically, in this project, we are developing an information extraction system of fine-grained methodological characteristics from RCTs to assist methodological quality assessment. The information items to be extracted are information in regards to how a clinical study (Randomized Control Trials) is conducted, including: what is the study design of the study? What are the blinding and randomization methods?, etc. The outcome of the project will be information extraction models that can create a structured methodological summary of a RCT from its publication. This is a guideline of the annotation study to support the project "Methodological information to assess quality of evidence from Randomized Control Trials", in which full-text RCTs will be annotated with a list of pre-defined methodological information items. This guideline is intended to provide guidance for annotators who participate in this study. The annotation will be carried out using the teamTat annotation tool. The outcome of this annotation study will be an annotated dataset that can be used as the training/testing data for machine learning models that extract methodological information from RCT publications.

A.2 Annotation tool introduction

Annotation Tool– Brat, is located at: http://ec2-3-144-241-74.us-east-2.compute.amazonaws.com/brat/. Login information for each annotator will be provided in the first training session. Each annotator will be assigned a separate folder which contains 50 annotation files, corresponding to 50 RCTs. In each annotation file, there are three sections of text: text from the title, the abstract, and the Method section of the paper. Each of them is broken down into sentences.

To open the annotation data file, go to the Brat link above, navigate to your annotation folder, open a specific annotation file. To annotate a particular information item, highlight the chunk of text that you want to annotate, a dropdown list of information items will be popped up in another window, select the appropriate information item.

A.3 List of information items

There are two types of information items:

- **Categorical information items:** are the items that you have to choose between predefined sub-categories. Based on the text in the article, the annotator needs to decide which sub-category the information item belongs to and annotate the text span that describes/supports the decision.
- **Free-text information items:** are the items that there are no predefined categorical values to choose. You rather highlight the text span that describes the information item in the text.

A.3.1 Trial Design Type

- **Definition:** this is a categorical information item which refers to how participants are assigned into different treatment groups.
- **Subcategories:**
  - **Parallel group**:
    - **Definition**: Parallel group trials allocate each participant to a single intervention for comparison with one or more alternative interventions
    - **Examples**: *"This multicenter, randomized, double-blind, placebo-controlled,**parallel-group** study compared self-administered low-dose colchicine and high-dose colchicine with placebo."*
    - **Complex cases and rule of thumb:**
      - The article doesn't explicitly mention "parallel-group" and nothing else about design type. However, it lists several treatment groups, then we can assume that it is a parallel design type. In such cases, you should highlight the names of the treatments as an annotation

of parallel trial design. E.g.: *"Young adults with elevated levels of depression symptoms and who habitually consume a poor diet were randomly allocated to* ***a brief 3-week diet intervention (Diet Group) or a habitual diet control group (Control Group)****."*

· "Placebo-controlled" only should NOT be sufficient enough to conclude that it is parallel-group trial design. So you should NOT annotate this phrase only and mark it as "Trial_Parallel_Group". However, if the trial lists several treatments including placebo, you should annotate the whole text span as an annotation of parallel trial design. E.g: *"Patients with probable laboratory-supported, probable or definite ALS were enrolled by 25 Italian centres and randomly assigned (1:1) to receive* ***intravenous rhEPO 40000IU or placebo*** *fortnightly as add-on treatment to riluzole 100mg daily for 12months."*

– **Crossover:**

* **Definition**: Cross-over trials allocate each participant to a sequence of interventions. In this design, over time, each participant receives (or does not receive) an intervention in a random sequence. The sequences should be determined a priori and the experimental units are randomized to sequences. The most popular crossover design is the 2-period, 2-treatment crossover design, with sequences AB and BA, sometimes called the 2×2 crossover design [2]. If the trial is crossover design, sometime, you will be able to annotate "Design_Crossover_Period_Treatment" information item.
* **Examples**: *"A* ***crossover*** *randomized controlled trial (RCT) for investigating the primary aim and a cross-sectional study for investigating the secondary aim of this study."*

– **Factorial:**

* **Definition**: Factorial clinical trials test the effect of two or more treatments simultaneously using various combinations of the treatments. The simplest factorial design is known as a 2x2 factorial design, whereby participants are randomly allocated to one of four combinations of two interventions (e.g. A & B). These combinations are A alone, B alone, both A and B; neither A nor B (control) [3]. If the trial is factorial design, sometime, you will be able to annotate "Design_Factorial_Factor_Treatment" information item.
* **Examples**: *"In this blinded* ***factorial*** *trial, we randomly assigned 1223 critically ill adults in 40 intensive care units (ICUs) in Canada, the United States, and Europe."*

– **N-of-1:**

* **Definition**: N-of-1 or single subject clinical trials consider an individual patient as the sole unit of observation in a study investigating the efficacy or side-effect profiles of different interventions
* **Examples**: *"The study was an* ***N-of-1*** *trial design, divided into 3 blocks of 10 weeks."*

- **Other design:**
  - **Definition**: other trial design types. Select this option if you see any text that describes how the study is designed but does not fall into any of the above subcategories. Make sure to note your thoughts and discuss with other annotators during the check-up meeting.

A.3.2 Design_Crossover_Period_Treatment

- **Definition:** If a trial has crossover design, sometimes, the trial will provide detailed characteristics of the crossover design, including number of periods and number of treatments (e.g. 2-period, 2-treatment crossover design, with sequences AB and BA, is called the 2 × 2 crossover design).
- **Example:** *"This was a randomized **3x3** crossover design study with 26 healthy overweight adults."*

A.3.3 Design_Factorial_Factor_Treatment

- **Definition:** If a trial has factorial design, sometime, the trial will provide detailed characteristics of the factorial design, including number of factors and number of treatments (e.g. 2-factor, 2-treatment factorial design, is called the 2 × 2 factorial design).
- **Example:** *"In this blinded **2-by-2** factorial trial, we randomly assigned 1223 critically ill adults in 40 intensive care units (ICUs) in Canada, the United States, and Europe."*

A.3.4 Comparative Intent

- **Definition:** refers to the intent of comparison made in a study with two or more interventions.
- **Subcategories:**
  - **Equivalence**:
    - **Definition**: An equivalence trial is designed to determine whether the response to two or more treatments differs by an amount that is clinically unimportant. This is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence level of clinically acceptable differences.
    - **Examples**: *"TWe did this randomised, phase 3, **equivalence** trial (NSABP B-39/RTOG 0413) in 154 clinical centres in the USA, Canada, Ireland, and Israel."*
  - **Non-inferiority**:

    - **Definition**: A non-inferiority trial is designed to determine whether the effect of a new treatment is not worse than a standard treatment by more than a pre-specified amount. A one-sided version of an equivalence trial.
    - **Examples**: *"We performed a multicenter,* ***non-inferiority*** *randomized trial comparing HFNT and noninvasive ventilation (NIV) in nine centers in Italy."*
  - **Superiority**:
    - **Definition**: When the aim of the study is to show that an experimental (E) treatment is superior to a control (C) treatment, the RCT is called a superiority trial and the associated statistical test is a superiority test. With a significant result, one concludes in a superiority trial that E is different in effect from C, and when the observed result is in favor of E, we conclude that E is statistically significantly better performing than C.
    - **Examples**: *"****Superiority*** *analysis was performed on the secondary outcome reductions in glucose-lowering medication"*
  - **N-of-1:**
    - **Definition**: N-of-1 or single subject clinical trials consider an individual patient as the sole unit of observation in a study investigating the efficacy or side-effect profiles of different interventions
    - **Examples**: *"The study was an* ***N-of-1*** *trial design, divided into 3 blocks of 10 weeks."*
  - **Other design:**
    - **Definition**: other trial design types. Select this option if you see any text that describes how the study is designed but does not fall into any of the above subcategories. Make sure to note your thoughts and discuss with other annotators during the check-up meeting.

A.3.5 Phase

- **Definition:** Phase describes the level of a trial required of drugs before (and after) they are routinely used in clinical practice.
- **Subcategories:**
  - **Phase 1**:
    - **Definition**: trials assess toxic effects on humans (not many people participate in them, and usually without controls)
  - **Phase 2**:
    - **Definition**: trials assess therapeutic benefit (usually involving a few hundred people, usually with controls, but not always)

- **Phase 3**:
    - **Definition**: trials compare the new treatment against standard (or placebo) treatment (usually a full randomised controlled trial). At this point, a drug can be approved for community use.
- **Phase 4**:
    - **Definition**: trials monitor a new treatment in the community, often to evaluate long term safety and effectiveness.trials monitor a new treatment in the community, often to evaluate long term safety and effectiveness.

- **Examples**: *"In this multicentre, open-label, **phase 3**, randomised controlled trial (the ENDURANCE trial; E1A11), we recruited patients aged 18 years or older."*

A.3.6 Blinding Method

- **Definition:** Blinding or masking (the process of keeping the study group assignment hidden after allocation) is commonly used to reduce the risk of bias in clinical trials with two or more study groups.

- **Subcategories:**
    - **Open Label**:
        - **Definition**: All parties are aware of the treatment the participant receives.
        - **Examples**: *"This is an **open label**, placebo controlled trial."*
        - **Complex cases and rule of thumb:**
            - The trial may not be explicitly mentioned as open-label. However, the trial might say "all participants are aware of the treatments". In such cases, the annotator should highlight the text span and assign it for the "blinding open label" information item. E.g: *"In this study, **both patients and physicians are aware of treatment allocation.**"*
    - **Single Blind**:
        - **Definition**: A single blind trial involves blinding of any one group of individuals. Usually, the subjects (patients) receiving the intervention are blinded to the intervention assignments.
        - **Examples**: *"This is a parallel-group, single-center, single-blind randomized controlled trial"*
        - **Complex cases and rule of thumb:**
            - The trial may not be explicitly mentioned as single blind. However, the trial might mention only ONE party of the object being blinded. In such cases, the annotator should annotate the text span and assign it for the "single blind" information item. E.g.: *"In this study, **only patients** are blinded."*
    - **Double Blind**:

    - **Definition**:In a double blind trial, any two groups of individuals are blinded.
    - **Examples**: *"**Double-blind**, randomized, fully remote (contactless) clinical trial of fluvoxamine vs placebo."*
    - **Complex cases and rule of thumb:**
        - The trial may not be explicitly mentioned as double blind. However, the trial might mention TWO parties of the object being blinded. In such cases, the annotator should annotate the text span and assign it for the "double blind" information item. E.g.: *"In this study, **patients and caregivers** are blinded."*
- **Triple Blind**:
    - **Definition**: Three groups of people are blinded to the intervention assignments in a triple blinded study.
    - **Examples**: *"This is a **triple-blind** randomized two-group clinical trial to evaluate the effect of Aloe Vera gel on the prevention of pressure ulcers in patients."*
    - **Complex cases and rule of thumb:**
        - The trial may not be explicitly mentioned as triple blind.. However, the trial might mention THREE parties of the object being blinded. In such cases, the annotator should annotate the text span and assign it for the "triple blind" item. E.g.: *"In this study, **patients, doctors, and investigators** are blinded."*
- **Quadruple Blind**:
    - **Definition**: Four groups of people are blinded to the intervention assignments in a triple blinded study.
    - **Examples**: *"This will be a parallel group, **quadruple blind**-randomised controlled pilot trial with an add on laboratory based study."*
    - **Complex cases and rule of thumb:**
        - The trial may not be explicitly mentioned as quadruple blind. However, the trial might mention FOUR parties of the object being blinded. In such cases, the annotator should annotate the text span and assign it for the "quadruple blind" item. E.g.: *"**Patients, intervention provider, outcome assessor and the data collection officer** will be blinded."*

A.3.7 Blinding Objects

- **Definition:** who are the people that were blinded [6].
- **Subcategories:**

- **Patients**:
  - **Definition**: patients (or participants) are the people who are recruited to take part in the trial.
  - **Examples**: *"In this study the **patients**, the trained nurse and the statistician did not know anything about the Aloe-Vera gel and placebo containers in two intervention and control groups."*
- **Care Providers (Caregivers)**:
  - **Definition**: care providers, in general, are doctors, nurses, clinical workers, pharmacists who provide health care diagnosis and treatment services, and are authorized to practice the trial. These are the people who are often involved directly in the actual implementation of the RCT (e.g. give treatments/medicines to the patients, take care of the patients during hospital stays, etc.).
  - **Examples**: *"In this study the patients, the **trained nurse** and the statistician did not know anything about the Aloe-Vera gel and placebo containers in two intervention and control groups."*
  - **Complex cases and rule of thumb:**
    - Sometimes, the article uses a more general term to describe care providers such as "clinical staff". If you are able to confirm the role of the object which is clinical related, you can annotate the object as "care providers". E.g: *"Patients were randomized by MEDUMO software, and **physicians, clinic staff**, and patients were blinded."*
- **Investigators**:
  - **Definition**:Investigators often are the researchers who are conducting and managing the trial.
  - **Examples**: *"The **investigator**, pharmacist, and trial participant were blind to group allocation."*
- **Outcomes Assessors**:
  - **Definition**: Outcomes assessors are often people who are involved in the analysis of the trial AFTER the outcomes of the trial are already collected. They are often statisticians, data analysts (however, that might not always be the case).
  - **Examples**: *"In this study the patients, the trained nurse and the **statistician** who analyzed outcome data did not know anything about the Aloe-Vera gel and placebo containers in two intervention and control groups."*
- **Other Blinded Objects**:
  - **Definition**: - It is important to understand what is the role of the object in the trial in order to categorize him/her/them into an appropriate subcategory of blinding objects.

- If you find any other objects that are being blinded in the trial but it is unclear what their roles are. Then assign them with Other Blinding Objects. Those often are described using general terms such as "staff", "site personnel" and their roles in the trial are unclear.

* **Examples**: *"**Study team members**, apart from the study pharmacist and the unblinded statistical staff, are blinded."*

A.3.8 Randomization Type

- **Definition:** is a random allocation scheme that describes how patients are assigned into different treatment groups.
- **Subcategories:**
  - **Simple Randomization**:
    * **Definition**: Randomisation based solely on a single, constant allocation ratio is known as simple randomisation [7]. In this type of randomization, there are no randomization restrictions or conditions mentioned.
    * **Examples**: *"In this study, we used simple randomization."* – note that here we also include "randomization" into the annotation if it is available.
    * **Complex cases and rule of thumb:**
      · E.g: *"After screening, patients were **randomised** to bosentan or placebo (1:1 ratio) by sequential allocation of randomisation numbers."*
  - **Block Randomization**:
    * **Definition**: Block randomization is done by creating blocks of sequences, which will ensure that the same number of participants will be allocated to the study groups within each block [8]. If it is block randomization, you should be able to annotate the "randomization block size" information item, too.
    * **Examples**: *"In this study the patients, the **trained nurse** and the statistician did not know anything about the Aloe-Vera gel and placebo containers in two intervention and control groups."*
    * **Complex cases and rule of thumb:**
      · Sometimes, the article uses a more general term to describe care providers such as "clinical staff". If you are able to confirm the role of the object which is clinical related, you can annotate the object as "care providers". E.g: *"Patients were randomized by MEDUMO software, and **physicians, clinic staff**, and patients were blinded."*
  - **Stratified Randomization**:

* **Definition**: When specific variables are known to influence the outcome, stratification of the sample is required to keep the variables (e.g., age, gender, weight, prognostic status) as similar as possible between the treatment groups [9]. If it is stratified randomization, you should be able to annotate the "randomization stratification criteria" information item, too.
* **Examples**: *"Randomization schedules were generated that were **stratified** by age (18-44, 45-54, 55-64, and 65 years)."*

– **Minimization Randomization**:

* **Definition**: Minimisation is a method of adaptive stratified sampling that is used in clinical trials. The aim of minimisation is to minimise the imbalance between the number of patients in each treatment group over a number of factors [10, 11]. Minimisation calculates the imbalance within each factor should the patient be allocated to a particular treatment group. The various imbalances are added together to give the overall imbalance in the study. The treatment group that would minimise the imbalance can be chosen directly, or a random element may be added. If it is minimization "randomization minimization criteria" information item, too.
* **Examples**: *"We used a **minimisation** algorithm to assign 62 women with early-onset pre-eclampsia (24+0 -31+6 weeks of gestation) to receive pravastatin 40 mg daily (n = 30) or matched placebo (n = 32), from randomisation to childbirth."*

– **Minimization Randomization**:

* **Definition**: other randomization types. Select this option if you see any text that describes how the randomization is conducted but does not fall into any of the above subcategories. Make sure to note your thoughts and discuss with other annotators during the check-up meeting.

A.3.9 Randomization Ratio

- **Definition:** Ratio of randomization into treatment groups.This attribute is not tied to any particular type of randomization.
- **Example:** *"We randomly assigned women aged 39-41 years, using individual randomisation, stratified by general practice, in a **1:2** ratio."– note that you should only annotate the ratio number (don't include the "ratio" phrase in the annotation)*

A.3.10 Randomization Sequence Generation Method

- **Definition:** How the randomized sequence is generated (e.g. using a computer random number generator; random number table; coin tossing; shuffling cards or envelopes; throwing dice.

- **Example:**

  *"The randomization sequence was **computer generated** by an experimenter who was not involved in recruitment."*

  *"Patients were randomly assigned (1:1:1:1) to one of three dosing regimens of quilizumab or placebo using an **interactive web response system.**" – note that you don't need to include "a" in your annotation here.*

- **Complex cases and rule of thumb:**

  – If the randomization sequence generation method is a combination of multiple devices, you should annotate the whole text span. E.g: "Clinicians entered baseline data via ***a telephone voice-activated or a secure web-based randomisation system.***"

A.3.11 Randomization Personnel

- **Definition:** refers to the person, people, organization, who is involved in creating/generating the randomization sequence.

- **Example:** *"The computer-generated sequentially numbered randomisation list (with variable block sizes) containing both allocations was pre-prepared by the **trial statistician.**"*

- **Complex cases and rule of thumb:**

  – Sometimes, randomization personnel is not a single person, but an organization/third party, you should annotate the name of the organization and assign it to the information item. E.g.: "The randomisation list was generated by **Boehringer Ingelheim Pharma GmbH & Co. KG**, Biberach an der Riss, Germany, using a validated pseudo-random number generator and a supplied seed number."

A.3.12 Randomization Block Size

- **Definition:** If the randomization type is "block randomization", the trial should provide information about the block size of randomization accordingly.

- **Example:** *"Treatments were randomly allocated using alternating block sizes of **2 and 4.**"*

A.3.13 Randomization Stratification Criteria

- **Definition:** If the randomization type is "stratified randomization", what are the criteria for stratification and what are the values of the criteria.

- **Example:** *"Randomization schedules were generated that were stratified by* ***age (18-44, 45-54, 55-64, and 65 years).*** *"*

- **Complex cases and rule of thumb:**

  - If there are multiple stratified criteria, annotate them separately. You should not include the "and" phrase in your annotation. E.g.: *"Both factorial randomisations were stratified by* ***centre*** *and* ***age (<7, 7–12, 13 years).*** *"*

A.3.14 Randomisation Minimisation Criteria

- **Definition:** If the randomization type is "minimisation randomization", what are the criteria for minimisation and what are the values of the criteria.

- **Example:** *"The system used a minimisation algorithm to achieve optimum balance for key prognostic factors by* ***world region*** *and on all the* ***other key factors within regions.*** *"*

A.3.15 Allocation Concealment Method

- **Definition:** Allocation concealment is performed when the treatment allocation system is set up so that the person enrolling participants does not know in advance which treatment the next person will get. Allocation concealment methods refer to the methods used to conceal the allocation. Such as: Sequentially labeled drug containers, sequentially labeled opaque sealed envelopes, telephone, web-based, etc.

- **Example:** *'This randomisation sequence was concealed by using* ***sequentially numbered, opaque, sealed, and stapled envelopes."***

A.3.16 Required Sample Size

- **Definition:** The number of patients based on the required sample size calculation.

- **Example:** *"Based on 80% power, an level of .05, a rate of 20% for clinical deterioration in the placebo group, a total sample size of* ***152*** *participants was required."*

- **Complex cases and rule of thumb:**

  - If the sample size information includes some specific details of the patients' characteristics (such as "children", "female", "male", you could include those details into your annotation by annotating the phrases). E.g.: *"Based on 80% power, an level of .05, a rate of 20% for clinical deterioration in the placebo group, a total sample size of* ***152 children*** *was required."*

A.3.17 Target Sample Size

- **Definition:** The target number of patients based on the required sample size.
- **Example:** *"Based on the required sample size, our target sample size is **200**."*

A.3.18 Actual Sample Size at Enrollment

- **Definition:** The number of patients who actually enrolled in the study at the beginning of the study.
- **Example:** *"Of 1337 patients screened, 834 (62%) were excluded, 322 (24%) were contacted and declined participation, and 181 (14%) were randomized and provided with study materials."*

A.3.19 Actual Sample Size at Outcome Analysis

- **Definition:** The number of patients who actually completed the study and collected data for analysis.
- **Example:** *"At the time of completion of the study, **78** participants had completed testing at baseline and Day 21."*

A.3.20 Sample Size Calculation Power Value

- **Definition:** What is the power value used to calculate required sample size.
- **Example:** *"Based on an estimated effect size of d = .80, alpha level = .01 (one-tailed as direction was hypothesised), power = **80%**, we estimated that we would require a total of n = 36 participants."* – note that you should only annotate the power value, don't include the "power =" phrase in your annotation.

A.3.21 Sample Size Calculation Alpha Value

- **Definition:** What is the alpha value used to calculate required sample size. Sometimes, this value is also provided as a significance level or p-value.
- **Example:** *"We considered a significance level of **0.05** and 80% power to detect a moderate difference (Standardized difference<0.05) of hospitalization duration across two groups. Based on these criteria, the sample size was calculated as 50 per group."* – note that you should only annotate the significant level value (don't include "significance level of" phrase in your annotation).
- **Complex cases and rules of thumbs:** Sometimes, instead of providing alpha value or p-value, the trial used "confidence interval" to infer this value. In such a case, we should annotate the whole phrase "confidence interval 95%" (not only "95%"). That way, we will know this number refers to "confidence interval" and could be used to infer the alpha value accordingly.

A.3.22 Sample Size Calculation Drop Out Rate Value

- **Definition:** The sample size estimation formula will provide a number of evaluable subjects required for achieving desired statistical significance for a given hypothesis. However in practice we may need to enroll more subjects to account for potential dropouts.
- **Example:** *"Dropout rate is expected to not exceed **15 to 20%** in the study."*

A.3.23 Settings - Multicenter/Single Center

- **Definition:** what is the setting of the study? Choose between two values: Single_center or Multi_center.
- **Subcategories:**
  - **Multicenter:**
    - **Definition:** the trial is conducted in multiple different settings/locations.
    - **Example:** *"This is a **multicenter**, randomized controlled trial."*
    - **Complex and edge cases:** Sometimes, an article doesn't explicitly indicate if it is a multi center setting. However, it mentions the number of centers/hospitals/locations where the trial is conducted. In such cases, highlight such information and annotate with the corresponding setting types. E.g.: *"**49 usual care primary care practices** in the Netherlands."*
  - **Single center:**
    - **Definition:** the trial is conducted in multiple different settings/locations.
    - **Example:** *"The study was conducted at **a level 1 trauma centre** in the Netherlands."* → *this should indicate "single_center" settings.*

A.3.24 Settings - Location

- **Definition:** Highlight the span of text that describes the location of the study. This information should be city, country, area names.
- **Example:** *"49 usual care primary care practices in the **Netherlands.**"*

A.4 Annotating rules

1. ***Minimal annotation:***

- **Rule:** Only annotate the minimum amount of text span that provides the exact information item that you are looking for. Surrounding general/non specific phrases should not be annotated.

Table A.1: Examples of minimal annotation rules

Information Item	What you SHOULD annotate	What you SHOULD NOT annotate
Sample_Size_Required	6000	6000 patients
Design_Crossover	crossover	crossover study
Sample_Size_Calculation_Alpha_Value	0.1	=0.10
Randomization_Sequence_ Generation_Method	computer generated list	a computer generated list
Sample_Size_Calculation_Power_Value	84 %	84 % power
Randomization_Ratio	1 to 1	1 to 1 allocation ratio

- **Examples:**

- **Exceptions:**

  - In some cases, if the text you want to annotate contains two components (e.g. two interventions), then you should also include the article "a" or "an" into your annotation. For example*: "Young adults with elevated levels of depression symptoms and who habitually consume a poor diet were randomly allocated to* **a brief 3-week diet intervention (Diet Group) or a habitual diet control group (Control Group)***."*

2. ***Annotating specific useful information:***

- **Rule:** Even though we try to follow the "minimal annotation" rule above, sometimes, if the surrounding text provides important/specific details about the information item, you should include those details into your annotation.

- **Examples:**

Table A.2: Examples of annotating specific useful information

Information Item	What you SHOULD annotate	What you SHOULD NOT annotate
Sample_Size_Required	6000 children	6000
Randomization_Type_Block	balanced incomplete-block	block

3. ***Annotate the same information with different values:***

- **Rule:** Sometimes, a trial can contain more than one values/implementations for a certain information item. For example: a trial can contain multiple phases. And in each phase, different blinding methods are applied or different sample size calculations are implemented. You should annotate ALL the information available.

Table A.3: Examples of annotating same information with different values

Information Item	What you SHOULD annotate	What you SHOULD NOT annotate
Sample_Size_Calculation_Alpha_Value	0.10 — 0.15	0.10
Sample_Size_Calculation_Power_Value	84 % — 70 %	84 %

- **Examples:**

4. ***Annotate information from different sections:***

- Annotate information items in the **Title**, **Abstract** and the **Body** of each paper separately. For example, if you see "blinding_type" information item in both sections **Abstract** and **Body**, you should annotate both places.

- Within the **Body** of the paper, start annotating the **Methods** section first. Then, if you cannot find certain information items in the **Methods** section, but you see the items in the **Results** or **Introduction** sections, you can annotate the information there (give priority to the Results section over the Introduction section). For example: if you cannot find "Actual Sample Size at Outcome Analysis" in the **Methods** section, but in the **Results** section, annotate it there.

- In each section (either Title, Abstract, Body), only annotate one time for the **same information item** with the **same value**. For example, if you see the trial design "parallel_group" information in multiple places in the **Methods** section in the body, only annotate the first place that you see the information.

- You need to annotate the most specific information regarding the trial. For example, it is common to see that a trial is only identified as "randomized" (which may suggest "simple randomization") in the **Title**. However, later on in the **Methods** section, you see "block randomization". In this case, you should only annotate "block randomization" in the Methods section. If you've already annotated "randomized" in the **Title**, you should delete it.

- Exception of "**Simple Randomization**": If you only see "randomized" phrase in ALL sections (Title, Abstract, Body), and nothing else in regard of randomization type, you can annotate the phrase "randomized" in ALL sections separately and assign them to "Simple Randomization".

5. ***Other rule of thumbs:***

- The annotator should not limit yourself to looking for the keywords that are often tied to a particular information item (for example: look for "open label" keyword for "open label" blinding type). Sometimes, the information item is described in a different context (for example, "no personnel is blinded during the trial" is also equivalent to "open label" blinding type).

- For information items that are meant to be numerical values (such as: sample size, power value, alpha value, etc.), it is ok to annotate the numerical values only. However, if the information items are not numerical values, you should annotate all the text span (including numerical values plus their units) as supporting text for your annotation. For example: for "Settings_multicenter" information item, you should annotate the text span "100 centers" or "100 sites" (not only "100") as your supporting text.

- To categorize blinding objects, you should consider the whole context and only put them into a specific category if the role is explicitly mentioned. If the role of the person is unclear, you can put him/her/them into the Blinding_Object_Others category instead.

- For categorical information items, if the information item is the same, but you can find multiple supporting text in different places, give priority to annotate the EXACT phrase that describes the corresponding category. For examples:

  - If you see two supporting texts "multicenter" and "100 sites" in the Methods section, both texts could be used to infer "multicenter settings". You can annotate BOTH. However, it is also ok to give priority to the phrase "multicenter" and annotate it only.
  - If you see two supporting texts "parallel group" and "treatment A, B and placebo" in the Methods section, both texts could be used to infer "parallel group design". You can annotate BOTH. However, it is also ok to give priority to the phrase "parallel group" and annotate it only.

- For free text information items, if the information item is the same, but you can find multiple supporting text in different places, you should annotate the most specific information regarding that information item. For example:

  - For Actual Sample Size at Enrollment, you find two supporting texts "100" and "50 for treatment A and 50 for treatment B", you should annotate BOTH supporting text and assign them to the information item.

Appendix B

RCT Methodological Characteristics Extraction - User Study

B.1 Description and methods

- **Description:** This is an evaluation template to assess the Information Extraction system of the RCT Methodological Characteristics. The goal is for a potential end users to evaluate if the prediction results from the system is correct or incorrect. And also if the extraction result is helpful for the end user in any other downstream task (e.g annotation, quality assessment).
- **Task:**
  - End user goes through 10 articles and fills in the evaluation table (one article is in one tab). Evaluation table glossary is provided below.
  - End user provides a summary feedback in the end about usefulness of the extraction (for example: does the end user find these extractions are helpful and can be used for any downstream relevant task?
- **Glossary:**
  - ***ID:*** Identifier number of the extracted span.
  - ***Predicted Information Item:*** the information item in our RCT Methodological Characteristics Data Model that the Information Extraction system predicts.
  - ***Predicted Text Span:*** the span of text that the Information Extraction system predicts.
  - ***Sentence:*** Sentence in which the predicted span located.

- ***Section:*** Section in which the predicted span located.
- ***Evaluator's judgement (correct or not correct):*** Evaluator indicates if the prediction is correct or incorrect.
- ***Evaluator's comment:*** This field is optional. If the judgment is incorrect, evaluator could elaborate his/her evaluation (e.g. if it is incorrect prediction, explain why do you think so if needed).

B.2 Results

Overall comments from evaluator:

"Overall, the algorithms performed better than I expected, and most errors seemed to occur when classifying details related to randomization and sample size. Given that the language describing randomization can vary substantially, this isn't a surprise and perhaps more examples would improve accuracy. There are a few things that could be improved probably quite easily. First, sometimes methods are picked up in the background or discussion that describe other studies. Restricting the screening to title, abstract, methods, results sections could improve this. Second, as I noted in a few places, there are some opposing classifications that could likely be dealt with with conditional logic. For example, if a classification is made as "Settings/Multicenter" with high confidence, a subsequent classification of "Settings/Single center' could be excluded. Another example is the Settings/Location; perhaps some location logic could be used to determine what is a broader country-level setting vs. local setting. An important next step would be to evaluate what the model did not pick up at all - there are likely relevant sentences from these articles that have methodological details important for evaluating the design and execution of each trial. I could see this being useful in several contexts. First, when extracting information for systematic reviews or meta-research projects, it is possible that this would save time so the screener could see the information in context beside the paper. Second, for general evaluation/peer review of the literature, it could help to evaluate the rigor/transparency of trials. An interesting next step could be to further classify sentences beyond just whether they reported something to whether it was implemented. For example for blinding, if it was explicitly stated that investigators or analysts were not blinded. Finally, automated extractions could be used to build a large database of trials so one could filter by randomization type, ratio, blinding, and so on if they wished to study a specific corpus of articles with certain methodological characteristics, or to track trends in methodological characteristics over time and over disciplines."