Doctoral Dissertation

A Study on Automating Meta-Analysis Statistical Analysis by Employing Natural Language Processing Techniques

Mutinda Faith Wavinya
Program of Information Science and Engineering
Graduate School of Science and Technology
Nara Institute of Science and Technology

Supervisor: Prof. Eiji Aramaki
Social Computing Lab. (Division of Information Science)

Submitted on March 15, 2023

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Mutinda Faith Wavinya

Thesis Committee:
Supervisor Eiji Aramaki
(Professor, Division of Information Science)
Taro Watanabe
(Professor, Division of Information Science)
Shoko Wakamiya
(Associate Professor, Division of Information Science)
Shuntaro Yada
(Assistant Professor, Division of Information Science)
Kongmeng Liew
(Assistant Professor, Division of Information Science)

A Study on Automating Meta-Analysis Statistical Analysis by Employing Natural Language Processing Techniques[*]

Mutinda Faith Wavinya

Abstract

Meta-analyses aggregate results of different clinical studies to assess the effectiveness of a treatment. Despite their importance, meta-analyses are time-consuming and labor-intensive as they involve reading hundreds of research articles and extracting data. The number of research articles is increasing rapidly and most meta-analyses are outdated shortly after publication as new evidence has not been included. Automatic extraction of data from research articles can expedite the meta-analysis process and allow for automatic updates when new results become available. In this research, we propose a system for automatically extracting data from research abstracts and performing statistical analysis.

First, we created a corpus consisting of 1011 PubMed abstracts of breast cancer randomized controlled trials annotated with the core elements of clinical trials: Participants, Intervention, Control, and Outcomes (PICO). We then proposed a BERT-based named entity recognition (NER) model to identify PICO information from research abstracts. After extracting the PICO information, we parse numeric outcomes to identify the number of patients having certain outcomes for statistical analysis.

The NER model extracted PICO elements with relatively high accuracy, achieving F1-scores greater than 0.80 in most entities. We assessed the performance of the proposed system by reproducing the results of an existing meta-analysis. The data extraction step achieved high accuracy, but the statistical

[*] Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, , .

analysis step achieved low performance because abstracts sometimes lack all the required information.

In this work, we proposed a system for automatically extracting PICO information from research abstracts for the purpose of performing meta-analysis statistical analysis. We evaluated the performance of the system by reproducing an existing meta-analysis and the system achieved a relatively good performance, though more substantiation is required.

Keywords:

Evidence-Based Medicine (EBM), Automatic Meta-analysis, Natural Language Processing (NLP), Randomized Controlled Trial (RCT), Named Entity Recognition (NER)

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Evidence-based medicine (EBM) is an approach where doctors and health care professionals use the best available research evidence to guide them in making clinical decision about the care of patients [49]. EBM involves incorporating individual clinical expertise with the best available external evidence such as relevant clinical research literature [49]. Meta-analyses are one of the essential tools in EBM and clinical and health policy decision-making because they provide the highest form of medical evidence [23, 15]. A meta-analysis is a type of a quantitative study that combines the results of different studies that are all focused on same disease, treatment, or outcome to determine if a treatment is effective or not. Regardless of their importance, meta-analyses tend to be time-consuming, labor-intensive, and expensive as they require domain experts to manually search, read, and extract data from hundreds of research articles written in unstructured natural language. The number of research articles is increasing exponentially and it is becoming almost impossible to keep up with the high number of biomedical literature [4]. For instance, a recent study showed that more than 50,000 research articles related to the COVID-19 pandemic have been published and more articles are being published every day [60]. The large number of research articles increases the time required to conduct a meta-analysis. Previous research showed that on average it takes about 67 weeks, from registration to publication, to finalize a meta-analysis [8]. This poses a challenge for practitioners in the infectious disease field where informed decisions have to be made promptly. Further, most meta-analyses are outdated shortly after publication as they have not incorporated new evidence which might alter the results [51].

Automatic meta-analysis systems have the benefit of reducing the time-taken in conducting a meta-analysis so as to help in timely dissemination of medical evidence and allow for automatic updates when new evidence becomes available. According to surveys on automation of meta-analysis, different strategies for automating the various meta-analysis stages (searching the databases for relevant literature, screening, data extraction, and statistical analysis) have been proposed [30, 39]. Marshall and Wallace [39] suggests that systems for searching literature, identifying randomized controlled trials (RCTs), and screening articles have attained a good performance and are ready for use. The systems for the data extraction and statistical analysis, on the other hand, are still not readily available. The scarcity of publicly available corpora, which are usually expensive to create, is one barrier to the development of high-performance systems.

Techniques for extracting core elements of clinical trials, i.e., Participants, Intervention, Control, and Outcomes (PICO) from research abstracts and full-text articles have been widely studied [30]. Although various methods for extracting PICO information from research articles have been proposed, fewer attempts have been made to extract detailed information for the outcomes, especially numeric texts identifying the number of patients having certain outcomes [48, 55]. Extraction of numeric texts is important for statistical analysis to determine the effectiveness of the intervention. Summerscales et al. [55] used conditional random field-based approach to extract various named entities including treatment groups, group sizes, outcomes, and outcome numbers from research abstracts. Their annotations are however less extensive and the corpus is not publicly available for reproducibility. Pradhan et al. [48] developed a Web application for extracting data from ClinicalTrials.gov, a clinical trials database. Although ClinicalTrials.gov is an important source of clinical trials data, it has a small number of studies and mainly focuses on clinical trials in the United States [48].

The goal of this work is to provide a system that automates data extraction in order to support meta-analysis statistical analysis. To achieve this, first we create a publicly corpus annotated with the core components of clinical trials, i.e., Participants, Intervention, Control, and Outcomes (PICO). We annotate in detail numeric texts especially those that identify the number of participants having certain outcomes. The annotation of the numeric texts is important for statistical analysis to determine the overall effect of an intervention. Currently, the corpus consists of 1011 research abstracts extracted from the PubMed database. The abstracts are of randomized controlled trials (RCTs) related to breast cancer, which is one of the leading causes of deaths in the world[1]. We focus on RCTs as they are considered the gold standard for clinical research methods.

This research utilizes the current state-of-the-art natural language process-

[1] https://www.who.int/news-room/factsheets/detail/cancer

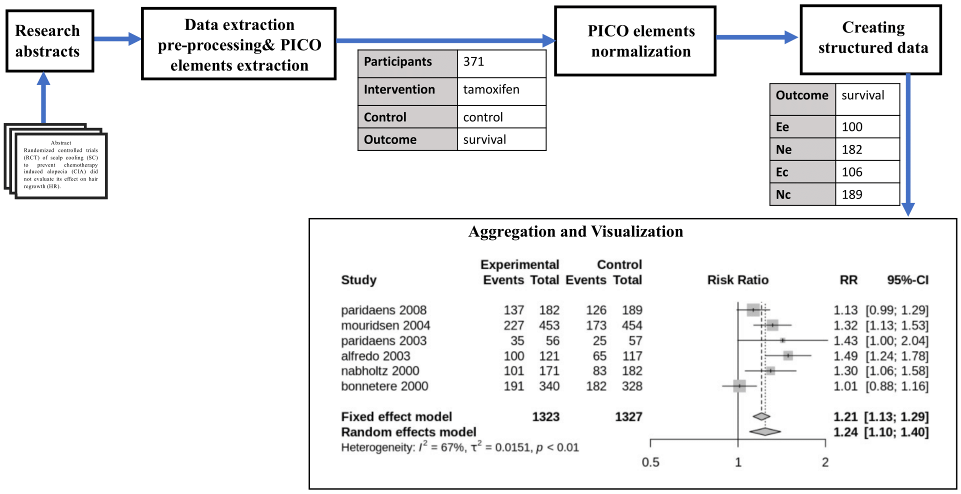

Research abstracts
Abstract
Randomized controlled trials (RCT) of scalp cooling (SC) to prevent chemotherapy induced alopecia (CIA) did not evaluate its effect on hair regrowth (HR).
Data extraction pre-processing& PICO elements extraction
Participants 371
Intervention tamoxifen
Control control
Outcome survival
PICO elements normalization
Creating structured data
Outcome survival
Ee 100
Ne 182
Ec 106
Nc 189
Aggregation and Visualization
Study
Experimental Events Total
Control Events Total
Risk Ratio
RR 95%-CI
paridaens 2008 137 182 126 189 1.13 [0.99; 1.29]
mouridsen 2004 227 453 173 454 1.32 [1.13; 1.53]
paridaens 2003 35 56 25 57 1.43 [1.00; 2.04]
alfredo 2003 100 121 65 117 1.49 [1.24; 1.78]
nabholtz 2000 101 171 83 182 1.30 [1.06; 1.58]
bonnetere 2000 191 340 182 328 1.01 [0.88; 1.16]
Fixed effect model 1323 1327 1.21 [1.13; 1.29]
Random effects model 1.24 [1.10; 1.40]
Heterogeneity: I^2 = 67%, τ^2 = 0.0151, p < 0.01
0.5
1
2

Figure 1: Proposed system architecture

ing (NLP) models to extract PICO information from research abstracts. We use abstracts because they are easily accessible and they provide a concise summary of the full-text article especially the main results. The proposed system (shown in Figure 1) performs various steps including extracting data from research abstracts, parsing numeric outcomes to identify the number of patients having specific outcomes, converting extracted data into a structured format for statistical analysis, and visualizing the results. We assess the performance of the proposed system by using it to reproduce the results of an existing meta-analysis. The results show potential in automating the tasks and hope to increase interest in research on automating the entire integrated meta-analysis process.

1.2 Objectives

One of the motivations of this research is the scarcity of publicly available corpora to train models for automatic extraction of PICO information from RCT literature. Most of the existing corpora are not publicly available. Furthermore, most of the existing corpora lack detailed annotations especially annotation of numeric values which are necessary for meta-analysis statistical analysis. To address this gap, we created a publicly available corpus consisting of 1011 research abstracts with detailed annotations of PICO information. Extraction of PICO information at a sufficient level of detail is crucial for the progress on automatic meta-analysis systems.

A second motivation is that previous studies do not go beyond PICO information extraction. In order to determine the effectiveness of an intervention/treatment, it is necessary to extract detailed information such as the number of participants who experienced certain outcomes. In this research, in addition to extracting PICO information from research abstracts, we parse numeric values to identify the number of patients who experienced specific outcomes for the purpose of statistical analysis.

The objectives of this research therefore include:

1. Create a publicly available corpora with detailed annotations of the core elements of clinical trials, i.e., Participants (P), Intervention (I), Control (C), and Outcomes (O) (Chapter 2).

2. Develop a model for automatic extraction of PICO elements from RCT research abstracts by utilizing natural language processing techniques (Chapter 3).

3. Transform the extracted PICO information into structured format by parsing numeric texts to identify the number of patients who experienced particular outcomes for the purpose of statistical analsyis (Chapter 4).

4. Evaluate our system by replicating the results of existing meta-analyses (Chapter 4).

1.3 Outline

The rest of the dissertation is structured as follows: Chapter 2 describes the creation of the PICO corpus including the source of the data, detailed annotation guidelines, and the corpus statistics. Chapter 3 describes natural language processing techniques for extraction of PICO information from RCT research abstracts. Chapter 4 outlines the approaches for PICO information normalization. We also explain the process of parsing numeric texts and converting the extracted information into structured format for statistical analysis. Chapter 5 concludes the dissertation, discusses the limitations and possible future directions.

2. A Publicly Available PICO Corpus to Support Automatic Data Extraction from RCT Literature

2.1 Background and related work

There are few manually annotated corpora to support automatic extraction of core PICO elements from clinical trials RCT studies. Most of these corpora, however, are not publicly available. Furthermore, most of the annotations are not detailed enough to support information extraction for meta-analysis. This is because they largely lack detailed annotation of the PICO elements and especially annotation of numerical values which are necessary for meta-analysis statistical analysis. The existing corpora annotate PICO elements both at the sentence-level and entity-level. Even though the sentence-level annotations can be used for tasks such as question answering and document retrieval, they are not sufficient for meta-analysis which requires more fine-grained annotations of the PICO elements.

2.1.1 Sentence-level annotations

Jin and Szolovits [26] extracted abstracts from the PubMed database whose study type was RCT. The sentences in the abstract were labeled with one of seven labels: participants (P), intervention (I), outcomes (O), aim (A), method (M), results (R), and conclusion (C). The corpus consists of 24,668 abstracts each of which contain at least one of the P/I/O labels. There are 21,198 abstracts with P label, 13,712 with I label, and 20,473 with O label.

Kim et al. [33] annotated 1,000 MEDLINE structured and unstructured abstracts. The sentences in the abstracts are labeled with seven labels which include background, population, intervention, outcome, study design, and other. The sentences can be assigned multiple classes.

Boudin et al. [9] created a dataset with about 15,000 PubMed abstracts. They extracted structured abstracts and auto-labeled the sentences with P, IC, and O labels. Structured abstracts contain distinctive sentence headings, and they selected sentences marked with corresponding PICO elements. Since all abstracts are not structured, the usefulness maybe limited to structured abstracts context only.

Demner-Fushman and Lin [18] extracted 633 abstracts from the MEDLINE database. The 633 abstracts were annotated with P, condition, and IC labels on a sentence level. One hundred of the 633 abstracts were annotated with population, problem, intervention, and comparison on an entity level.

Zhao et al. [66] extracted medical abstracts from journal websites. They randomly selected 2,000 sentences from the abstracts and annotated them with patient, intervention, result, study design, and research goal labels. In addition to the sentence level annotations, they did entity level annotations where they annotated the gender, age, race, condition, intervention, and study design.

Chabou and Iglewski [12] created a corpus containing about 3,000 abstracts. The sentences in the abstracts were annotated manually and automatically. The automatically labeled sentences relied on structured abstracts with explicitly mentioned headings, that is, patient, intervention, and main outcome.

Chung [13] extracted both structured and unstructured RCTs abstracts from PubMed. They filtered abstracts on asthma, angina, breast cancer, diabetes, prostate cancer, heart failure erectile dysfunction, and cardiovascular. They annotated sentences both manually and automatically by using the headings in structured abstracts. The sentences in the unstructured abstracts were labeled as one of aim, method, results, and conclusion. Sentences in both the structured and unstructured abstracts were further annotated as P, IC, O sentences. The corpus contains more than 344 abstracts. Other than Jin and Szolovits [26] and Kim et al. [33], the rest of the corpora are not publicly available.

2.1.2 Entity-level annotations

De Bruijn et al. [17] retrieved 88 full-text articles from five medical journals; *JAMA*, *PLoS Clinical Trials*, *Annals of Internal*, *NEJM*, *Lancet*, and *Lancet Medicine*. They annotated the sentences in each abstract according to its section (or subsection), i.e., abstract, methods, and so on. They also performed entity level annotation and labeled various PICO elements which included eligibility, intervention and control treatments, intervention and control features (medication dosage, frequency, route, etc), study start and end dates, primary and secondary outcomes, funding, and so on.

Brassey et al. [10] created a corpus consisting of abstracts of 1,750 RCTs randomly selected from PubMed/MEDLINE database. They annotated PIC elements in the title and abstracts.

Kang et al. [31] retrieved 170 RCT research publications from the MEDLINE database. They annotated entities which included population, intervention & control, and outcome. They also classify each of the abstracts into one of treatment, prevention, diagnosis, prognosis, and etiology categories.

Bui et al. [11] created a dataset consisting of 48 full-text research articles which were RCTs. They annotated the texts on both sentence and entity level. They labeled with labels which identified the N (sample/group size), population (P), study arm (intervention or control, IC), and outcome (O).

Kiritchenko et al. [34] developed a dataset containing 182 full-text articles. They annotated 21 entities such as study dates, treatment, control, treatment dosage, treatment frequency, primary outcomes, secondary outcomes, outcome time point, funding organization, grant number, and so on.

Summerscales et al. [55] created a corpus consisting of 263 RCT abstracts of British Medical Journal (BMJ) extracted through PubMed. They annotated the treatment groups, outcomes, group sizes, and outcome numbers. Their work is close to our study as they attempted to identify outcome numbers and group sizes for the purpose of calculating summary statistics, such as absolute risk reduction. The annotations are however less extensive and the corpus is not publicly available.

Since constructing large corpora is expensive, Wallace et al. [59] employed a distant supervision approach to create a large corpora consisting of full-text articles. They also manually annotated 133 articles for evaluation. Although distant supervision is a cheap way to construct large datasets, the dataset's quality might be low.

Nye et al. [43] developed the EBM-NLP corpus with the aim of facilitating development of automatic extraction of PICO information from RCT abstracts. The corpus consists of about 5,000 abstracts of RCTs mostly related to cardiovascular diseases, cancer, and autism. The abstracts were annotated by crowdsourcing through Amazon Mechanical Turk and a small part (200 abstracts) was done by medical professionals. The corpus contains fine-grained annotation of

PICO elements compared to the other previous corpora. They however do not annotate numeric texts that identify the number of participants who had certain outcomes. Since most of the previously developed corpora are not publicly available (as shown in Table 1), the EBM-NLP corpus is one of the largest publicly available corpora.

Table 1: Existing PICO annotated corpora

	Publication	**Description**	**Size**	**Availability**
Sentence-level annotations	Jin and Szolovits [26]	P, I, O, aim, method, results, conclusion	24,668 abstracts	Yes
	Kim et al. [33]	background, P, I, O, study design	1,000 abstracts	Yes
	Boudin et al. [9]	P, IC, O	15,000 abstracts	No
	Demner-Fushman and Lin [18]	P, condition, IC	633 abstracts	No
	Zhao et al. [66]	P, I, result, study design, research goal	2,000 sentences	No
	Chabou and Iglewski [12]	P, I, O	3,000 abstracts	No
	Chung [13]	aim, method, results, conclusion, P, IC, O	344 abstracts	No
Entity-level annotations	De Bruijn et al. [17]	P, I, C,O, eligibilty, dosage, duration, funding, e.t.c.	88 full-text articles	No
	Brassey et al. [10]	P, I, C	1,750	No
	Kang et al. [31]	P, IC,O	170 abstracts	No
	Bui et al. [11]	N (sample/group size), P, IC, O	48 full-text articles	No
	Kiritchenko et al. [34]	study dates, treatment, control, dosage, frequency, outcomes, funding , e.t.c	182 full-text articles	No
	Summerscales et al. [55]	treatment groups, outcomes, group sizes, outcome numbers	263 abstracts	No
	Nye et al. [43]	P, I, C, O	5,000 abstracts	Yes

2.2 Corpus annotation

2.2.1 Dataset collection

The corpus in this work consists of abstracts extracted from PubMed[2], which is a free search engine that provides access to the MEDLINE database[3]. MEDLINE is one of the largest bibliographic databases maintained by the the U.S. National Library of Medicine (NLM) and is considered an authoritative source of clinical literature [18]. It indexes over 29 million references to journal articles in biomedical and life sciences. Each MEDLINE citation contains basic meta-data such as the article title, authors, affiliations, publication date, abstract text, and so on.

For this research, we extracted English research abstracts related to breast cancer and whose study type is RCT. Abstracts which were meta-analyses or systematic-reviews were excluded. This was achieved by using keywords such as "breast cancer," "randomized controlled," "randomised controlled," "meta-analysis," and "systematic review."

The abstracts were extracted from the PubMed database using the Bio.Entrez package[4] which provides access to several National Center for Biology Information (NCBI) databases such as PubMed, GenBank, and so on. The Bio.Entrez package has various functions such as *esearch* which retrieves the PMID's (PubMed unique indentifier number) of documents related to a search query (keywords). The abstracts were extracted in XML format and we used the Beautiful Soup library[5] to extract data from the XML files. The XML documents include many tags such as the *Journal, PubDate, Author, AffiliationInfo, Title, AbstractText*, and so on. The PubMed database mainly provides free access to abstracts, and to access full-texts mostly one has to retrieve them from external links (some of which are not open-access).

2.2.2 Annotation process

The extracted research abstracts were manually annotated. The annotation was performed using BRAT, an open-source web annotation tool [53]. The annota-

[2] https://pubmed.ncbi.nlm.nih.gov/

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

[4] https://biopython.org/docs/1.75/api/Bio.Entrez.html

[5] https://beautiful-soup-4.readthedocs.io/en/latest/

tors were asked to read and label text spans that identify the PICO elements. The annotators were required to annotate the shortest possible phrase which can be considered as the building block for the PICO elements. For each PICO category, we developed sub-categories to capture detailed information within each category. The PICO label hierarchy is shown in Figure 2. Figure 3 shows examples of abstracts with PICO elements annotated. In total, we annotated 26 sub-categories (entities) which are described below.

- **Participants (P)**

  We annotate text snippets that describe the characteristics of the participants in a study. We annotate eight entities in the participants category that include:

  - **Total participants**: the total number of participants in the study.

    Examples:

    - *<total-participants>One hundred and seventy-six </total-participants>* metastatic breast cancer patients were randomised to receive docetaxel (100 mg m(-2)) every 3 weeks or 5-fluorouracil+vinorelbine.
    - We randomly assigned *<total-participants >2972</ total-participants>* women, aged 30-70 years, with surgically removed stage I breast cancer or ductal carcinoma in situ to receive 5 years either fenritinide orally or no treatment.

  - **Intervention participants**: the number of participants in the intervention group.

    Examples:

    - *<intervention-participants >Eighty-six</intervention-participants>* patients received 516 cycles of docetaxel; 90 patients received 476 cycles of 5-fluorouracil+vinorelbine
    - Patients were randomized to undergo (*<intervention-participants >10</intervention-participants>*) or not undergo (10) concomitant resection.

  - **Control participants**: the number of participants in the control group.

  Examples:

  - Eighty-six patients received 516 cycles of docetaxel; *<control-participants >90</control-participants>* patients received 476 cycles of 5-fluorouracil+vinorelbine
  - Patients were randomized to undergo (10) or not undergo (*<control-participants>10</ control-participants>*) concomitant resection.

- **Age**: age of the participants.

  Examples:

  - Fifty-three white women, aged *<age>36 to 55 years</age>*, with breast cancer and artificially induced menopause were stratified
  - Twenty consecutive women (age range *<age>43-61 yrs</age>*)

- **Eligibility**: the selection criteria (inclusion or exclusion) for study participants.

  Examples:

  - One hundred and seventy-six *<eligibility>metastatic breast cancer patients</eligibility>* were randomised to receive docetaxel (100 mg m(-2)) every 3 weeks or 5-fluorouracil+vinorelbine
  - Fifty-three white women, aged 36 to 55 years, *<eligibility>with breast cancer and artificially induced menopause</eligibility>* were stratified

- **Ethnicity**: the racial/ethnic group of the participants

  Examples:

  - Fifty-three *<ethnicity>white</ethnicity>* women, aged 36 to 55 years, with breast cancer and artificially induced menopause were stratified.
  - Safety and efficacy results from *<ethnicity>Asian</ethnicity>* patients in BOLERO-2 are reported.

- **Condition**: although breast cancer is the main condition, some studies focus on conditions associated with breast cancer such as hair loss, bone loss, depression, pain, and vomiting.

  Examples:

        - Effect of tamoxifen on *<condition>venous thromboembolic events</condition>* in a breast cancer prevention trial.
        - Effect of shugan liangxue compound for relieving *<condition>hot flashes</condition>* in breast cancer patients

    - **Location**: the location where the study was conducted.

      Examples:

        - A total of 703 women from a Basic Health Area of *<location>Barcelona</location>*, and with a mobile phone number registered, were invited to participate in a breast cancer screening programme
        - This study was performed at the University of Florence (*<location>Florence, Italy</location>*).

- **Intervention and Control (IC)**

  There are only two entities in this category.

    - **Intervention**: the intervention treatment which includes the medications (e.g., drugs, chemicals), diagnostic tests (e.g., screening), therapy, and lifestyle changes (e.g., exercise, diet).

      Examples:

        - *<intervention>Docetaxel</intervention>* vs 5-fluorouracil plus vinorelbine in metastatic breast cancer after anthracycline therapy failure.
        - Within each stratum, patients were randomly assigned to receive *<intervention>risedronate</intervention>* (n = 27) or placebo (n = 26)

    - **Control**: control treatment which is the alternative to the main intervention.

      Examples:

        - Docetaxel vs *<control>5-fluorouracil</control>* in metastatic breast cancer after anthracycline therapy failure.
        - Within each stratum, patients were randomly assigned to receive risedronate (n = 27) or *<control>placebo</control>* (n = 26)

- **Outcomes (O)**:

  We annotate the outcome measures (primary and secondary end-points), outcomes that were measured, and intervention events and control events. Intervention events and control events refer to the number of participants who experienced a particular outcome in the intervention group and control group respectively. We aim to capture detailed information for the outcomes especially the numeric texts that identify the number of participants who experienced a particular outcome. In meta-analysis statistical analysis, these numeric texts are important for calculating summary statistics to ascertain the effectiveness of the intervention.

  In the annotation of outcomes and their events, we mainly consider two types of outcomes, i.e., *binary outcomes* and *continuous outcomes.* Binary outcomes take two values such as the treatment was successful or failed, or survival (alive or dead). Continuous outcomes are not as straightforward as binary outcomes. Continuous outcomes such as pain are measured on a numerical scale (for instance, pain scores on a scale of 0 and 10). Continuous outcomes are usually measured at different time points (such as at baseline and at followup) and the results reported as mean, standard deviation, median, or quartiles.

  We created labels to capture the various types of numeric texts in the intervention and control groups. We use "*iv*," "*cv*," "*bin*," and "*cont*" to represent intervention group, control group, binary outcome, and continuous outcome, respectively. In addition, binary outcomes numeric texts tend to be absolute values or percentage values. We use "*abs*" and "*percent*" to label absolute and percentage values respectively. Further, for the continuous outcomes, we also designed labels to capture the different types of numeric texts. We use "*mean*," "*sd*," "*median*," "*q1*," and "*q3*" to represent mean, standard deviation, median, first quartile, and third quartile respectively. In total, we have 16 entities for the outcomes category.

  - Outcome measure examples:

    - The primary end point was the *<outcome-measure>incidence of contralateral breast cancer</outcome-measure>* 7 years after randomization.

- *<outcome-measure>Overall survival</outcome-measure>* was a secondary endpoint.

- Binary outcome examples:
  - *<iv-bin-abs>Four</iv-bin-abs>* patients in the intervention group and *<cv-bin-abs>two</cv-bin-abs>* in the control group were *<outcome>*lost to follow-up*</outcome>*.
  - After 20 years, *<iv-bin-percent>50.4%</iv-bin-percent>* of the women in the XRT group *<outcome>*died*</outcome>* compared with *<cv-bin-percent>54.0%</cv-bin-percent>* in the non-XRT group.
- Continuous outcome examples:
  - The *<outcome>median PFS</outcome>*of test group was significantly longer than that of control group, *<iv-cont-median>39.1 weeks</iv-cont-median>* vs *<cv-cont-median>14.0 weeks</cv-cont-median>*.
  - *<outcome>Depression scores</outcome>* at follow-up were significantly lower in the exercise group (M = *<iv-cont-mean>4.78</iv-cont-mean>* SD = *<iv-cont-sd>3.56</iv-cont-sd>*) compared to the control group (M= *<cv-cont-mean>6.91</cv-cont-mean>*, SD =*<cv-cont-sd>5.86</cv-cont-sd>*).

2.3 Corpus statistics

The corpus contains 1011 manually annotated abstracts. The abstracts were annotated by two annotators. One of the annotators was hired from an annotation company and has extensive experience annotating medical documents and the second annotator is one of the authors. The first annotator annotated all the abstracts while the second annotator annotated 45% of the abstracts. The inter-annotator agreement was calculated based on Cohen Kappa and achieved a score of 0.72. Cohen Kappa, κ is calculated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{1}$$

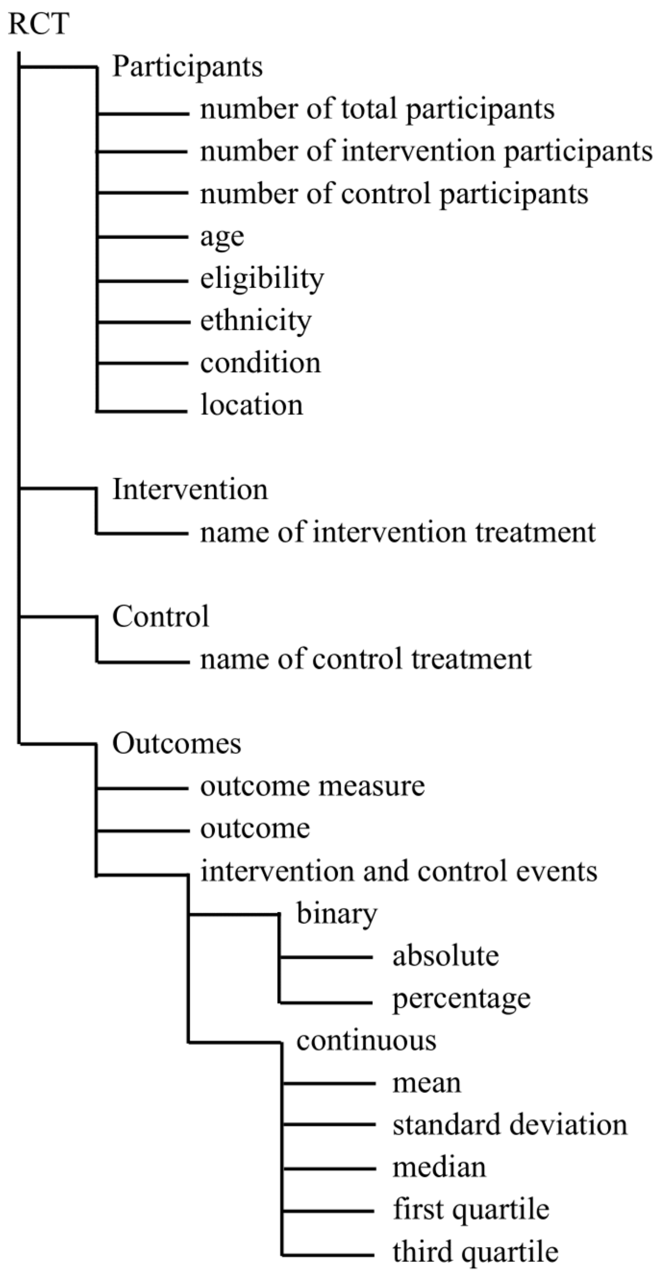
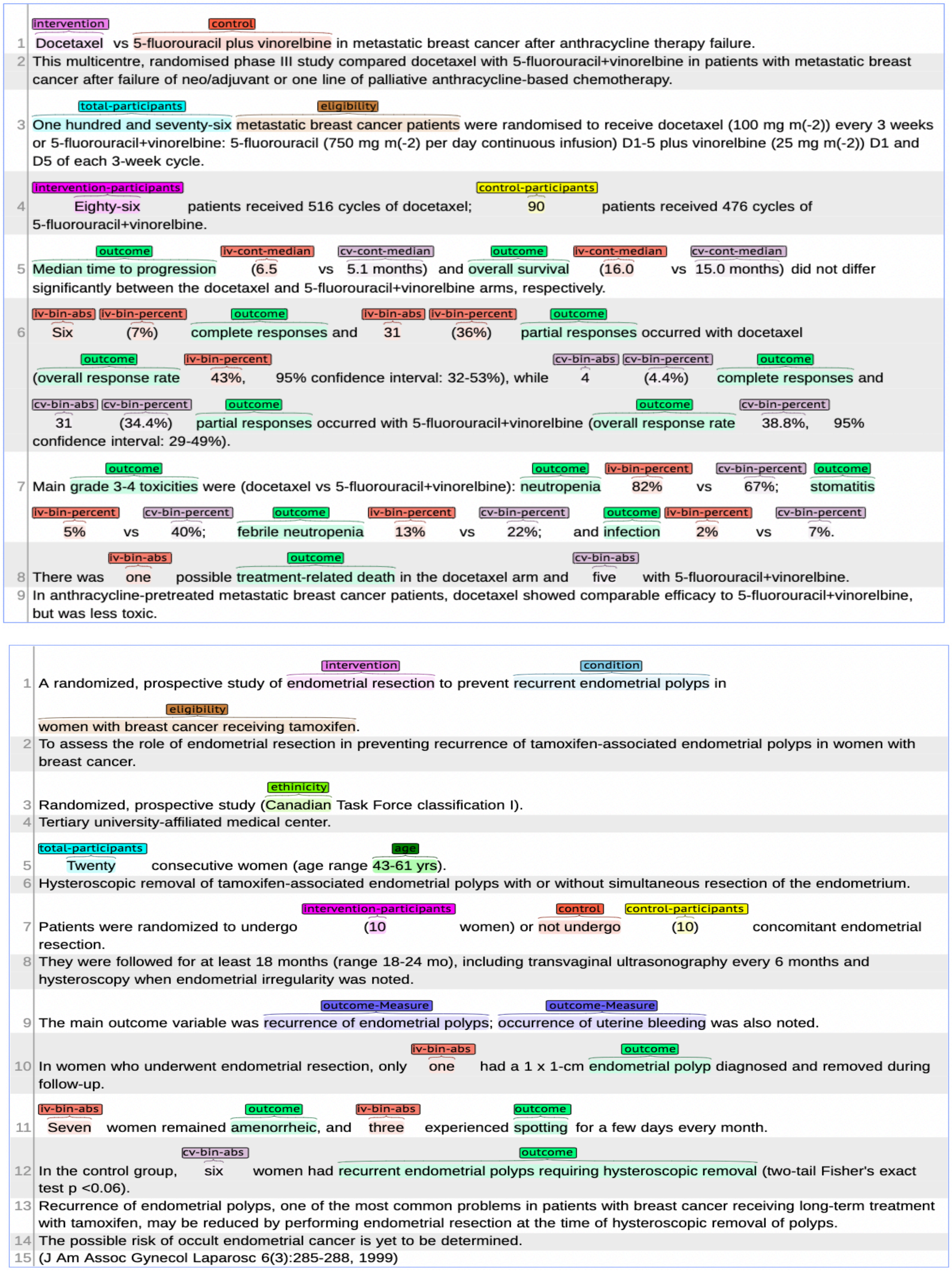


Figure 2: Annotation label hierarchy



1 Docetaxel vs 5-fluorouracil plus vinorelbine in metastatic breast cancer after anthracycline therapy failure.

2 This multicentre, randomised phase III study compared docetaxel with 5-fluorouracil+vinorelbine in patients with metastatic breast cancer after failure of neo/adjuvant or one line of palliative anthracycline-based chemotherapy.

3 One hundred and seventy-six metastatic breast cancer patients were randomised to receive docetaxel (100 mg m(-2)) every 3 weeks or 5-fluorouracil+vinorelbine: 5-fluorouracil (750 mg m(-2) per day continuous infusion) D1-5 plus vinorelbine (25 mg m(-2)) D1 and D5 of each 3-week cycle.

4 Eighty-six patients received 516 cycles of docetaxel; 90 patients received 476 cycles of 5-fluorouracil+vinorelbine.

5 Median time to progression (6.5 vs 5.1 months) and overall survival (16.0 vs 15.0 months) did not differ significantly between the docetaxel and 5-fluorouracil+vinorelbine arms, respectively.

6 Six (7%) complete responses and 31 (36%) partial responses occurred with docetaxel (overall response rate 43%, 95% confidence interval: 32-53%), while 4 (4.4%) complete responses and 31 (34.4%) partial responses occurred with 5-fluorouracil+vinorelbine (overall response rate 38.8%, 95% confidence interval: 29-49%).

7 Main grade 3-4 toxicities were (docetaxel vs 5-fluorouracil+vinorelbine): neutropenia 82% vs 67%; stomatitis 5% vs 40%; febrile neutropenia 13% vs 22%; and infection 2% vs 7%.

8 There was one possible treatment-related death in the docetaxel arm and five with 5-fluorouracil+vinorelbine.

9 In anthracycline-pretreated metastatic breast cancer patients, docetaxel showed comparable efficacy to 5-fluorouracil+vinorelbine, but was less toxic.

1 A randomized, prospective study of endometrial resection to prevent recurrent endometrial polyps in women with breast cancer receiving tamoxifen.

2 To assess the role of endometrial resection in preventing recurrence of tamoxifen-associated endometrial polyps in women with breast cancer.

3 Randomized, prospective study (Canadian Task Force classification I).

4 Tertiary university-affiliated medical center.

5 Twenty consecutive women (age range 43-61 yrs).

6 Hysteroscopic removal of tamoxifen-associated endometrial polyps with or without simultaneous resection of the endometrium.

7 Patients were randomized to undergo (10 women) or not undergo (10) concomitant endometrial resection.

8 They were followed for at least 18 months (range 18-24 mo), including transvaginal ultrasonography every 6 months and hysteroscopy when endometrial irregularity was noted.

9 The main outcome variable was recurrence of endometrial polyps; occurrence of uterine bleeding was also noted.

10 In women who underwent endometrial resection, only one had a 1 x 1-cm endometrial polyp diagnosed and removed during follow-up.

11 Seven women remained amenorrheic, and three experienced spotting for a few days every month.

12 In the control group, six women had recurrent endometrial polyps requiring hysteroscopic removal (two-tail Fisher's exact test p <0.06).

13 Recurrence of endometrial polyps, one of the most common problems in patients with breast cancer receiving long-term treatment with tamoxifen, may be reduced by performing endometrial resection at the time of hysteroscopic removal of polyps.

14 The possible risk of occult endometrial cancer is yet to be determined.

15 (J Am Assoc Gynecol Laparosc 6(3):285-288, 1999)

Figure 3: Abstracts with PICO elements annotated

where P_o is the relative agreement between the annotators and P_e is the hypothetical probability that the annotators agree by chance. The cohen Kappa score is a number between -1 and 1 where high value indicates high agreement between the annotators and lower score means chance agreement [14].

Currently the corpus has 17,739 entities and the frequencies of the annotated entities are shown in Table 2. The most frequent entity type is *outcome*, which comprises about 28% of all the annotations. Continuous outcomes quartile values (*q1* and *q3*) are the least frequent entity types. Table 2 also shows the number of abstracts containing each of the entities. The entities found in most abstracts are *intervention*, *outcome*, and *control* which are in 100%, 97%, and 94% of the abstracts, respectively. Most abstracts do not contain continuous outcomes values (*mean*, *median*, *sd*, *q1*, *q3*), *ethnicity*, and *location*.

Annotator disagreement: Annotator disagreements were mainly found in the *outcome* and *eligibility* entities. The *outcome* and *eligibility* entities mostly contain more than two words. The main source of disagreement in the annotation was mainly due to the annotators identifying different limits of the start and end spans. The disagreement for the other entities was lower since the entities could be identified by one or two words. Numerical entities had the fewest annotation disagreements.

2.4 Conclusion

In this work, we presented a publicly available corpus consisting of 1011 abstracts related to breast cancer RCTs. The corpus provides detailed annotation of PICO elements. For the outcomes we especially annotate in detail numeric texts that identify the number of participants having certain outcomes. This is important for statistical analysis to determine the effectiveness of a treatment. The corpus will facilitate NLP research on automatic information extraction from biomedical literature and contribute towards evidence-based medicine. Since the corpus consists of breast cancer related abstracts, one of the future works is to extend it to include other types of cancer. Since most of the intervention treatments, outcomes, and outcome measures are common across different types of cancer, the corpus can be extended using various machine learning techniques. The corpus is publicly available at `https://github.com/sociocom/PICO-Corpus`.

Table 2: Corpus statistics: The frequency of each entity (sub-category) and the number of abstracts in which each entity is found.

Sub-category	Tag count	Number of abstracts
Participants (P)		
total-participants	1094 (6%)	847 (84%)
intervention-participants	887 (5%)	674 (67%)
control-participants	784 (4%)	647 (64%)
age	231 (1%)	210 (21%)
eligibility	925 (5%)	864 (85%)
ethinicity	101 (1%)	83 (8%)
condition	327 (2%)	321 (32%)
location	186 (1%)	168 (17%)
Intervention & Control (IC)		
intervention	1067 (6%)	1011 (100%)
control	979 (6%)	949 (94%)
Outcomes (O)		
outcome	5053 (28%)	978 (97%)
outcome-measure	1081 (6%)	413 (41%)
iv-bin-abs	556 (3%)	288 (28%)
cv-bin-abs	465 (3%)	258 (26%)
iv-bin-percent	1376 (8%)	561 (55%)
cv-bin-percent	1148 (6%)	520 (51%)
iv-cont-mean	366 (2%)	154 (15%)
cv-cont-mean	327 (2%)	154 (15%)
iv-cont-median	270 (2%)	140 (14%)
cv-cont-median	247 (1%)	133 (13%)
iv-cont-sd	129 (1%)	69 (7%)
cv-cont-sd	124 (1%)	67 (7%)
iv-cont-q1	4 (0%)	3 (0%)
cv-cont-q1	4 (0%)	3 (0%)
iv-cont-q3	4 (0%)	3 (0%)
cv-cont-q3	4 (0%)	3 (0%)

3. Extracting PICO Elements from RCT Literature

3.1 Background and related work

Information extraction is a task whose objective is to extract information from data. Named Entity Recognition (NER) is an information extraction task aiming to extract specific entities from text and classify them into predefined categories, such as disease, medication, symptom, etc. In NER, given an input sequence $x = (x_1, x_2, ..., x_n)$, the task is to predict a predict label sequence $y = (y_1, y_2, ..., y_n)$, where n is the number of words in the sequence. NER is a common natural language processing (NLP) task and various approaches have been studied over time. Previous studies on extraction of PICO elements have proposed various models including rule-based, Support Vector Machines (SVM), Hidden Markov Models (HMM), Conditional Random Fields (CRF), and deep learning-based models [30].

3.1.1 Rule-based and machine learning models

Rule-based approaches are one of the earliest approaches used to extract PICO information from research abstracts. Demner-Fushman and Lin [18] proposed a rule-based approach where they first identified the sentences containing the PICO information and then created different rule patterns to extract the PICO elements. Kelly and Yang [32] used regular expressions to extract number of participants, gender, ethnicity, age, study duration, and so on. Regular expressions and rules are useful when finding patterns that adhere to a particular structure. However, they rely heavily on hand-crafted rules and therefore have some limitations. First limitation is that they are a brute force approach where one needs to be aware of all the possible patterns. Second, creating sets of rules/patterns for each named entity class is time consuming. Third, the rules tend to be domain specific and cannot be transferable to other domains. Moreover, sometimes they require domain knowledge and expertise for their development.

SVM [25, 16] based approach has also been used to extract participants information from RCTs abstracts [24]. SVM classifiers make binary decision on whether a token belongs to one of pre-defined classes. The basic foundation of SVM is to learn a linear hyperplane that separates positive samples from negative samples by a large margin. Given an input training sequence $x = (x_1, x_2, ..., x_n)$ with label sequence $y = (y_1, y_2, ..., y_n)$, SVM learns a function $f(x) = wx + b = 0$ of a separating hyperplane with maximum margin. Both w and b are parameters learned from the training dataset, where w is a weight vector and b is a bias that determines the offset of the hyperplane from the origin. Margin is the separation between the hyperplane and the support vectors (data points that are closest to the hyperplane). A data sample x is classified as positive if $f(x) = wx+b > 0$ and negative otherwise. When there exists more than two classes, multiple classifiers are used to classify the samples.

HMM models are some of the earliest approaches for solving NER tasks [20, 6]. Xu et al. [64] used HMM-based approach to extract participants demographics, diseases, symptoms and so on from research abstracts. HMM is a generative statistical model which uses the Viterbi algorithm [57] to assign the most likely target sequence to each word sequence. HMM can be represented with three parameters: $\lambda = (A, B, \Pi)$, where, Π represents the start probability, A represents the transition probability, and B represents the emission probability. Start probability (Π) is the probability that a particular tag will appear first in a sentence. Transition probability ($A = a_{ij}$) is the probability that the subsequent tag j will appear in a sentence given the current tag i. Emission probability $B = b_j(m)$ is the probability of an output sequence occurring given state j. During the training phase, HMM takes annotated training data as input and outputs the three parameters. In the testing phase, HMM takes sentence and the obtained three parameters and outputs the sequence of states from which named entities can be detected.

Summerscales et al. [54] utilized CRF-based models [35] to extract treatments, groups, and outcomes from research abstracts. CRF models are probabilistic models that take into consideration neighboring examples as contextual features. Given an input sequence $x = (x_1, x_2, ..., x_n)$ with label sequence $y = (y_1, y_2, ..., y_n)$, CRF models the conditional probability as:

$$P(y|x, \lambda) = \frac{1}{Z(x)} exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_i(x, i, y_{i-1}, y_i).$$

$Z(x)$ is a normalization factor defined as:

$$Z(x) = \sum_{y \in Y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_i(x, i, y_{i-1}, y_i).$$

Here, $f_i(x, i, y_{i-1}, y_i)$ is a feature function which computes probability by taking into account the current and previous class labels. λ is a learning weight attributed to the feature function and is calculated during training.

Rule-based, SVM, HMM, and CRF-based models are useful in information extraction. However, they heavily rely on hand-crafted features. Designing hand-crafted features is time-consuming and might require domain knowledge in determining useful features. In recent years, deep learning-based models gained popularity because they can learn hidden features automatically.

Jin and Szolovits [27] proposed a long-short-term memory (LSTM) model to extract PICO elements and later proposed model an improved model that consists of bidirectional LSTM (bi-LSTM) model with a CRF layer on top (bi-LSTM-CRF) [28]. Bi-LSTM-CRF model can capture dependencies in both left and right directions of the input sequence. The bi-LSTM-CRF model consists of an embedding layer, a bi-LSTM layer, and a CRF layer. Given an input sequence $s = (s_1, s_2, ..., s_n)$ with a label sequence $y = (y_1, y_2, ..., y_n)$, the embedding layer maps each token to a vector representation $x = (x_1, x_2, ..., x_n)$, where x_i is the token embedding of s_i. The bi-LSTM layer takes the embedding layer token/word embeddings as input and outputs a contextualized vector consisting of two hidden states (forward and backward). The CRF layer conditional probability is calculated as:

$$P(y|x) = \frac{e^{S(x,y)}}{\sum_{\hat{y} \in y} e^{S(x,\hat{y})}},$$

where x is the input sequence, y is the label of the sequence, and S is the score of the prediction result sequence. For a given sequence, the probability score is calculated as:

$$s(x, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i},$$

where $A_{y_i, y_{i+1}}$ is the transition scores and P_{i, y_i} is the emission scores. In training phase, the maximum likelihood of probability of gold label sequences are maximized. The final predicted label is calculated based on the highest score which is expressed as:

$$y^* = \sum \arg\max S(x, \hat{y}).$$

3.1.2 Pre-trained language models

Pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) have recently attracted attention due to their state-of-the-art performance in various NLP tasks including NER [19]. Pre-trained language models learn general language representation from lots of general domain corpus such as Wikipedia and BooksCorpus. BERT is pre-trained on large corpus of unlabeled data and the bidirectional nature of the model means that it learns information from left to right and right to left, and this is what makes it a powerful language model.

Pre-trained language models consists of two stages; pre-training and fine-tuning. The pre-training stage consists of self-supervised tasks which include masked language modelling (MLM) and next sequence prediction (NSP). In MLM, words are randomly masked (hidden) and the language model predicts the masked words to complete the sentence. In NSP, the language model learns relationships between sentences and predicts the next sentence in a pair. In fine-tuning stage, the models are trained with a small amount of labeled data and adapted to various NLP tasks such as NER.

BERT model takes a sequence of tokens as input. It has special tokens, [CLS] and [SEP]. [CLS] is a classification which is the first token of the input sequence and [SEP] token is the last token of the input sequence. The maximum input sequence size of the BERT model is 512 tokens including the [CLS] and [SEP] tokens. Sequences longer than 512 tokens are usually truncated and shorter sequences are padded with the [PAD] token to fill the unused slots.

The BERT model architecture for NER task is as shown in Figure 4. Given a sequence of input tokens $x = (x_1, x_2, ..., x_n)$ and the token gold labels $y = (y_1, y_2, ..., y_n)$, each input token is mapped to its embeddings $e = (e_1, e_2, ..., e_n)$. The hidden layer outputs the hidden vector $h = (h_1, h_2, ..., h_n)$ which is then passed to the fully connected layer for prediction. The prediction for the i-th token is calculated as

$$P(\hat{y}_i|x_i) = softmax(Wh_i + b),$$

where y_i si the gold label, x_i is the ith-token, h_i is the final hidden state of the i-th token, and W and b are hyperparamaters. During training, the model minimizes the cross-entropy loss:

$$CrossEntropy = -\sum_{c=1}^{n} y_c log P(\hat{y}_c|x_c),$$

where y_c is the gold label and $P(\hat{y}_c|x_c)$ is the softmax probability for the c^{th} class. During testing, the model predicts the class as:

$$Prediction_i = argmax(log P(\hat{y}_c|x_c)).$$

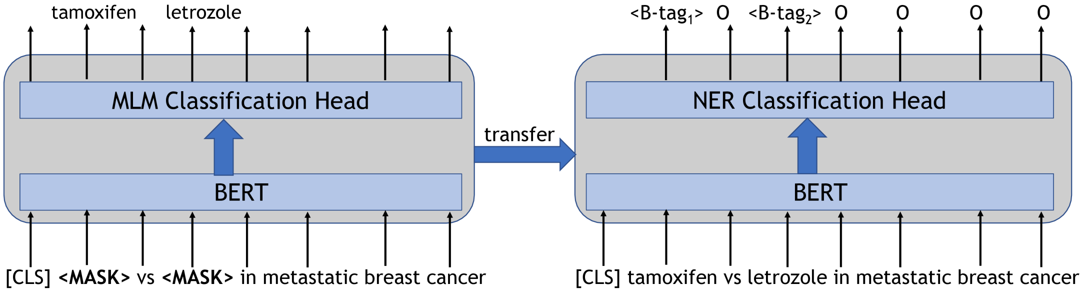

tamoxifen
letrozole
MLM Classification Head
BERT
[CLS] <MASK> vs <MASK> in metastatic breast cancer
transfer
<B-tag$_1$> O <B-tag$_2$> O O O O
NER Classification Head
BERT
[CLS] tamoxifen vs letrozole in metastatic breast cancer

Figure 4: BERT model for NER task

Traditional BERT models cannot attend to long sequences and are limited to a maximum of 512 tokens at a time. This is due to the self-attention operation which grows quadratically with sequence length. Modified transformer models, such as Longformer [5], have been created to overcome this problem. In Longformer model, the self-attention pattern scales linearly with sequence length enabling it to process longer documents. It can attend to long sequences of up to 4096 tokens, which is eight times longer than BERT.

Longformer uses a sliding window self attention mechanism (known as local attention) to capture context. For instance, if we assume a context window of length w, each token attends to $w/2$ tokens to the left and to the right of the current token. If the sequence length is n, the complexity is $O(n.w)$ which scales linearly. In addition, the authors of the Longformer model also applied a dilation to the sliding window so as to increase the size of w without using extra memory. Dilation refers to the ability to skip a token to allow the attention to reach further tokens. For instance, if the dilation size is d, the number of gaps between each token in the window will be d. This does not affect model performance since transformer architecture includes multiple attention heads across multiple layers which can learn and attend to multiple tokens and texts properties.

3.1.3 Evaluation Metrics

To evaluate NER tasks, standard evaluation metrics of Precision, Recall, and F1-measure are commonly used.

- Precision is the ratio of the correctly identified entities to the number of all identified entities.

$$P = \frac{TP}{TP + FP}$$

- Recall is the ratio of correctly identified entities to the number of actual entities in the gold set.

$$R = \frac{TP}{TP + FN}$$

- F1-measure is the weighted mean of precision and recall.

$$F1 = 2 * \frac{P * R}{P + R}$$

TP (true positive) is the number of entities that were correctly identified, FP (false positive) is the number of entities that were incorrectly identified, and FN (false negative) is the entities that the model failed to identify.

3.2 Methods

3.2.1 Data pre-processing

The pre-processing step mainly involves acronym expansion. In research articles, acronyms are frequently used to avoid repeating long terms and save space. Even though acronyms simplify writing and reading, they are a major obstacle to natural language text understanding tasks [47]. Generally, acronyms can have multiple common expansions which depend on a particular context. Acronyms commonly occur in the words preceding their first occurrence in parentheses, for example, “Randomized controlled trials (RCT) of scalp cooling (SC) to prevent chemotherapy induced alopecia (CIA)”. In this research, we employ a rule-based method using regular expressions for acronym expansion. The first step in identifying acronyms is to look for terms in parenthesis that are between two and ten characters long. Regular expressions are then used to find expansion candidates in the surrounding text.

3.2.2 PICO elements extraction

Data extraction aims to extract PICO elements from research abstracts. This task is formulated as a sequence labelling task, i.e., given a token, classify it as one of pre-defined named entity recognition (NER) tags. As deep learning models have gained a lot of attention in NLP tasks, we adopt BERT-based models for this task. BERT has achieved state-of-the-art performance in NER tasks and has also proven to be effective for small datasets [19]. BERT is a language model pre-trained on huge amounts of unlabeled data and can be fine-tuned to specific tasks such as NER.

We chose three pre-trained transformer-based models, i.e., BioBERT [36], BlueBERT [46], and Longformer [5]. BioBERT is pre-trained on different combinations of general and biomedical domain corpora. It is initialized with BERT [19] and further pre-trained on biomedical domain texts (PubMed abstracts and PubMed Central full-text articles). BlueBERT is also initialized with BERT and further pre-trained on PubMed abstracts and clinical notes from MIMIC-III [29]. Longformer is initialized with the RoBERTa model [37] and further pre-trained with books, wikipedia, realnews, and stories.

Moreover, we developed a web-based system[6] for extracting PICO information from RCTs abstracts. The system was developed using Python. When using the system, shown in Figure 5, a user inputs free-text and selects the model to use for information extraction. Currently, two models, that is, BioBERT and Longformer, are available. The system then extracts PICO information from the

[6] https://aoi.naist.jp/autometa-demo-v2/

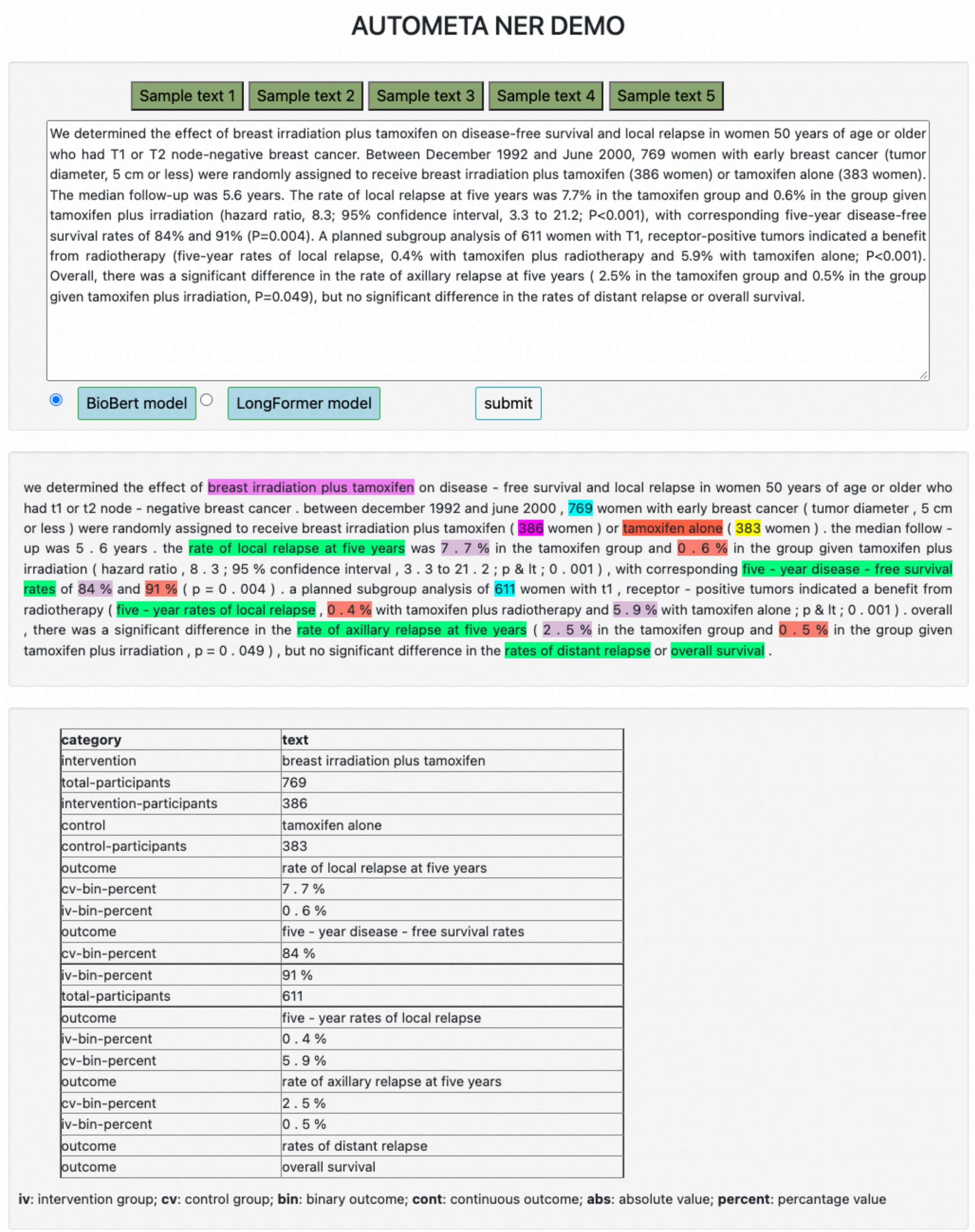
input text and outputs the results in a table as shown in the Figure 5.

3.2.3 Experimental settings

Our corpus consists of 1011 PubMed abstracts annotated with PICO elements (discussed in Chapter 2). The dataset was split into 80% training set and 20% test set. We developed BERT-based models for data extraction (NER) and compared the performance of general-purpose (Longformer) and biomedical domain (BioBERT, BlueBERT) BERT models. The BioBERT and BlueBERT models cannot attend to sequences longer than 512 tokens. BERT uses WordPiece [62] tokenization and a word can be broken down into more than one sub-words. In the corpus, some abstracts were found to have more than 512 tokens after the WordPiece tokenization process. The default strategy for the BioBERT and BlueBERT models is to truncate long sequences and ignore the tokens after the maximum number is reached. Since truncation leads to loss of information, we split sequences longer than the maximum length into multiple chunks so as to preserve all the information. The split was done in a sentence-wise manner, i.e., if the number of tokens in an abstract is more than 512, we split the abstract into individual sentences, then split the sentences into two halves to create two almost equal chunks. If the number of tokens is greater than 1024, the abstracts are split into three chunks and so on. The abstracts were split into sentences using the NLTK sentence tokenizer package[7].

In the experiments, we followed the standard pre-trained BERT models for sequence classification. The pre-trained models were fine-tuned on our corpus. The fine-tuning was done by setting the maximum sequence length to 512 tokens for the BioBERT and BlueBERT models and 4096 tokens for the Longformer model. The number of epochs was set to 10, batch size was set to 2, and the learning rate was set to 2e-5 for the BioBERT model and 5e-5 for BlueBERT and Longformer models. Moreover, since neural networks provide different results when initialized with different seeds, we trained each of the models with five different seeds and averaged the results.

[7] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html



AUTOMETA NER DEMO

We determined the effect of breast irradiation plus tamoxifen on disease-free survival and local relapse in women 50 years of age or older who had T1 or T2 node-negative breast cancer. Between December 1992 and June 2000, 769 women with early breast cancer (tumor diameter, 5 cm or less) were randomly assigned to receive breast irradiation plus tamoxifen (386 women) or tamoxifen alone (383 women). The median follow-up was 5.6 years. The rate of local relapse at five years was 7.7% in the tamoxifen group and 0.6% in the group given tamoxifen plus irradiation (hazard ratio, 8.3; 95% confidence interval, 3.3 to 21.2; P<0.001), with corresponding five-year disease-free survival rates of 84% and 91% (P=0.004). A planned subgroup analysis of 611 women with T1, receptor-positive tumors indicated a benefit from radiotherapy (five-year rates of local relapse, 0.4% with tamoxifen plus radiotherapy and 5.9% with tamoxifen alone; P<0.001). Overall, there was a significant difference in the rate of axillary relapse at five years (2.5% in the tamoxifen group and 0.5% in the group given tamoxifen plus irradiation, P=0.049), but no significant difference in the rates of distant relapse or overall survival.

we determined the effect of breast irradiation plus tamoxifen on disease - free survival and local relapse in women 50 years of age or older who had t1 or t2 node - negative breast cancer . between december 1992 and june 2000 , 769 women with early breast cancer (tumor diameter , 5 cm or less) were randomly assigned to receive breast irradiation plus tamoxifen (386 women) or tamoxifen alone (383 women) . the median follow - up was 5 . 6 years . the rate of local relapse at five years was 7 . 7 % in the tamoxifen group and 0 . 6 % in the group given tamoxifen plus irradiation (hazard ratio , 8 . 3 ; 95 % confidence interval , 3 . 3 to 21 . 2 ; p & lt ; 0 . 001) , with corresponding five - year disease - free survival rates of 84 % and 91 % (p = 0 . 004) . a planned subgroup analysis of 611 women with t1 , receptor - positive tumors indicated a benefit from radiotherapy (five - year rates of local relapse , 0 . 4 % with tamoxifen plus radiotherapy and 5 . 9 % with tamoxifen alone ; p & lt ; 0 . 001) . overall , there was a significant difference in the rate of axillary relapse at five years (2 . 5 % in the tamoxifen group and 0 . 5 % in the group given tamoxifen plus irradiation , p = 0 . 049) , but no significant difference in the rates of distant relapse or overall survival .

category	text
intervention	breast irradiation plus tamoxifen
total-participants	769
intervention-participants	386
control	tamoxifen alone
control-participants	383
outcome	rate of local relapse at five years
cv-bin-percent	7 . 7 %
iv-bin-percent	0 . 6 %
outcome	five - year disease - free survival rates
cv-bin-percent	84 %
iv-bin-percent	91 %
total-participants	611
outcome	five - year rates of local relapse
iv-bin-percent	0 . 4 %
cv-bin-percent	5 . 9 %
outcome	rate of axillary relapse at five years
cv-bin-percent	2 . 5 %
iv-bin-percent	0 . 5 %
outcome	rates of distant relapse
outcome	overall survival

iv: intervention group; **cv**: control group; **bin**: binary outcome; **cont**: continuous outcome; **abs**: absolute value; **percent**: percantage value

Figure 5: NER system for automatic extraction of PICO elements

3.3 Results and discussion

The performance of the NER model was evaluated using Precision, Recall, and F1 score in the test set and the results are shown in Table 3. BioBERT_split and BlueBERT_split are the model results where sequences longer than 512 tokens were split into multiple chunks. The Longformer model did not require splitting of abstracts because the maximum sequence length for Longformer is 4096 tokens and there were no abstracts with tokens exceeding the maximum number.

The performance was relatively high with sub-categories such as total-participants and outcome-measure achieving F1-scores greater than 0.90. Most of the other sub-categories achieved F1-scores greater than 0.80. F1-score was zero for the entities with lowest frequency such as cont-q1-iv, cont-q1-cv, cont-q3-iv, and cont-q3-cv. In overall, BioBERT and Longformer models achieved the highest performance in almost all of the entities.

The Longformer model, which is a general purpose model, performed well compared to the biomedical domain BERT models (BioBERT and BlueBERT). One likely explanation is that the biomedical domain BERT models have a maximum sequence length of 512 tokens and longer sequences are truncated resulting in loss of important contextual information. The Longformer model has a maximum sequence length of 4096 tokens and could therefore build contextual representation of the entire context.

The input of the NER models was the entire abstract. It is common practice in tasks like these to use a sentence as the input. In this task, identification of most entities depends on data from the entire abstract, and hence using data from a single sentence would be insufficient. By using the entire abstract as input, we can incorporate context clues from other sentences to enhance the model performance. However, since some abstracts were longer than the standard input length for BERT models (512 tokens), long abstracts were split into multiple chunks sentence-wise.

The splitting of long sequences was expected to increase model performance, however, there was no change in the model performance. This could be attributed to loss of useful contexts caused by splitting. In this research, it is necessary to extract information from the entire abstract. The default strategy for BERT models is to truncate long texts hence leading to loss of important information.

The purpose of splitting the abstracts into multiple chunks was to enable extraction of information from the entire abstracts. Even though splitting the abstracts did not improve the performance, we were able to avoid loss of information due to truncation.

The confusion matrices for each of the models are as shown in Figure 6. Since there are many sub-categories, the confusion matrices show the results for the major categories for clarity purposes. The BioBERT model has a true-positive rate of 0.943, false-positive rate of 0.002, false-negative rate of 0.057, and true-negative rate of 0.998. The BioBERT_split model has a true-positive rate of 0.943, false-positive rate of 0.002, false-negative rate of 0.057, and true-negative rate of 0.998. The BlueBERT model has a true-positive rate of 0.935, false-positive rate of 0.002, false-negative rate of 0.06, and true-negative rate of 0.998. The BlueBERT_split model has a true-positive rate of 0.937, false-positive rate of 0.002, false-negative rate of 0.006, and true-negative rate of 0.998. The longformer model has a true-positive rate of 0.961, false-positive rate of 0.001, false-negative rate of 0.04, and true-negative rate of 0.999. All the models achieved relatively high performance with longformer model having the highest true-positive rate and lowest false negatives.

Error analysis

We performed an error analysis and identified misclassified entities, boundary detection, and missed entities as the major types of errors.

- Misclassified entities: this is where the model detected the correct boundaries for entities but assigned them the wrong classes. For example, the model sometimes misclassified intervention events (e.g., *bin-abs-iv*) as control events (e.g., *bin-abs-cv*) and vice versa. Example (i) in Table 4(a) shows a situation where the model identified the entities but misclassified *cv-cont-median* as *iv-cont-median* and vice-versa. The reason might be because intervention events tend to be reported before control events in most of the samples in our corpus. The model might have then learned the pattern that intervention events are reported before control events when no other clues are available.

Table 3: NER models results. Bold texts represent the best score for each sub-category.

(a) BioBERT model results

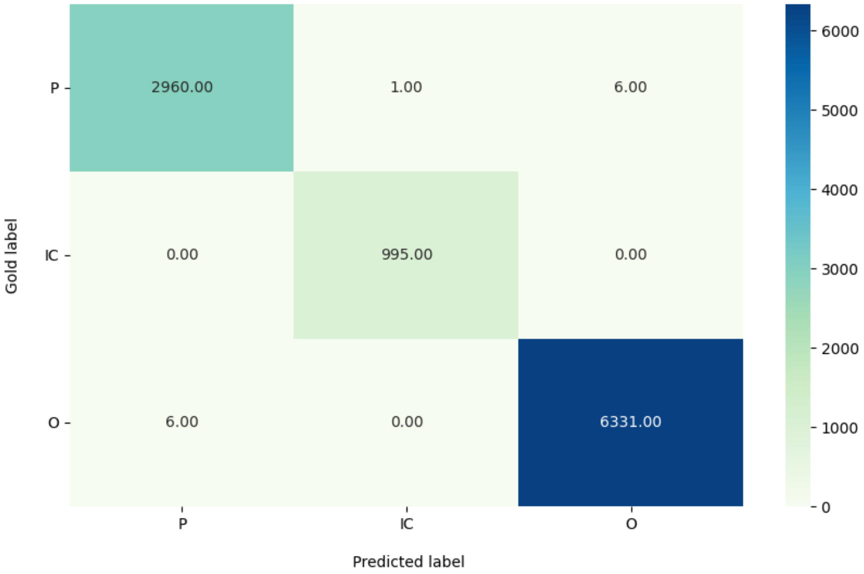
	BioBERT			**BioBERT_split**		
Sub-category	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Total-participants	0.95	**0.95**	**0.95**	0.94	0.94	0.94
Intervention-participants	**0.80**	0.91	**0.85**	0.78	**0.93**	**0.85**
Control-participants	0.87	**0.91**	**0.89**	0.85	**0.91**	0.88
Age	0.66	0.97	0.79	0.66	0.96	0.78
Eligibility	0.75	0.77	0.76	0.77	0.74	0.76
Ethnicity	0.82	0.89	0.86	0.82	**0.96**	**0.88**
Condition	0.86	**0.81**	**0.84**	0.84	0.75	0.79
Location	0.75	**0.85**	0.80	0.73	0.81	0.77
Intervention	0.85	0.82	0.84	0.85	0.82	0.84
Control	0.78	0.80	0.79	0.77	0.76	0.77
Outcome	0.82	0.81	0.81	0.84	0.80	0.82
Outcome-measure	0.79	0.90	0.84	0.81	0.88	0.84
bin-abs-iv	0.75	0.78	0.77	0.81	0.78	0.79
bin-abs-cv	0.79	**0.87**	0.83	0.77	0.80	0.79
bin-percent-iv	**0.87**	0.88	0.87	0.83	0.86	0.84
bin-percent-cv	**0.88**	**0.90**	**0.89**	0.87	0.82	0.84
cont-mean-iv	0.78	**0.90**	0.83	0.80	0.86	0.83
cont-mean-cv	**0.86**	0.86	**0.86**	0.81	0.84	0.83
cont-median-iv	**0.70**	0.80	0.75	**0.70**	**0.86**	**0.78**
cont-median-cv	0.76	**0.81**	**0.78**	**0.83**	0.74	**0.78**
cont-sd-iv	0.68	**0.93**	0.79	0.80	0.85	0.82
cont-sd-cv	0.76	0.84	0.80	0.72	0.85	0.78
cont-q1-iv	0.00	0.00	0.00	0.00	0.00	0.00
cont-q1-cv	0.00	0.00	0.00	0.00	0.00	0.00
cont-q3-iv	0.00	0.00	0.00	0.00	0.00	0.00
cont-q3-cv	0.00	0.00	0.00	0.00	0.00	0.00

(b) BlueBERT model results

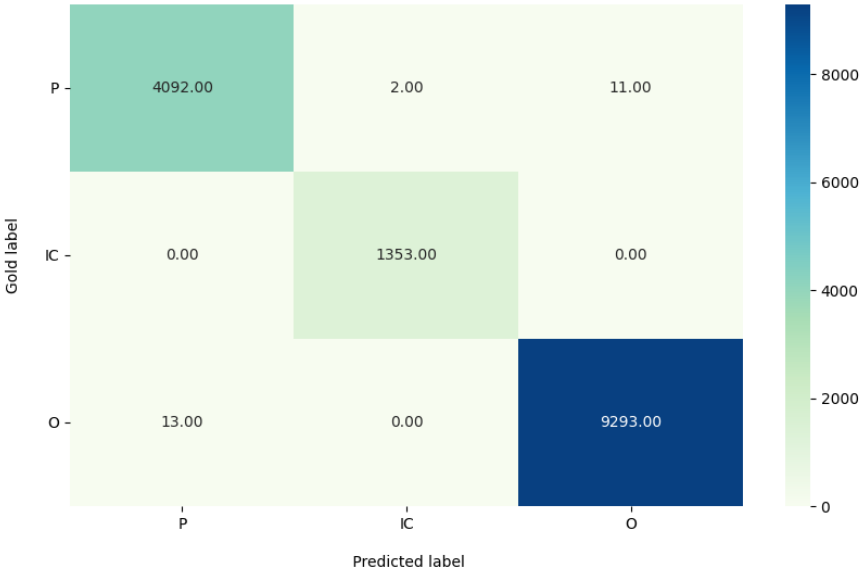
	BlueBERT			**BlueBERT_split**		
Sub-category	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Total-participants	0.94	0.91	0.92	0.95	0.92	0.94
Intervention-participants	0.72	0.90	0.80	0.73	0.91	0.81
Control-participants	0.81	0.85	0.83	0.79	0.89	0.84
Age	0.67	0.97	0.79	0.66	0.97	0.79
Eligibility	0.73	0.74	0.73	0.73	0.70	0.72
Ethnicity	0.90	0.72	0.80	**0.91**	0.78	0.84
Condition	**0.90**	0.70	0.79	0.82	0.77	0.79
Location	0.77	0.67	0.71	0.76	0.76	0.76
Intervention	0.80	0.81	0.81	0.84	0.83	0.83
Control	0.72	0.68	0.70	0.78	0.71	0.74
Outcome	0.81	0.79	0.80	0.81	0.80	0.80
Outcome-measure	0.73	0.84	0.78	0.76	0.86	0.81
bin-abs-iv	0.77	0.75	0.76	0.67	0.76	0.71
bin-abs-cv	0.75	0.79	0.77	0.72	0.84	0.78
bin-percent-iv	0.74	0.85	0.79	0.79	0.81	0.80
bin-percent-cv	0.83	0.73	0.78	0.82	0.79	0.80
cont-mean-iv	0.72	0.74	0.73	0.61	0.81	0.69
cont-mean-cv	0.77	0.74	0.75	0.73	0.76	0.74
cont-median-iv	0.65	0.78	0.71	0.67	0.62	0.64
cont-median-cv	0.80	0.66	0.72	0.75	0.66	0.70
cont-sd-iv	0.62	0.68	0.65	0.59	0.60	0.59
cont-sd-cv	0.67	0.68	0.67	0.56	0.70	0.63
cont-q1-iv	0.00	0.00	0.00	0.00	0.00	0.00
cont-q1-cv	0.00	0.00	0.00	0.00	0.00	0.00
cont-q3-iv	0.00	0.00	0.00	0.00	0.00	0.00
cont-q3-cv	0.00	0.00	0.00	0.00	0.00	0.00

(c) Longformer model results

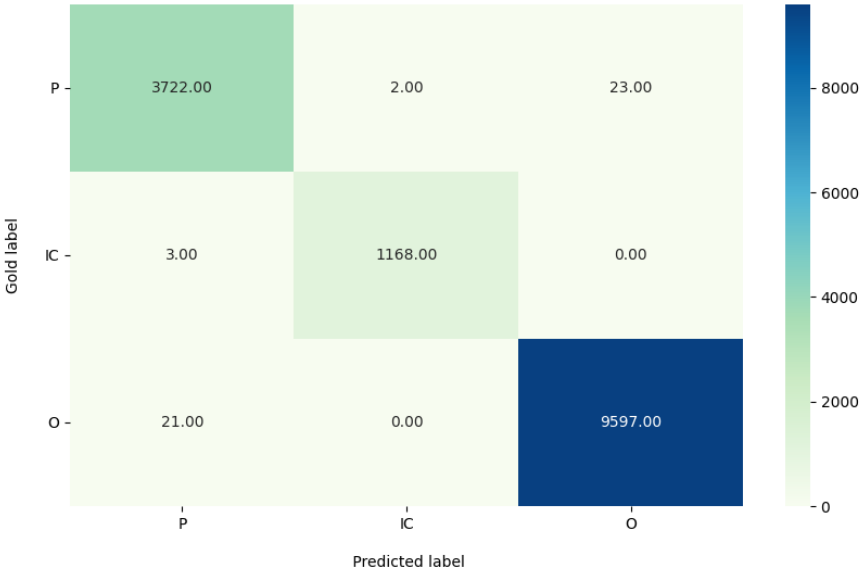
Sub-category	**Precision**	**Recall**	**F1**
Total-participants	**0.96**	0.94	**0.95**
Intervention-participants	0.79	0.92	**0.85**
Control-participants	**0.89**	0.89	**0.89**
Age	**0.78**	**0.98**	**0.87**
Eligibility	**0.89**	**0.86**	**0.88**
Ethnicity	0.75	0.83	0.78
Condition	0.83	0.79	0.81
Location	**0.91**	0.79	**0.85**
Intervention	**0.86**	**0.85**	**0.86**
Control	**0.81**	**0.86**	**0.83**
Outcome	**0.85**	**0.86**	**0.86**
Outcome-measure	**0.85**	**0.95**	**0.90**
bin-abs-iv	**0.83**	**0.83**	**0.83**
bin-abs-cv	**0.8**4	0.85	**0.84**
bin-percent-iv	0.85	**0.90**	**0.88**
bin-percent-cv	**0.88**	0.85	0.87
cont-mean-iv	**0.85**	0.87	**0.86**
cont-mean-cv	0.78	**0.91**	0.84
cont-median-iv	0.65	0.76	0.70
cont-median-cv	0.75	0.76	0.75
cont-sd-iv	**0.83**	0.86	**0.85**
cont-sd-cv	**0.77**	**0.92**	**0.84**
cont-q1-iv	0.00	0.00	0.00
cont-q1-cv	0.00	0.00	0.00
cont-q3-iv	0.00	0.00	0.00
cont-q3-cv	0.00	0.00	0.00



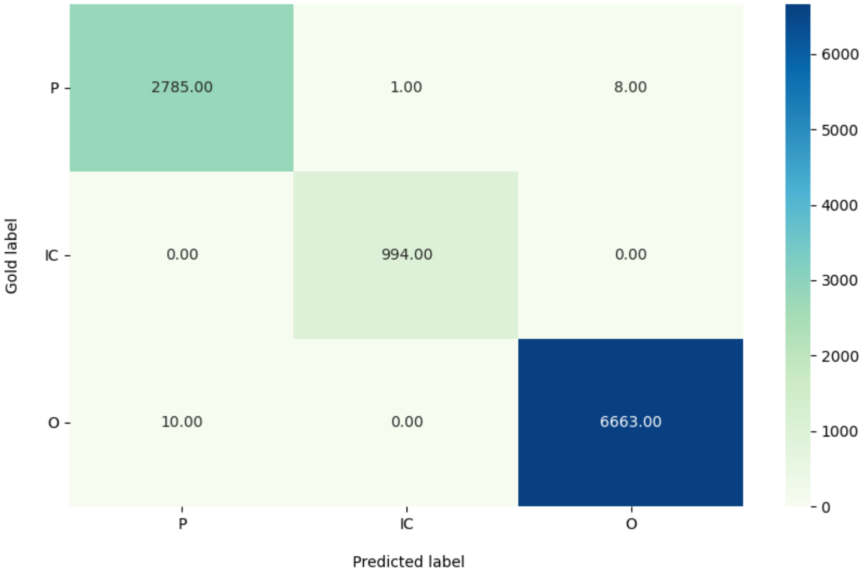
(a) BioBERT model



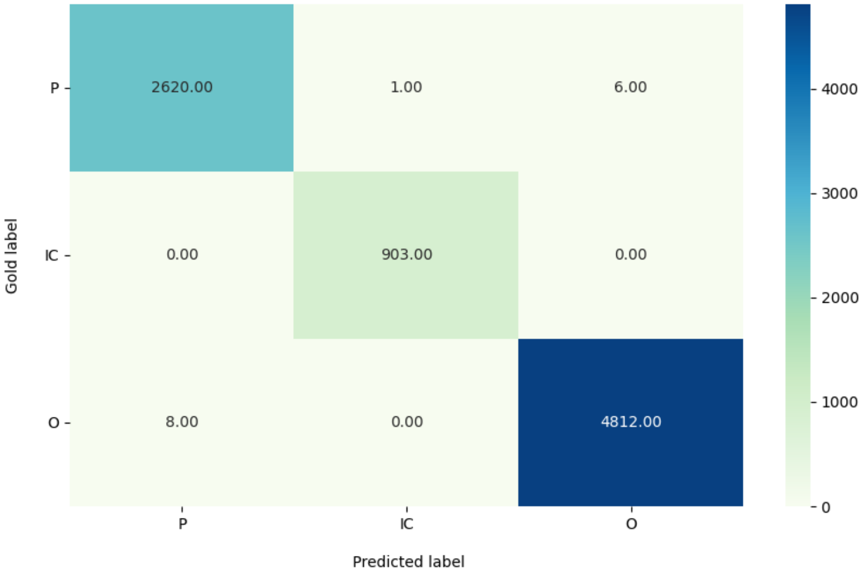
(b) BioBERT_split model



(c) BlueBERT model



(d) BlueBERT_split model



(e) Longformer model

Figure 6: Confusion matrices for the NER models

- Boundary detection: here the model identifies shorter or longer entities than those marked in the gold set. Human annotation could contribute to this error, because sometimes it is difficult to decide the start and end spans of some entities. Examples (ii) and (iii) in Table 4(a) show cases where the models and the annotators identified different spans. These type of error was common in entities which require more than two words to be identified especially the *outcome* and *eligibility* entities.

  Furthermore, the BERT tokenization process also contributed to this type of errors. We found out that the BERT tokenizer tokenizes decimal numbers into multiple individual tokens. For example, 56.3 is tokenized to ‘56’, ‘.’, ‘3’. Also, percentage values such as 15% are tokenized to ‘15’, ‘%’. In some cases, the model predicted the numbers and ‘%’ as different entities.

- Missed entities: this is where the model fails to identify the entities. Example (iv) in Table 4(a) shows an example where the model failed to identify the *control* entity. In the training set, many *control* entities are identified by terms such as placebo and control and hence the reason the model did not identify it. Example (v) shows a case where the model captured an entity that was missed during the annotation process.

Table 4(b) shows the number of errors in each of the NER models. The values were calculated as the percentage of entities out of the total entities whose predictions were incorrect. Boundary detection errors were the most frequent errors whereas missed entities were the fewest. The *outcome* and *eligibility* entities contributed to majority of the errors. These entities need more than two words to be identified, and it was also challenging for human annotators to determine their spans, as mentioned in Section 2.3.

The Longformer and BioBERT_split model had the least errors whereas the BlueBERT model had the most errors. These results are consistent with the results discussed in the previous section, where the Longformer and BioBERT models achieved the best performance.

Table 4: NER Error analysis

(a) Example sentences with prediction errors

	Example sentence	**Predicted label**	**Gold label**	**Comment**
(i)	...median time to death was *17.4 months* and 16.0 months...	iv-cont-median	cv-cont-median	Misclassified entities
	...median time to death was 17.4 months and *16.0 months*...	cv-cont-median	iv-cont-median	
(ii)	...women between 18 and 75 years...	age *18 and 75 years*	age *between 18 and 75 years*	Boundary detection
(iii)	...incidence of grade 3/4 anemia...	outcome *grade 3/4 anemia*	outcome *incidence of grade 3/4 anemia*	Boundary detection
(iv)	..were randomized to ... or *treatment as usual*	_	control	Missed entity
(v)	...significant decrease in *fatigue*...	outcome	_	Annotation error

(b) Number of errors in the NER models

Model	Misclassified entities	Boundary detection	Missed entities
BioBERT	5.67%	8.09%	2.28%
BioBERT_split	5.68%	5.32%	2.46%
BlueBERT	6.47%	9.15%	2.66%
BlueBERT_split	6.28%	8.44%	2.64%
Longformer	3.88%	4.74%	1.61%

3.4 Conclusion

We proposed BERT-based NER models for PICO extraction from RCT literature. The NER models extracted PICO elements with relatively high accuracy. Some of the entities achieved F1-scores higher than 0.90 and most of the other entities achieved F1-scores greater than 0.80. Some entities could not be detected due to low frequency in the dataset, and hence some of the future work is to increase the training set to include these entities. In addition, since traditional BERT models can only process a maximum of 512 tokens, we proposed a technique of splitting long texts into multiple chunks. This technique avoided information loss due to truncation, however, the models performance did improve. In future, it is important to investigate more approaches on effective text splitting to improve the models performance.

4. Automating Meta-Analysis Statistical Analysis

4.1 Background and related work

Natural language processing techniques to accelerate data extraction from research abstracts and full-text articles have been widely studied [30]. Most of the proposed approaches extract the different information which describe the PICO elements such as the age, study design, medications, medication dosage, medication frequency, outcomes, and so on. These previous studies, however, fail to extract detailed information for the outcomes, especially numeric texts identifying the number of patients having certain outcomes [48, 55]. Extraction of numeric texts is important for statistical analysis to determine the effectiveness of the intervention. Extracting this information is difficult due to lack of uniformity and different studies report their results differently.

Summerscales et al. [55] used conditional random field-based (CRF) approach to extract various named entities including treatment groups, group sizes, outcomes, and outcome numbers from research abstracts. They created a corpus with 263 abstract and annotated different entities including treatment groups, group sizes, outcomes, and outcome numbers. The proposed system first filters sentences likely to contain the relevant information so as to reduce the amount of texts to be processed. The selected sentences were those which at contain at least one integer as such sentences are considered most likely to contain outcome mentions and outcome numbers. Their approach to extract outcomes and outcome numbers for the purpose of calculating summary statistics is similar to this research. However, their annotations are less extensive and their corpus is not publicly available for reproducibility.

Pradhan et al. [48] developed EXACT, a Python based web application tool extracting data from ClinicalTrials.gov. ClinicalTrials.gov is a United States clinical trials registry database supported by the United States National Library of Medicine (NLM)[8]. EXACT parses the database and extracts data from the stored clinical trials. EXACT can extract upto 30 different data elements which include

[8] https://clinicaltrials.gov/

the baseline information (such as study type, intervention, control, condition, etc.), outcomes and their events, outcome measures, and adverse events. The system was evaluated by reproducing the results of three meta-analyses containing a total of fifteen clinical trials. The data extracted by the EXACT system were compared to manually extracted data. The system reduced the data extraction time by 60% and the data elements were extracted with 100% accuracy. EXACT uses data mining techniques rather than machine learning, hence the data extracted is 100% accurate with no extraction errors. The reason for high accuracy is because the data was extracted from the ClinicalTrials.gov database where data is recorded in a structured format. Although the ClinicalTrials.gov database is an important source of clinical trials data, it has mainly focuses on clinical trials in the United States. This substantially decreases the number of studies available for data extraction [48].

The goal of this work is to provide a system that supports meta-analysis statistical analysis. The proposed system (shown in Figure 1) performs various steps including extracting data from research abstracts (discussed in Chapter 2), parsing numeric outcomes to identify the number of patients having specific outcomes, converting extracted data into a structured format for statistical analysis, and visualizing the results. We assess the performance of the proposed system by using it to reproduce the results of existing meta-analysis studies and show potential in automating the meta-analysis statistical analysis task.

4.2 Methods

4.2.1 PICO elements normalization

Meta-analysis involves combining similar studies to assess the effectiveness of the intervention (treatment). To automatically group similar studies together and compare them within a meta-study, it is necessary to normalize the extracted PICO elements. We focus on the normalization of the intervention, control, and outcome elements. Since our corpus consists of RCTs related to breast cancer, all participants are breast cancer patients.

We utilize the UMLS Metathesaurus for the normalization of intervention and control elements. UMLS comprehensively covers most of the interventions and control, especially medications, and hence we did not need to create a normalization dictionary manually. We use MetaMap [3], which is a state-of-the-art NLP tool that maps biomedical text to concepts in the UMLS Metathesaurus. For each text, MetaMap splits the text into phrases and identifies possible mappings for each phrase based on lexical look-up and variants.

A dictionary-based approach was employed for outcome normalization. We extracted all the outcomes from the corpus and manually created a dictionary of the outcomes and their normalizations. For example, pain, breast pain, less pain, and mild pain are all normalized to pain. After creating the dictionary in this manner, we use dictionary string matching techniques to match outcomes and their normalized versions.

The task of matching an outcome with its normalization is defined as follows. Given a predefined set of normalized outcomes N, and an input string o (outcome), find normalized outcome $n \in N$ that is most similar to o. For this task, we utilize a technique that combines Term-Frequency Inverse Document Frequency (TF-IDF), n-grams, and cosine similarity. TF-IDF creates features from text by multiplying the frequency of a term in a document (term frequency) by the importance (inverse document frequency) of the term in the entire corpus. In TF-IDF, usually the term is a word, but depending on the corpus, n-grams have been shown to achieve high performance. For each outcome, we represent the outcome as a vector using TF-IDF and calculate the cosine similarity between the outcome vector and the normalized outcomes vectors and select the normalized outcome with the highest cosine similarity score.

Even though BERT-based models are currently widely used for NLP tasks we utilized a traditional string matching approach for outcome normalization. The current corpus contains many different outcomes which vary greatly with some occurring frequently and others occurring less frequently. Although the BERT models achieve high performance for the outcomes with high frequency, they fail for the outcomes with less frequency. Therefore, we adopted the approach of TF-IDF with cosine similarity, which achieves relatively good performance for both high-frequency and low-frequency outcomes.

4.2.2 Outcome event matching and creating structured data

Once PICO elements are extracted (as discussed in Chapter 2) and normalized, studies with the same intervention and outcome are pooled together so as to compute the overall effect of the intervention. Before calculating the overall effect of the intervention, each study's treatment effect is determined first. The effect is usually calculated using summary statistics such as risk ratio, odds ratio, or risk difference. In this research, the extracted and normalized PICO elements are converted into a structured format as shown in Figure 7. To compute the summary statistics, for each outcome four values are required, i.e., Ee, Ne, Ec, and Nc.

- Ee is the number of participants in the intervention group that demonstrated effect of the treatment (intervention events).
- Ne is the total number of participants in the intervention group.
- Ec is the number of participants in the control group that demonstrated effect of the treatment (control events).
- Nc is the total number of participants in the control group.

The summary statistics (risk ratio (RR), odds ratio (OR), and risk difference (RD)) used in this study are intended for binary outcomes.

$$RR = \frac{Ee/Ne}{Ec/Nc}$$

$$OR = \frac{Ee/(Ne - Ee)}{Ec/(Nc - Ec)}$$

$$RD = \frac{Ee}{Ne} - \frac{Ec}{Nc}$$

Ee and Ec are absolute values that correspond to bin-abs-iv and bin-abs-cv respectively (Table 2). Ee and Ec can also be calculated from bin-percent-iv and bin-percent-cv as explained in an example further down.

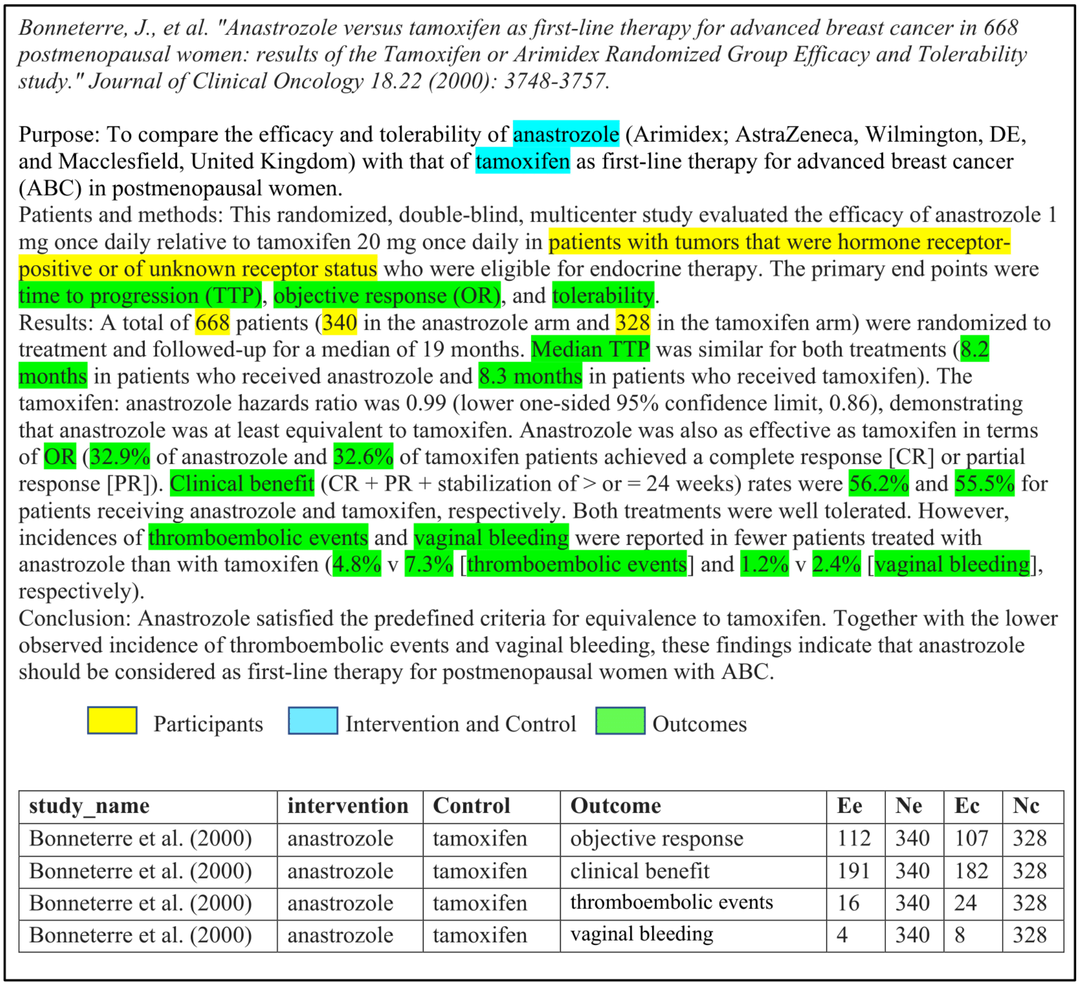
Extraction of the number of participants having certain outcomes is challenging because of lack of uniformity in reporting of results in different articles.

We use a rule-based approach for this task and assume that an outcome and its events are reported within the same sentence. If only one outcome is present in a sentence, we assume that the intervention and control events reported in that sentence belong to that outcome. If two or more outcomes are present in a sentence, the first occurrence of intervention events and control events are assigned to the first outcome, the second occurrence of intervention and control events are assigned to the second outcome, and so on. For example, "Overall survival (100% treated, 90.6% controls at 5 years) and disease-free survival (96.2% treated, 86.8% controls at 5 years) were not significantly different in the 2 groups", we extract (outcome: overall survival, intervention events: 100%, control events: 90.6%) and (outcome: disease-free survival, intervention events: 96.2%, control events: 86.8%). In this example, only percentage values are reported and hence we require knowledge of the number of participants in the intervention and control groups to calculate the absolute values (Ee and Ec). In some studies, the number of participants in the intervention and control groups (Ne and Nc) are reported in a different sentence within the abstract (as shown in the sample abstract in Figure 2) while in other studies they are not reported at all. In the rule-based approach, if the number of participants are not mentioned in the outcome sentence, we check if they are mentioned in the other sentences. Moreover, in some studies words instead of numbers are used, for instance, "Sixty-three percent achieved a complete response ...", and hence we need to convert the words to numbers. Once the abstracts have been processed in this manner, we get structured data as shown in the bottom part of Figure 7.

4.2.3 Meta-analysis results visualization system

We developed a web-based visualization system[9] for visualizing meta-analysis results. The system was developed using Python and R. R is a powerful and flexible tool that is commonly used when conducting meta-analyses. The calculations of summary statistics were implemented using meta [50], which is an R package commonly used when conducting standard meta-analysis. The results are visualized using forest plots which provide a summary and the extent to which results from different studies overlap. In the forest plot, the effect size of each study is

[9] https://aoi.naist.jp/autometavisualization/



Bonneterre, J., et al. "Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: results of the Tamoxifen or Arimidex Randomized Group Efficacy and Tolerability study." Journal of Clinical Oncology 18.22 (2000): 3748-3757.

Purpose: To compare the efficacy and tolerability of anastrozole (Arimidex; AstraZeneca, Wilmington, DE, and Macclesfield, United Kingdom) with that of tamoxifen as first-line therapy for advanced breast cancer (ABC) in postmenopausal women.
Patients and methods: This randomized, double-blind, multicenter study evaluated the efficacy of anastrozole 1 mg once daily relative to tamoxifen 20 mg once daily in patients with tumors that were hormone receptor-positive or of unknown receptor status who were eligible for endocrine therapy. The primary end points were time to progression (TTP), objective response (OR), and tolerability.
Results: A total of 668 patients (340 in the anastrozole arm and 328 in the tamoxifen arm) were randomized to treatment and followed-up for a median of 19 months. Median TTP was similar for both treatments (8.2 months in patients who received anastrozole and 8.3 months in patients who received tamoxifen). The tamoxifen: anastrozole hazards ratio was 0.99 (lower one-sided 95% confidence limit, 0.86), demonstrating that anastrozole was at least equivalent to tamoxifen. Anastrozole was also as effective as tamoxifen in terms of OR (32.9% of anastrozole and 32.6% of tamoxifen patients achieved a complete response [CR] or partial response [PR]). Clinical benefit (CR + PR + stabilization of > or = 24 weeks) rates were 56.2% and 55.5% for patients receiving anastrozole and tamoxifen, respectively. Both treatments were well tolerated. However, incidences of thromboembolic events and vaginal bleeding were reported in fewer patients treated with anastrozole than with tamoxifen (4.8% v 7.3% [thromboembolic events] and 1.2% v 2.4% [vaginal bleeding], respectively).
Conclusion: Anastrozole satisfied the predefined criteria for equivalence to tamoxifen. Together with the lower observed incidence of thromboembolic events and vaginal bleeding, these findings indicate that anastrozole should be considered as first-line therapy for postmenopausal women with ABC.

🟨 Participants 🟦 Intervention and Control 🟩 Outcomes

study_name	intervention	Control	Outcome	Ee	Ne	Ec	Nc
Bonneterre et al. (2000)	anastrozole	tamoxifen	objective response	112	340	107	328
Bonneterre et al. (2000)	anastrozole	tamoxifen	clinical benefit	191	340	182	328
Bonneterre et al. (2000)	anastrozole	tamoxifen	thromboembolic events	16	340	24	328
Bonneterre et al. (2000)	anastrozole	tamoxifen	vaginal bleeding	4	340	8	328

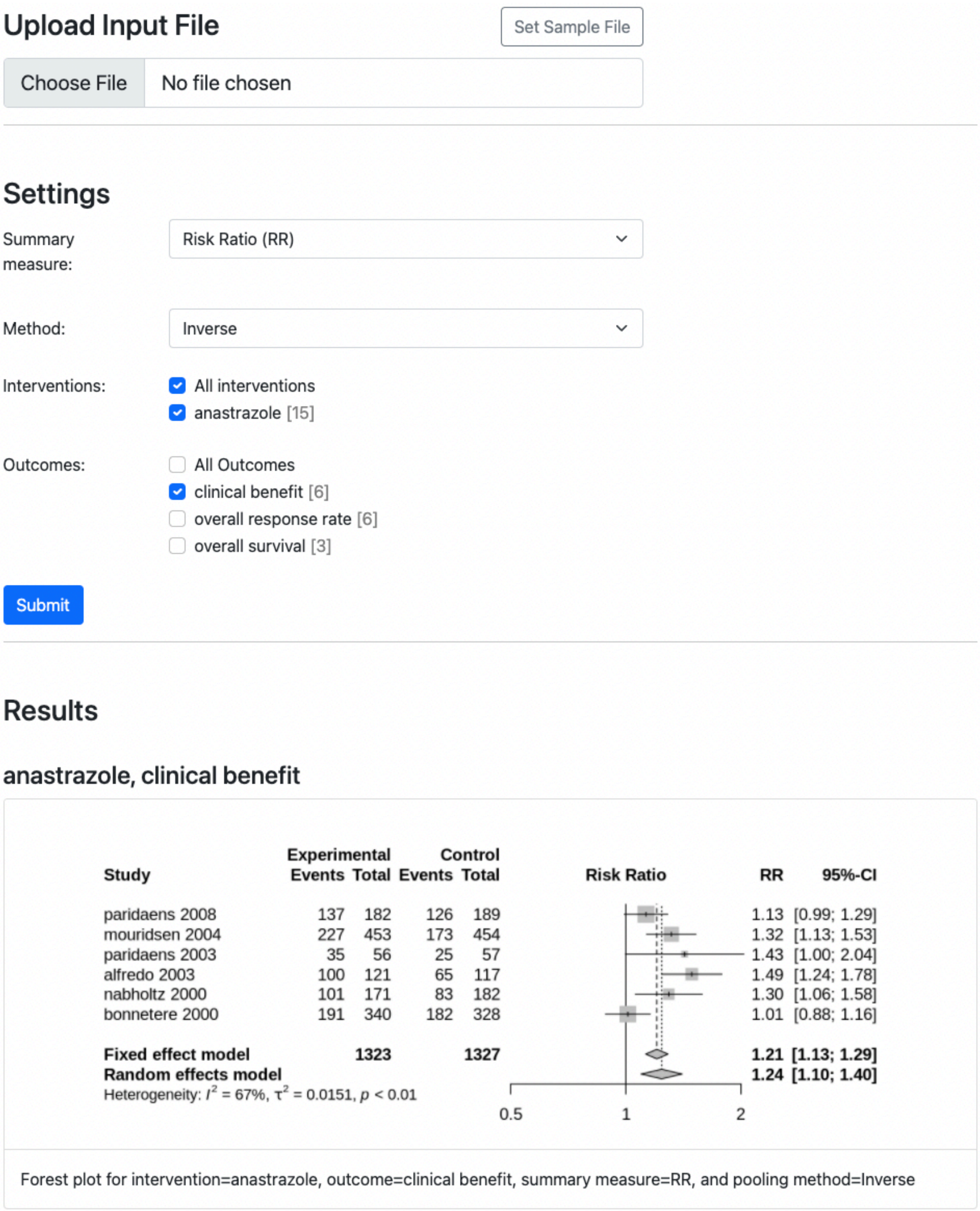
Figure 7: A sample abstract with PICO elements highlighted. The top part shows the abstract while the bottom part shows the PICO elements transformed into a structured format.

shown and the average effect is shown at the bottom of the plot. Also, in the forest plot, each study is represented by a square whose area represents the weight of the study in the meta-analysis and horizontal line (95% confidence interval).

When using the visualization system, shown in Figure 8, a user first uploads a csv file. The file must contain columns for study_name, intervention, control, outcome, Ee, Ne, Ec, and Nc as shown in the bottom part of Figure 7. After uploading the file, the user then selects a summary measure and a method for pooling the studies. The available summary measures include risk ratio, odds ratio, and risk difference which are commonly used for binary outcomes. The available pooling methods include inverse variance (Inverse), Mantel-Haenszel (MH), Peto, generalised linear mixed model (GLMM), and sample size method (SSW). For risk ratio and risk difference, only the Inverse or MH pooling methods are used. For odds ratio, inverse, MH, Peto, GLMM, or SSW pooling methods are used. In addition, the user selects the interventions and outcomes for which they would like the results to be visualized. The system groups together similar studies depending on the selected intervention(s) and outcome(s), computes the summary statistics, and returns forest plots. Each forest plot is a summary of studies with the same intervention and the same outcome.

4.3 Results and discussion

Even though automatic extraction of PICO elements from abstracts has been studied widely, only a few studies have attempted extraction of numeric texts that identify the number of patients experiencing specific outcomes. We developed a rule-based approach (discussed in Section 4.2.2 above) to parse numeric texts to identify the patients having certain outcomes. The rule-based approach was able to extract outcomes and their events from 77% of the outcome sentences in the gold test set. The rule-based approach however cannot extract outcomes and their events in cases where the outcomes and events are reported in different sentences or in studies other than double-arm studies (one intervention group and one control group).



Upload Input File

No file chosen

Settings

Summary measure: Risk Ratio (RR)

Method: Inverse

Interventions:
☑ All interventions
☑ anastrazole [15]

Outcomes:
☐ All Outcomes
☑ clinical benefit [6]
☐ overall response rate [6]
☐ overall survival [3]

Results

anastrazole, clinical benefit

Study	Experimental Events	Experimental Total	Control Events	Control Total	RR	95%-CI
paridaens 2008	137	182	126	189	1.13	[0.99; 1.29]
mouridsen 2004	227	453	173	454	1.32	[1.13; 1.53]
paridaens 2003	35	56	25	57	1.43	[1.00; 2.04]
alfredo 2003	100	121	65	117	1.49	[1.24; 1.78]
nabholtz 2000	101	171	83	182	1.30	[1.06; 1.58]
bonnetere 2000	191	340	182	328	1.01	[0.88; 1.16]
Fixed effect model		**1323**		**1327**	**1.21**	**[1.13; 1.29]**
Random effects model					**1.24**	**[1.10; 1.40]**

Heterogeneity: $I^2 = 67\%$, $\tau^2 = 0.0151$, $p < 0.01$

Forest plot for intervention=anastrazole, outcome=clinical benefit, summary measure=RR, and pooling method=Inverse

Figure 8: Visualization system interface

4.3.1 System evaluation

To evaluate the performance of the proposed system, we selected two published meta-analyses and used our system to reproduce the results. The first selected meta-analysis was conducted by Feng et al. [21] and examines the effect of platinum-based neoadjuvant chemotherapy on resectable triple-negative breast cancer patients. The meta-analysis consists of nine studies, Alba et al. [1], Ando et al. [2], Gluz et al. [22], Loibl et al. [38], Sikov et al. [52], Tung et al. [56], Minckwitz et al. [58], Wu et al. [61], and Zhang et al. [65]. The results are shown in Table 5(a). The NER model successfully extracted data from the abstracts of the nine studies. There was a NER model prediction error in one study as shown in bold underlined text in Table 5(a). For the study Gluz et al. (2018) and pathological complete response outcome, the model misclassified Ne as Nc and vice-versa. In this research, the Ee and Ec values were reported as percentage values. The absolute values of Ee and Ec were therefore calculated based on the Ne and Nc values. Since the system extracted Ne and Nc values were incorrect, the calculated Ee and Ec values were also incorrect.

Although the NER model had high accuracy, there were other factors that prevented the full reproduction of the meta-analysis. The italic and underlined texts represent studies where extra post-processing steps were required. For instance, for the studies Loibl et al. (2018) and Sikov et al. (2015), and pathological complete response, the studies have multiple intervention and control groups. The Gluz et al. (2018) and Minckwitz et al. (2014) studies, for the pathological complete response outcome, the abstracts report results for different sub-groups. The current system considers only double-arm studies (studies with one intervention group and one control group) and does not perform subgroup analysis, and these will be one of our important future works. Moreover, in some studies, the total number of participants in the intervention and control groups (Ne and Nc) were not reported in the abstracts. The studies where the numbers were not reported are indicated as NA in Table 5(a). In the Sikov et al. (2015) and Tung et al. (2020) studies, we were not able to calculate the absolute values for Ee and Ec because their calculation depends on the Ne and Nc values which were not reported in the abstracts.

The second selected meta-analysis was conducted by Xu et al. [63] and compares aromatase inhibitor to tamoxifen in women with advanced breast cancer. The meta-analysis consists of six studies, Bonneterre et al. [7], Nabholtz et al. [42], Alfredo et al. [40], Paridaens et al. [44, 45], and Mouridsen et al. [41]. The outcomes include overall response rate, clinical benefit, and overall survival. The results are as shown in Table 5(b). The NER model successfully extracted data from the abstracts of the six studies. There was only one NER model prediction error as shown in bold underlined text in Table 5(b). For the study Bonneterre et al. (2000) and overall response rate outcome, the model misclassified control events (Ec) as intervention events (Ee).

Similarly to the first selected meta-analysis, despite the great accuracy of the NER model, several issues hindered a complete replication. In some studies, the total number of participants in the intervention and control groups (Ne and Nc) were not reported in the abstracts as indicated as NA in Table 5(b). The italic and underlined texts represents studies where extra post-processing steps were required. For instance, for study Mouridsen et al. (2004) and overall response rate outcome, our system identified two intervention events and two control events. In the Mouridsen et al. (2004) study, the abstract reports the overall response rate for two sub-groups i.e., younger participants (<70 years) and older participants (>70 years). Also, in the Alfredo et al. (2003) study and overall survival outcome, the abstract reported the number of patients that died, and hence extra calculation to find the overall survival is required.

Table 5: Results of selected meta-analyses

(a) Results of first selected meta-analysis conducted by Feng et al. [21]

		Gold values				System extracted values			
Study	**Outcome**	**Ee**	**Ne**	**Ec**	**Nc**	**Ee**	**Ne**	**Ec**	**Nc**
Alba et al. (2012)	pathological complete response	14	47	16	46	14	48	16	46
Ando et al. (2014)		23	37	10	38	23	37	10	38
Gluz et al. (2018)		70	154	52	182	**44, 30**	**182**	**84, 81**	**154**
Loibl et al. (2018)		92	160	49	158	92, 168	160	49	158
Sikov et al. (2015)		60	110	43	105	60%, 59%, 54%	NA	44%, 48%, 41%	NA
Tung et al. (2020)		9	40	10	36	18%	NA	26%	NA
Minckwitz et al. (2014)		90	158	67	157	129, 84, 45	137, 158	108, 58, 50	136, 157
Wu et al. (2018)		24	62	8	63	24	62	8	63
Zhang et al. (2016)		18	47	6	44	18	47	6	44
Alba et al. (2012)	objective response rate	36	47	32	46	37	48	32	46
Wu et al. (2018)		58	62	46	63	58	62	46	63
Zhang et al. (2016)		42	47	34	44	42	47	34	44

Ee is the number of events in the intervention group, Ne is the number of participants in the control group, Ec is the number of events in the control group, and Nc is the number of participants in the control group. NA indicates where the information was not available in the abstract. **Bold underlined texts** are NER model prediction errors while *italic underlined texts* are values where extra pre-processing was required.

(b) Results of second selected meta-analysis conducted by Xu et al. [63]

		Gold values				System extracted values			
Study	**Outcome**	**Ee**	**Ne**	**Ec**	**Nc**	**Ee**	**Ne**	**Ec**	**Nc**
Bonneterre et al. (2000)	overall response rate	112	340	107	328	111.86, **106.928**	340	—	328
Nabholtz et al. (2000)		36	171	31	182	21%	NA	17%	NA
Alfredo et al. (2003)		43	121	31	117	NA	NA	NA	NA
Paridaens et al. (2003)		25	56	10	57	41%	NA	17%	NA
Mouridsen et al. (2004)		137	453	92	454	*117.78, 172.14*	453	*99.88, 81.72*	454
Paridaens et al. (2008)		83	182	59	189	83.72	182	58.59	189
Bonneterre et al. (2000)	clinical benefit	191	340	182	328	191.08	340	182.04	328
Nabholtz et al. (2000)		101	171	83	182	59%	NA	46%	NA
Alfredo et al. (2003)		100	121	65	117	100.43	121	65.52	117
Paridaens et al. (2003)		35	56	25	57	57%	NA	42%	NA
Mouridsen et al. (2004)		227	453	173	454	NA	453	NA	454
Paridaens et al. (2008)		137	182	126	189	NA	NA	NA	NA
Alfredo et al. (2003)	overall survival	48	121	13	117	*72.6*	121	*104.130*	117
Mouridsen et al. (2004		368	453	322	454	NA	NA	NA	NA
Paridaens et al. (2008)		100	182	106	189	NA	NA	NA	NA

Ee is the number of events in the intervention group, Ne is the number of participants in the control group, Ec is the number of events in the control group, and Nc is the number of participants in the control group. NA indicates where the information was not available in the abstract. **Bold underlined texts** are NER model prediction errors while *italic underlined texts* are values where extra pre-processing was required.

4.4 Conclusion

This research proposed a system for automating meta-analysis statistical analysis. The proposed system extracts PICO elements from research abstracts, parses numeric outcomes to extract the number of patients experiencing certain outcomes, transforms the extracted information into a structured format, performs statistical analysis, and visualizes the results in forest plots. We evaluated the performance of the system by attempting to reproduce the results of existing meta-analyses. The system extracted PICO elements from the studies with high accuracy. The statistical analysis step did not perform well owing to lack of some information in the abstracts and lack of uniformity in the research abstracts were some abstracts required extra pre-processing. These results however show that there is potential to automate these tasks and wish to motivate more research towards fully automating the entire meta-analysis process.

5. Conclusion

5.1 Limitations and future work

Our research has several limitations. The corpus created for this research consists of breast cancer related abstracts (Chapter 2), and one of the future works is to extend it to include other diseases. Even though the current corpus consists of breast cancer only articles, most of the intervention treatments, control treatment, outcomes, and outcome measures are commonly used in the other types of cancers. The corpus can be therefore be extended to include other types of diseases/cancers by employing machine learning techniques.

The corpus in this study consists of abstracts only and as seen in Section 4.3.1, abstracts sometimes lack information that are present in the full-text document. For instance, a manual check of our corpus found that a significant number of abstracts do not mention the number of participants in the intervention and control groups. This presents a challenge when determining the number of patients having certain outcomes for statistical analysis (Section 4.3.1). We also do not account for participants who drop out of a study and this might affect the final results. For future work, it is important to consider extracting information from full-text articles.

We proposed a rule-based system for matching outcomes and their events (Section 4.2.2). The rule-based approach considers only double-arm studies, i.e., studies with one intervention group and one control group. Single-arm studies and studies with more than multiple intervention or control groups are ignored. In future, it is necessary to explore other approaches such as relation extraction.

In the statistical analysis step, we consider only binary outcomes. The summary statistics (odds ratio, risk ratio, and risk difference) used in our results visualization system (Section 4.2.3) are only focused on binary outcomes. Incorporating continuous outcomes and their summary statistics is important future work. In addition, the current approach calculates the summary statistics from absolute values. A review of the corpus revealed that some of the abstracts report the summary statistics that have already been computed. Annotation of these already calculated summary statistics and incorporating them to the current system is a challenging task but an important future work.

Moreover, some meta-analyses perform subgroup analysis where they compare the results of different subgroups of participants either by age or cancer type. Annotation and incorporation of such information is also necessary in future. Finally, we assessed the performance of the proposed system by replicating the results of two existing meta-studies (Section 4.3.1). To substantiate the usefulness of the system, it is important to test it on larger and more complex meta-studies.

5.2 Summary

In this dissertation, we proposed a system for automating data extraction to support meta-analysis statistical analysis. Our objective is to provide a system that automates data extraction and statistical analysis, to shorten the time it takes to carry out a meta-analysis and allow for automatic updates when new results becomes available. The proposed system extracts PICO elements from research abstracts, parses numeric outcomes to extract the number of patients experiencing certain outcomes, transforms the extracted information into a structured format, performs statistical analysis, and visualizes the results in forest plots. We evaluated the performance of the system by attempting to reproduce the results of existing meta-analyses. The system extracted PICO elements from the studies with high accuracy. The statistical analysis step did not perform well owing to lack of some information in the abstracts and lack of uniformity in the research abstracts were some abstracts required extra pre-processing. These results however show that there is potential to automate these tasks and wish to motivate more research towards fully automating the entire meta-analysis process.

In addition, we created a publicly available corpus with detailed annotation of the PICO elements. The corpus contains 1011 abstracts related to breast cancer RCTs. The corpus provides detailed annotation for outcomes especially numeric texts to identify the number of participants having certain outcomes. This is important for statistical analysis to determine the effectiveness of a treatment. The corpus will facilitate NLP research on automatic information extraction from biomedical literature and contribute towards evidence-based medicine. The corpus is publicly available at `https://github.com/sociocom/PICO-Corpus`.

Acknowledgements

I would like to express my sincere and deep gratitude to my supervisors, Professor Eiji Aramaki, Associate Professor Shoko Wakamiya, Assistant Professor Shuntaro Yada, and Assistant Professor Kongmeng Liew for their guidance throughout my research and creating a perfect atmosphere for research. Great thanks to Professor Taro Watanabe for taking his time to reviewing my dissertation. I would also like to thank Kaneko san for her support with administrative work. Great thanks also to my fellow labmates who supported me in my daily life.

I would like to express deep gratitude to MEXT for their monetary support which enabled me pursue my study in Japan comfortably. I would also like to thank NAIST International Student Affairs for their great support with administrative and daily life. Lastly, I thank my dearest family and friends for their love and encouragement.

Bibliography

[1] Emilio Alba, JI Chacon, A Lluch, A Anton, L Estevez, B Cirauqui, E Carrasco, L Calvo, MA Segui, N Ribelles, et al. A randomized phase ii trial of platinum salts in basal-like breast cancer patients in the neoadjuvant setting. results from the geicam/2006-03, multicenter study. *Breast cancer research and treatment*, 136(2):487–493, 2012.

[2] Masashi Ando, Hideko Yamauchi, Kenjiro Aogi, Satoru Shimizu, Hiroji Iwata, Norikazu Masuda, Naohito Yamamoto, Kenichi Inoue, Shinji Ohono, Katsumasa Kuroi, et al. Randomized phase ii study of weekly paclitaxel with and without carboplatin followed by cyclophosphamide/epirubicin/5-fluorouracil as neoadjuvant chemotherapy for stage ii/iiia breast cancer without her2 overexpression. *Breast cancer research and treatment*, 145(2):401–409, 2014.

[3] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[4] Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326, 2010.

[5] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[6] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1):211–231, 1999.

[7] J Bonneterre, B Thürlimann, JFR Robertson, M Krzakowski, L Mauriac, P Koralewski, Ignace Vergote, A Webster, M Steinberg, M Von Euler, et al. Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: results of the tamoxifen or arimidex randomized group efficacy and tolerability study. *Journal of Clinical Oncology*, 18 (22):3748–3757, 2000.

[8] Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7 (2):e012545, 2017.

[9] Florian Boudin, Lixin Shi, and Jian-Yun Nie. Improving medical information retrieval with pico element detection. In *European Conference on Information Retrieval*, pages 50–61. Springer, 2010.

[10] Jon Brassey, Christopher Price, Jonny Edwards, Markus Zlabinger, Alexandros Bampoulidis, and Allan Hanbury. Developing a fully automated evidence synthesis tool for identifying, assessing and collating the evidence. *BMJ Evidence-Based Medicine*, 26(1):24–27, 2021.

[11] Duy Duc An Bui, Guilherme Del Fiol, John F Hurdle, and Siddhartha Jonnalagadda. Extractive text summarization system to aid data extraction from full text in systematic review development. *Journal of biomedical informatics*, 64:265–272, 2016.

[12] S Chabou and M Iglewski. Pico extraction by combining the robustness of machine-learning methods with the rule-based methods. In *2015 World Congress on Information Technology and Computer Applications (WCITCA)*, pages 1–4. Ieee, 2015.

[13] Grace Y Chung. Sentence retrieval for abstracts of randomized controlled trials. *BMC medical informatics and decision making*, 9(1):1–13, 2009.

[14] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL `https://doi.org/10.1177/001316446002000104`.

[15] Deborah J Cook, Cynthia D Mulrow, and R Brian Haynes. Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine*, 126(5):376–380, 1997.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[17] Berry De Bruijn, Simona Carini, Svetlana Kiritchenko, Joel Martin, and Ida Sim. Automated information extraction of key trial design elements from clinical trial publications. In *AMIA Annual Symposium Proceedings*, volume 2008, page 141. American Medical Informatics Association, 2008.

[18] Dina Demner-Fushman and Jimmy Lin. Knowledge extraction for clinical question answering: Preliminary results. In *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 9–13. AAAI Press (American Association for Artificial Intelligence) Pittsburgh, PA, 2005.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[21] Wuna Feng, Yujing He, Jingsi Xu, Hongya Zhang, Yuexiu Si, Jiaxuan Xu, and Shengzhou Li. A meta-analysis of the effect and safety of platinum-based neoadjuvant chemotherapy in treatment of resectable triple-negative breast cancer. *Anti-cancer drugs*, 33(1):e52–e60, 2022.

[22] Oleg Gluz, Ulrike Nitz, Cornelia Liedtke, Matthias Christgen, Eva-Maria Grischke, Helmut Forstbauer, Michael Braun, Mathias Warm, John Hackmann, Christoph Uleer, et al. Comparison of neoadjuvant nab-paclitaxel+ carboplatin vs nab-paclitaxel+ gemcitabine in triple-negative breast cancer: randomized wsg-adapt-tn trial results. *JNCI: Journal of the National Cancer Institute*, 110(6):628–637, 2018.

[23] S Gopalakrishnan and P Ganeshkumar. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of family medicine and primary care*, 2(1):9, 2013.

[24] Marie J Hansen, Nana Ø Rasmussen, and Grace Chung. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. *Journal of Telemedicine and Telecare*, 14(7):354–358, 2008.

[25] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[26] Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2308. URL `https://aclanthology.org/W18-2308`.

[27] Di Jin and Peter Szolovits. Pico element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, 2018.

[28] Di Jin and Peter Szolovits. Advancing pico element detection in biomedical text via deep neural networks. *Bioinformatics*, 36(12):3856–3862, 2020.

[29] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[30] Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):1–16, 2015.

[31] Tian Kang, Shirui Zou, and Chunhua Weng. Pretraining to recognize pico elements from randomized controlled trial literature. *Studies in health technology and informatics*, 264:188, 2019.

[32] Cassidy Kelly and Hui Yang. A system for extracting study design parameters from nutritional genomics abstracts. *Journal of integrative bioinformatics*, 10(2):82–93, 2013.

[33] Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central, 2011.

[34] Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1): 1–17, 2010.

[35] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[36] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36 (4):1234–1240, 2020.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[38] Sibylle Loibl, Joyce O'Shaughnessy, Michael Untch, William M Sikov, Hope S Rugo, Mark D McKee, Jens Huober, Mehra Golshan, Gunter von Minckwitz, David Maag, et al. Addition of the parp inhibitor veliparib plus carboplatin or carboplatin alone to standard neoadjuvant chemotherapy in triple-negative breast cancer (brightness): a randomised, phase 3 trial. *The lancet oncology*, 19(4):497–509, 2018.

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

[40] Alfredo Milla-Santos, Lidon Milla, Jordi Portella, Lidon Rallo, Maria Pons, Esther Rodes, Jose Casanovas, and Margarita Puig-Gali. Anastrozole versus tamoxifen as first-line therapy in postmenopausal patients with hormone-dependent advanced breast cancer: a prospective, randomized, phase iii study. *American journal of clinical oncology*, 26(3):317–322, 2003.

[41] H Mouridsen and HA Chaudri-Ross. Efficacy of first-line letrozole versus tamoxifen as a function of age in postmenopausal women with advanced breast cancer. *The oncologist*, 9(5):497–506, 2004.

[42] JM Nabholtz, A Buzdar, M Pollak, W Harwin, G Burton, A Mangalik, M Steinberg, A Webster, and M Von Euler. Anastrozole is superior to tamoxifen as first-line therapy for advanced breast cancer in postmenopausal women: results of a north american multicenter randomized trial. *Journal of Clinical Oncology*, 18(22):3758–3767, 2000.

[43] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[44] Robert Paridaens, L Dirix, C Lohrisch, LVAM Beex, Marianne Nooij, D Cameron, Laura Biganzoli, T Cufer, Luc Duchateau, Andrew Hamilton, et al. Mature results of a randomized phase ii multicenter study of exemestane versus tamoxifen as first-line hormone therapy for postmenopausal women with metastatic breast cancer. *Annals of Oncology*, 14(9):1391–1398, 2003.

[45] Robert J Paridaens, Luc Y Dirix, Louk V Beex, Marianne Nooij, David A Cameron, Tanja Cufer, Martine J Piccart, Jan Bogaerts, and Patrick Therasse. Phase iii study comparing exemestane with tamoxifen as first-line hormonal treatment of metastatic breast cancer in postmenopausal women: the european organisation for research and treatment of cancer breast cancer cooperative group. *Journal of Clinical Oncology*, 26(30):4883, 2008.

[46] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

[47] Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Leo Anthony Celi. Acronym identification and disambiguation

shared tasks for scientific document understanding. *arXiv e-prints*, pages arXiv–2012, 2020.

[48] Richeek Pradhan, David C Hoaglin, Matthew Cornell, Weisong Liu, Victoria Wang, and Hong Yu. Automatic extraction of quantitative data from clinicaltrials. gov to conduct meta-analyses. *Journal of clinical epidemiology*, 105:92–100, 2019.

[49] David L Sackett. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier, 1997.

[50] Guido Schwarzer et al. meta: An r package for meta-analysis. *R news*, 7(3): 40–45, 2007.

[51] Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233, 2007.

[52] William M Sikov, Donald A Berry, Charles M Perou, Baljit Singh, Constance T Cirrincione, Sara M Tolaney, Charles S Kuzma, Timothy J Pluard, George Somlo, Elisa R Port, et al. Impact of the addition of carboplatin and/or bevacizumab to neoadjuvant once-per-week paclitaxel followed by dose-dense doxorubicin and cyclophosphamide on pathologic complete response rates in stage ii to iii triple-negative breast cancer: Calgb 40603 (alliance). *Journal of clinical oncology*, 33(1):13, 2015.

[53] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi 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*, pages 102–107, 2012.

[54] Rodney Summerscales, Shlomo Argamon, Jordan Hupert, and Alan Schwartz. Identifying treatments, groups, and outcomes in medical abstracts. In *The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009)*. Indiana University Bloomington, IN, USA, 2009.

[55] Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377. IEEE, 2011.

[56] Nadine Tung, Banu Arun, Michele R Hacker, Erin Hofstatter, Deborah L Toppmeyer, Steven J Isakoff, Virginia Borges, Robert D Legare, Claudine Isaacs, Antonio C Wolff, et al. Tbcrc 031: randomized phase ii study of neoadjuvant cisplatin versus doxorubicin-cyclophosphamide in germline brca carriers with her2-negative breast cancer (the inform trial). *Journal of Clinical Oncology*, 38(14):1539, 2020.

[57] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13 (2):260–269, 1967.

[58] Gunter Von Minckwitz, Andreas Schneeweiss, Sibylle Loibl, Christoph Salat, Carsten Denkert, Mahdi Rezai, Jens U Blohmer, Christian Jackisch, Stefan Paepke, Bernd Gerber, et al. Neoadjuvant carboplatin in patients with triple-negative and her2-positive early breast cancer (geparsixto; gbg 66): a randomised phase 2 trial. *The lancet oncology*, 15(7):747–756, 2014.

[59] Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17 (1):4572–4596, 2016.

[60] Lucy Lu Wang and Kyle Lo. Text mining approaches for dealing with the rapidly expanding literature on covid-19. *Briefings in Bioinformatics*, 22(2): 781–799, 2021.

[61] Xiujuan Wu, Peng Tang, Shifei Li, Shushu Wang, Yueyang Liang, Ling Zhong, Lin Ren, Ting Zhang, and Yi Zhang. A randomized and open-label phase ii trial reports the efficacy of neoadjuvant lobaplatin in breast cancer. *Nature communications*, 9(1):1–8, 2018.

[62] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[63] Hong-Bin Xu, Yu-Jin Liu, and Ling Li. Aromatase inhibitor versus tamoxifen in postmenopausal woman with advanced breast cancer: a literature-based meta-analysis. *Clinical breast cancer*, 11(4):246–251, 2011.

[64] Rong Xu, Yael Garten, Kaustubh S Supekar, Amar K Das, Russ B Altman, Alan M Garber, et al. Extracting subject demographic information from abstracts of randomized clinical trial reports. *Medinfo*, 129:550–4, 2007.

[65] Pin Zhang, Yi Yin, Hongnan Mo, Bailin Zhang, Xiang Wang, Qing Li, Peng Yuan, Jiayu Wang, Shan Zheng, Ruigang Cai, et al. Better pathologic complete response and relapse-free survival after carboplatin plus paclitaxel compared with epirubicin plus paclitaxel as neoadjuvant chemotherapy for locally advanced triple-negative breast cancer: a randomized phase 2 trial. *Oncotarget*, 7(37):60647, 2016.

[66] Jin Zhao, Praveen Bysani, and Min-Yen Kan. Exploiting classification correlations for the extraction of evidence-based practice information. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1070. American Medical Informatics Association, 2012.

List of Publications

1. Mutinda, F. W., Liew K., Yada, S., Wakamiya, S., & Aramaki, E. (2022). PICO Corpus: A Publicly Available Corpus to Support Automatic Data Extraction from Biomedical Literature. In Proceedings of the First Workshop on Information Extraction from Scientific Publications. Asia-Pacific Chapter of the Association for Computational Linguistics (pp. 26-31).

2. Mutinda, F. W., Yada, S., Wakamiya, S., & Aramaki, E. (2022). AUTOMETA: Automatic Meta-Analysis System Employing Natural Language Processing. In MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation (pp. 612-616). IOS Press.

3. Mutinda, F. W., Liew, K., Yada, S., Wakamiya, S., & Aramaki, E. (2022). Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer. BMC Medical Informatics and Decision Making, 22(1), 1-13.

4. Mutinda, F. W., Yada, S., Wakamiya, S., & Aramaki, E. (2021). Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT. Methods of Information in Medicine, 60(S 01), e56-e64.

5. Raithel, L.*, Mutinda, F. W.*, Andrade, G. H. B.*, Yeh, H. S. *, Nishiyama, T., Laï-King, M., Yada, S., Roller, R., Grouin, C., Savary, A., Névéol, A., Lavergne, T., Aramaki, E., Möller, S., Matsumoto, Y., & Zweigenbaum, P. (2022): KEEPHA at n2c2 2022: Track 1 Contextualized Medication Event Extraction, 2022 n2c2 Shared Task and Workshop.