Automation of Title and Abstract Screening for Clinical Systematic Reviews

by

Mihir Prafullsinh Parmar

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2021 by the
Graduate Supervisory Committee:

Chitta Baral, Co-Chair
Murthy Devarakonda, Co-Chair
Irbaz Bin Riaz

ARIZONA STATE UNIVERSITY

August 2021

ABSTRACT

Systematic Reviews (SRs) aim to synthesize the totality of evidence for clinical practice and are important in making clinical practice guidelines and health policy decisions. However, conducting SRs manually is a laborious and time-consuming process. This challenge is growing due to the increase in the number of databases to search and the papers being published. Hence, the automation of SRs is an essential task. The goal of this thesis work is to develop Natural Language Processing (NLP)-based classifiers to automate the title and abstract-based screening for clinical SRs based on inclusion/exclusion criteria. In clinical SRs, a high-sensitivity system is a key requirement. Most existing methods for SRs use binary classification systems trained on labeled data to predict inclusion/exclusion. While previous studies have shown that NLP-based classification methods can automate title and abstract-based screening for SRs, methods for achieving high-sensitivity have not been empirically studied. In addition, the training strategy for binary classification has several limitations: (1) it ignores the inclusion/exclusion criteria, (2) lacks generalization ability, (3) suffers from low resource data, and (4) fails to achieve reasonable precision at high-sensitivity levels.

This thesis work presents contributions to several aspects of the clinical systematic review domain. First, it presents an empirical study of NLP-based supervised text classification and high-sensitivity methods on datasets developed from six different SRs in the clinical domain. Second, this thesis work provides a novel approach to view SR as a Question Answering (QA) problem in order to overcome the limitations of the binary classification training strategy; and propose a more general abstract screening model for different SRs. Finally, this work provides a new QA-based dataset for six different SRs which is made available to the community.

DEDICATION

Dedicated to, my loving parents, family and friends for the love, patience, and faith in this short journey and in much more to come...

ACKNOWLEDGEMENTS

Writing this thesis is a remarkable journey of my graduate life in terms of learning from experts, exploring new arenas, improving my research abilities and much more. I would like to acknowledge everyone who played a role in my academic accomplishments. I take this opportunity to convey my sincere gratitude to all those who helped me directly and indirectly for making this Master of Science thesis possible. Foremost, I would like to thank my supervisors, Dr. Chitta Baral and Dr. Murthy Devarakonda for their guidance and motivation through each phase of this thesis work, which made me enthusiastic and confident about proper execution and completion of this thesis work. Moreover, I would like to thank Dr. Chitta Baral for providing the best research environment and high-quality equipment in the lab. I would like to extend my special gratitude towards Mayo Clinic and team, Dr. Irbaz Bin Riaz (MD) and Dr. M. Hassan Murad (MD) for providing high-quality research data.

I would also like to express my gratitude to my colleagues and researchers behind this thesis, Ms. Man Luo, Mr. Hong Guan, Mr. Ashwin Karthik Ambalavanan, and Mr. Rishab Banerjee. I would like to thank them all for their outlines, motivation, and supervision to give excellent shape to my thesis. I am grateful to all the members of the Cogint Lab, Arizona State University. I would also like to thank my former colleague, Ms. Mirali Purohit for her help in keeping me motivated for this thesis work.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

1.1 Systematic Review 2

1.2 Motivation 4

1.3 Research Value and Contributions 5

1.4 Structure of Thesis 7

2 LITERATURE REVIEW 8

2.1 Systematic Reviews 8

2.2 Biomedical Tasks as Question-Answering 11

3 DATASETS 12

3.1 Data sources and Search Strategies 12

3.2 Manual Screening Process 13

3.3 Statistical Analysis 14

3.4 Inclusion and Exclusion Criteria 15

4 EMPIRICAL STUDY OF HIGH-SENSITIVITY METHODS 16

4.1 Background 16

4.2 Methods 18

4.2.1 Abstract-Screening Methods 18

4.2.2 High-sensitivity Techniques 20

4.2.3 Experiments 22

4.3 Results 24

4.4 Discussion 29

CHAPTER Page

4.5 Chapter Summary 31
5 SYSTEMATIC REVIEW AS QUESTION ANSWERING 32
5.1 Background 32
5.2 Methods 34
5.2.1 Abstract Screening Systems 34
5.2.2 Experiments 37
5.3 Results and Discussion 39
5.3.1 P-R Curves 41
5.3.2 Analysis 42
5.4 Chapter Summary 44
6 CONCLUSIONS 46
6.1 Summary 46
6.2 Limitations and Future Research 46
REFERENCES 49
APPENDIX
A DATASETS 56
A.1 INCLUSION/EXCLUSION CRITERIA 57
A.2 DATA SAMPLES 58
B EMPIRICAL STUDY OF HIGH-SENSITIVITY METHODS 59
B.1 QUERIES FOR BM25 60

LIST OF TABLES

Table Page

3.1 Statistics of Datasets Developed from Six Different SRs 14

4.1 Optimum F-measure Results for Supervised Text Classification (SL), Information Retrieval - BM25, and Hybrid Approaches - BM25+SciBERT and SciBERT+BM25. All the Cross-validation Results Are Presented in the Table Has Mean (Standard Deviation) Format. **Bold** Results Indicate the Best Method for Each Dataset. 25

4.2 Results for High-sensitivity Strategies - Th: Lowering the Probability Threshold for Inclusion, USP: Up-Sampling Positive, DSN: Down-Sampling Negative, USP and DSN: A Combination of Up-Sampling Positive and Down-Sampling Negative, and CSL: Cost-Sensitive Learning. All the Cross-validation Results Are Presented in the Table Has a Mean (Standard Deviation) Format. Highlighted Results Indicate the Best Method and Bold Results Indicate the Second-best Method for Each Dataset. 26

4.3 % Of Eventually Included a Study That Is Missed after Abstract-screening at Various High-Sensitivity along with Corresponding Precision. For Each Dataset, the First Row Presents the Sensitivity at Optimum F-measure, and the Last Row Represents the Maximum Achievable Sensitivity for Each Dataset. 29

5.1 Set of Boolean Questions Created Manually from the Inclusion Criteria for Each Dataset 36

Table Page

5.2 The Comparison of Optimum F-measure (First Block) and High-sensitivity (Second Block) Results Between Baseline BCM and Proposed Models (QA and General-QA). The Third Block Presents Relative Performance (RP) of Proposed Models Compared to Baseline at High-sensitivity (Green Highlighted % Indicates Improvement and Red Highlighted % Indicates Degradation). P: Precision, R: Sensitivity, F: F-measure, BCM: Binary Classification Model, QA: Question-Answering Model and General-QA: QA Model Trained on Six Datasets. **Bold** Results Indicate the Best Method for Each Dataset 39

5.3 Optimum F-measure Results for Pre-training and Fine-tuning Experiments. I: ICT, H: HRT, C: Cooking, ACC: Accelerometer, ACR: Acromegaly, COV: COVID. * Means Fine-tuning a Model on a New Dataset Improves the Performance on Other Datasets 41

5.4 Relative Performance (in %) of Optimum F-measure Results Between Fine-tuning Model Pre-trained on Other Five Datasets and Model Trained Only Using Particular Dataset (Data-Specific QA Model) 42

5.5 Optimum F-measure Results for the Data-specific QA Model to Analyze the Robustness of the Model Towards Different Question Formats. Q1 Indicates Set of Questions given in Table 5.1, and Q2 Indicates Set of New Semantically Equivalent but Syntactically Different Questions Prepared from Q1 43

A.1 Inclusion and Exclusion Criteria for Each Dataset. RCT: Randomized Control Trial, SRMA: Systematic Review and Meta Analysis 57

Table Page

A.2 Manually Annotated Positive (i.e., Include) and Negative (i.e., Exclude) Samples from ICI Dataset. Here, the Data Sample is Concatenation of Title and Abstract from the Candidate Article 58

LIST OF FIGURES

Figure Page

1.1 PRISMA Flowchart of Clinical SR Workflow. After Liberati *et al.* (2009). 3

3.1 (a) Probability and (b) Cumulative Distributions of Title and Abstract Sequences for the Datasets. 15

4.1 Schematic Representation of Supervised Text Classification (SL) Method. 19

4.2 P-R Curves for Different High-Sensitivity Methods Presented in Table 4.2. Threshold: Lowering the Probability Threshold for Inclusion, USP: Up-Sampling Positive, DSN: Down-Sampling Negative, Hybrid: Combination of Up-Sampling Positive and Down-Sampling Negative, and CSL: Cost-Sensitive Learning. 28

5.1 Schematic Representation of (a) Binary Classification Model (BCM), and (b) Proposed QA Approach. 34

5.2 P-R Curves for Different Methods Presented in Table 5.2. BCM: Binary Classification Model, QA: Question Answering Model and General-QA: General QA Model Trained on Combined Data 43

Chapter 1

INTRODUCTION

Biomedical research is the core of modern healthcare, and it is important in developing effective medication and treatments. Without this research, the prevention and cure of disease would be practically impossible, and it can help stimulate the development of healthcare and biomedical infrastructure. Recently, computer science is playing a significant role in accelerating biomedical research. With the advent of Artificial Intelligence (AI), it became easier to simulate patient behaviour and visualize the complex biological model. Nowadays, AI is playing a vital role in the automation of various biomedical tasks such as biomedical text classification, ontology harmonization, systematic reviews, and many more (Thakur *et al.* (2020)). In the last few decades, Machine Learning (ML), Deep Learning (DL), Information Retrieval (IR), Computer Vision (CV), and Natural Language Processing (NLP) have changed the face of biomedical research. Researchers are using these learning algorithms to simplify and automate many biomedical tasks. Because of the fruitful collaboration between computer science and biomedical, research in the clinical domain increased exponentially.

The growth of research in healthcare has become significant in the last few decades. With this increasing clinical research, it is difficult for busy clinicians and physicians to keep up with ongoing research in various domain. Hence, systematic study and summary of this massive abundance of studies in the various clinical domains is a necessary task. In the biomedical field, well-conducted Systematic Reviews (SRs) and Meta-Analysis (MA) are considered a feasible solutions for keeping physicians abreast to ongoing research in biomedical field. The next section describes the basic overview

of SR workflow in the clinical domain. These days, SRs are mostly carried out manually which becomes a time-consuming, laborious, and costly task of reviewing thousands of research articles. The relentless growth in clinical research and published articles have created a need for automation of SRs to expedite the process.

The main aim of this thesis work is to use the strength of recent NLP techniques to automate the SRs workflow for title and abstract based screening. In this work, we present a detailed empirical study of datasets developed from six different clinical SRs. We demonstrate the effectiveness of classical IR techniques, advanced NLP techniques and their ensemble in achieving a high-sensitivity system for automation of SRs. This work presents SR as a Question Answering (QA) problem for the first time to overcome the limitations of the binary classification training strategy, and propose a *more general* abstract screening model for different SRs. Finally, this thesis work provides a new QA-based dataset for six different SRs which will be made available to the community.

1.1 Systematic Review

Systematic Reviews (SRs) help in facilitating easy access to evidence for busy clinicians. They have become very important in the healthcare domain. SRs aim to synthesize the totality of evidence for clinical practice and are important in making clinical practice guidelines and health policy decisions. Clinical practice guidelines and health policy decisions depend on well-conducted clinical SRs (Tawfik *et al.* (2019)).

The process for conducting clinical SR is typically done in three steps: (1) search bibliographic datasets using inclusion/exclusion criteria, (2) screening of obtained articles based on their title and abstract, and (3) full-text review of an included articles from step (2) (Liberati *et al.* (2009)). PRISMA (Preferred Reporting Items for Sys-

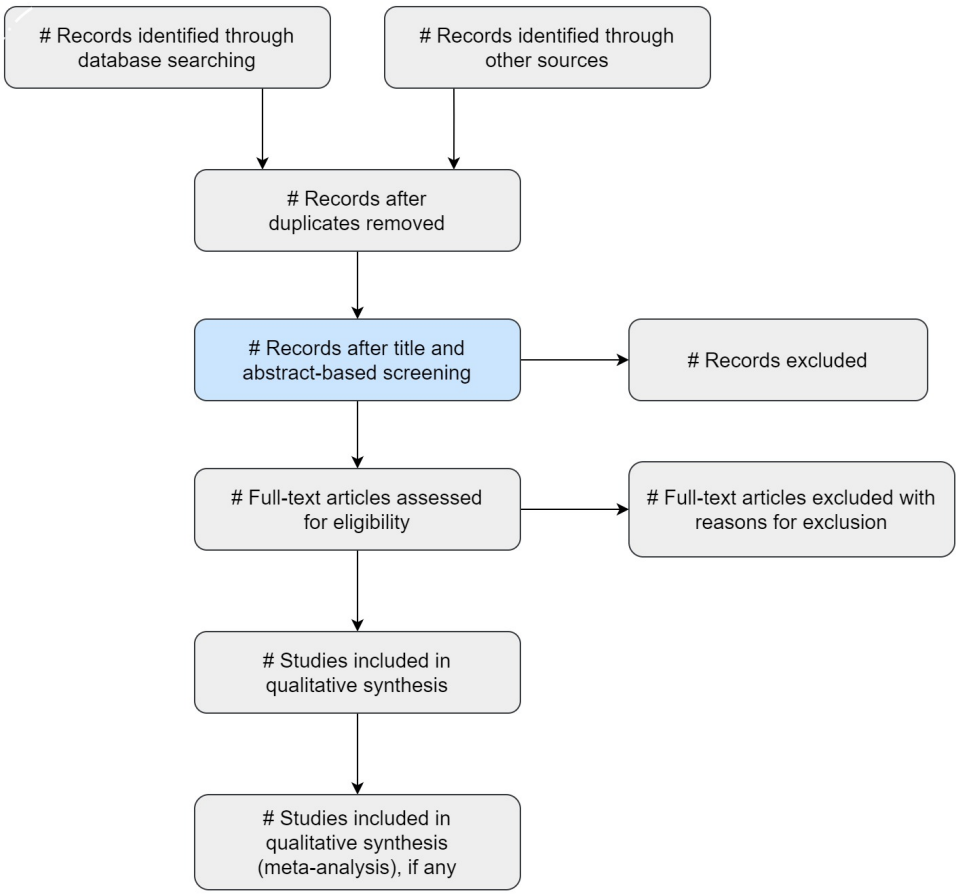


Figure 1.1: PRISMA Flowchart of Clinical SR Workflow. After Liberati *et al.* (2009).

tematic reviews and Meta-Analyses) flow diagram of clinical SR is shown in Figure 1.1. In this work, we have used data from six different clinical SRs. First, inclusion/exclusion criteria for all datasets are defined using PICO (Population, Intervention, Comparison, Outcome) approach. Detailed explanation of inclusion/exclusion criteria are given in Chapter 3. After designing inclusion/exclusion criteria, a search strategy is used to search various datasets to get relevant articles based on controlled vocabulary provided by experts. After getting articles, title and abstract based screening (simply called abstract screening in the rest of the thesis) is done by two reviewers independently (in this step, duplicate articles are also removed). As shown in Figure

1.1, only articles included from abstract screening are reviewed again using their full text for deciding final inclusion/exclusion from SR.

Research on methods for automating or semi-automating SR via ML and NLP now constitutes its own (small) subfield (Marshall and Wallace (2019)). The major work of this thesis revolves around the automation of abstract screening (highlighted block with light blue color in Figure 1.1) for clinical SR using advanced NLP techniques. As known, the US Institute of Medicine explicitly favors high sensitivity of literature searches and literature screening over high specificity (Morton *et al.* (2011)). However, the techniques needed to achieve high sensitivity have not been systematically studied before. In addition, this thesis work presents different methods to achieve high sensitivity for clinical SR.

1.2 Motivation

SRs search, appraise and collate all relevant empirical evidence to provide a complete interpretation of research results. Because of exponentially increasing research in clinical domain, we must expedite the SR process (Tsafnat *et al.* (2013)). Nowadays, abstract, and full-text screening are usually carried out manually (although some tools exist) (Marshall *et al.* (2017)). These steps require a large amount of time and a considerable workforce with expertise (Allen and Olkin (1999)). According to Allen and Olkin (1999), conducting a single review requires over 1000 hours of highly skilled labors. Moreover, existing manual SR process is not sustainable because review of current evidence goes out of date quickly. It is tiresome and challenging task to keep updating SR. Hence, automation of SR is an essential task. Moreover, clinicians do not want to miss any relevant article or study for any domain. Hence, designing systems with high sensitivity is key requirement in clinical SRs.

As mentioned in Bagheri *et al.* (2018), abstract screening is one of the most time-

consuming steps in the production of SR. At the abstract screening time, reviewers must review thousands of documents coming from various datasets. However, at the time of full-text screening, you have to review only few hundreds, or a smaller number of articles compared to abstract screening because many articles will be excluded during abstract screening. Hence, the scope of this study is limited to automate abstract-based screening of articles to reduce manual efforts.

1.3 Research Value and Contributions

This thesis work is the part of ongoing research project with Mayo Clinic. The goal of this project is to develop system that can automate or semi-automate the SR workflow and use that system to conduct "Live Systematic Review" for the real-life use for hospitals. Having the ability to do SR without or with less human intervention can be helpful to clinicians in numerous ways. For instance, clinical sites in developing nations, offshore sites, or areas with limited resources, may not have the ability to afford workforce and time for SRs to maintain an up-to-date repository of medical best-practices for a variety of situations. Moreover, automation of SR can save tremendous amount of time and energy of medical practitioners. Doing so would obviate the need of exhaustive manual review of thousands of new scientific articles and allow clinicians to focus only on most relevant literature and spend more time on making clinical and healthcare decisions. In addition, automation of SR can be helpful in figuring out most relevant literature related to ongoing healthcare situations in the time of catastrophic situations such as pandemics and many more.

Research Evaluation: This work formulates automation of SR as classification problem. Approaches of the abstract screening are evaluated based on standard classification metrics, i.e., Precision, Sensitivity, and F-measure. Moreover, this work

introduces new metric, the percentage of *eventually-included*[1] articles that are missed at various high-recall values along with precision.

Contribution: This work presents a series of contributions related to automation of abstract screening of clinical SR:

1. **High-Sensitivity NLP Methods for Abstract Screening**

   This work presents empirical study of NLP-based supervised text classification, classical IR methods and their ensemble for achieving effective automation of abstract screening. This is the first study to propose five different high-sensitivity methods and new evaluation metric for SR systems. The proposed methods are analyzed on datasets developed from six different SRs in the clinical domain. This work concludes that achieving sensitivity beyond 95% is still challenging task. However, 90% sensitivity may seem like a good compromise where healthy precision can be achieved while losing typically on an average 5% or fewer eventually-included articles and this system can be used as the alternate screener in the SR process.

2. **Systematic Review as Question Answering**

   To the best of our knowledge, this thesis work provides a novel approach to view SR as a QA problem for the first time in order to overcome the limitations of the binary classification training strategy; and propose a *more general* abstract screening model for different SRs. This general model gives an advantage for low resource data since it is trained on a combination of different datasets.

[1]Eventually included means that these are the studies that were in the SR after the full-text review.

3. **QA Dataset**

   This thesis work provides a new QA-based dataset for six different SRs which will be made available to the community.

1.4 Structure of Thesis

This work will first review some background and existing work on the topic of Systematic Reviews and Question Answering in Chapter 2. In the next Chapter, the creation, and statistics of all six datasets are discussed in detail. Chapter 4 will provide an empirical study of high-sensitivity methods for datasets developed from six different clinical SRs. Chapter 5 will discuss a novel approach to view SR as a QA problem, propose a more general abstract screening model for different SRs and evaluation of proposed approach w.r.t. baseline. In conclusion, Chapter 6 will summarize the thesis work with proposed methods, limitations, contributions and future research directions for automation of clinical SR process.

Chapter 2

LITERATURE REVIEW

This chapter discusses some background and existing work in the field of systematic reviews and application of QA in various biomedical tasks. In Section 2.1, I will briefly describe history of SRs, and review various text mining, ML, and DL techniques proposed to automate SR over the past several decades. In addition, I will discuss recent neural breakthroughs in NLP and their applicability to SRs. In Section 2.2, I will briefly review the existing work where researcher have reformulated various biomedical tasks as QA approach.

2.1 Systematic Reviews

With increasing plethora of studies, SRs have become more popular in clinical domain to provide concise summary of evidence. However, Gough *et al.* (2017) states the importance of SRs in various domains not limited to biomedical. SRs have become more popular in last few decades, however, Lind (2014) conducted the first SR to record and assess the state of knowledge on scurvy disease in 1753. It was not until the 20^{th} century that more attention was paid to SRs to improve the process of synthesized research evidence. After Greenhalgh (2004) drew attention to the importance of Randomized Control Trial (RCT) in determining treatments, it significantly impacted field of SR and pointed out the need to improve the process of SR (Chalmers *et al.* (2002)). To improve the SR workflow, QUOROM (QUality Of Reporting Of Meta-analysis) statement guidelines are proposed by Moher *et al.* (2000), and these guidelines improved in 2009 as PRISMA by Moher *et al.* (2009). The SR dataset creation (as discussed in next chapter) for our task in this thesis work strictly follow

the PRISMA guidelines.

Nowadays, SRs are mainly done manually by expert clinicians and team. However, the amount of information and published studies continue to increase at tremendous rate which in turn increases the time and cost for conducting SR. Many attempts have been made to reduce SR workload by automating the SR process or part of SR workflow. The ongoing development in learning algorithms such as ML, DL, IR, NLP, CV, etc. has paved a path to automate SR in various ways. As discussed in Bannach-Brown *et al.* (2019), SRs of many clinical problems are implemented using these approaches. Cohen *et al.* (2006, 2012); Liu *et al.* (2018) used learning algorithms for drug class efficacy assessment, Wallace *et al.* (2012) for genetic associations and cost-effectiveness analyses, Miwa *et al.* (2014); Shemilt *et al.* (2014) for public health, Howard *et al.* (2016) for toxicology, Wallace *et al.* (2010); Rathbone *et al.* (2015) for treatment effectiveness, Wallace *et al.* (2010) for nutrition, Howard *et al.* (2016); Liao *et al.* (2018) for preclinical animal studies, and many more studies exist. According to Michelson *et al.* (2019), there are several tools available for automatizing SRs which can be grouped into three categories. The first category is text visualization such as Covidence (Adams *et al.* (2013)), Early Review Organizing Software (Glujovsky *et al.* (2011)), PICO portal (Miller and Forrest (2001)), etc. The second category is term frequency and inverse document frequency weighting methods such as SWIFT-Active Screener (Howard *et al.* (2020)). The third category is semi-automate or automate screening and selections tools (i.e., text classification methods) such as Support Vector Machine (SVM)-based models (Yu *et al.* (2008); Gates *et al.* (2018); Ouzzani *et al.* (2016)). Aim of this thesis work is to build the third category screener for abstract screening of SRs.

With the advent of advance NLP-based techniques, many past attempts have been made to automate or semi-automate abstract screening in SRs (Tsafnat *et al.* (2014,

2018); Reddy *et al.* (2020); Gates *et al.* (2019, 2020); Jaspers *et al.* (2018); Ros *et al.* (2017); Rathbone *et al.* (2015); Kanoulas *et al.* (2019)). NLP-based text classification and data extraction techniques are used extensively in automation of abstract screening (Marshall and Wallace (2019)). A recent review by O'Mara-Eves *et al.* (2015) found 44 algorithms which use NLP to determine the probability that candidate paper should be included/excluded from the SR. NLP and ML-based approaches have become very popular in past decades for automatic clinical SR. Machine Learning models such as SVM, k-Nearest Neighbour (kNN), Latent Dirichlet Allocation (LDA), etc., had been proven very effective in the first decade of 21^{st} century for text classification (Cohen (2006); Cohen *et al.* (2006); Bekhuis and Demner-Fushman (2012); Miwa *et al.* (2014)). For the fist time, Cohen (2006) evaluated the effectiveness of SVM models, class imbalance problem, and high-sensitivity requirement for clinical SRs. After this, Ma (2007) evaluated Active Learning (AL) and Naive Bayes approaches for the first time for SRs. Fiszman *et al.* (2008) proposed use of semantic models for the first time. First developed system for clinical SRs, i.e., GAPscreener was proposed by Yu *et al.* (2008). Nowadays, many ML-based developed tools exist to automate screening process in SRs and reducing manual workload (Marshall *et al.* (2017)). Nevertheless, ML-based approaches rely heavily on arbitrarily set sample features and are unstable and labor-intensive (Qin *et al.* (2021)).

In the past decade, DL-based models have emerged with efficient text classification ability (Qin *et al.* (2021)) in the field of NLP such as recurrent neural networks (Tang *et al.* (2015); Poon *et al.* (2019)), attention mechanisms (Vaswani *et al.* (2017)), transformers (Wolf *et al.* (2019)), Bidirectional Encoder Representations from Transformers (BERT)-based models (Devlin *et al.* (2018)), etc. However, only a few studies exist which analyze the effect of these state-of-the-art NLP models on SRs (Brockmeier *et al.* (2019); Qin *et al.* (2021); Schmidt *et al.* (2020); Begert *et al.* (2020); Wang

and Lo (2021)). In this work, our focus is to use the BERT-based model (SciBERT proposed by Beltagy *et al.* (2019)) and leverage the advance NLP-based QA approach for abstract screening of SR.

2.2 Biomedical Tasks as Question-Answering

Over the past years, there has been a trend of reformulating NLP tasks as Question Answering (QA) tasks. Levy *et al.* (2017) transforms a relation extraction task to a QA task by generating a question for a relation type and if an answer can be extracted from a sentence, then such relations exist in the sentence. Li *et al.* (2019) reformulated Name Entity Recognition (NER) task to QA format in general domain. McCann *et al.* (2018) transformed ten tasks to QA format and built a general model to solve multiple tasks. Inspired by this, many attempts have been made in reformulating biomedical tasks as QA. Wang *et al.* (2020) presented biomedical event extraction as QA task, and proposed QA system based on domain-specific language model SciBERT. Nguyen (2019) proposed QA system which view patient related medical queries as questions to perform self-diagnosis. Similarly, Banerjee *et al.* (2019) made the same attempt but specifically on the biomedical domain in addition to leverage knowledge to guide model learning for NER task. Similar to these previous works, our focus is to reformulate the SR as QA task in this study.

Chapter 3

DATASETS

The purpose of this chapter is to get familiar with the datasets used for this thesis work. This work is developed to automate abstract screening of six different clinical SRs. The datasets used for this study are created manually by the expert physicians of Mayo Clinic. The scope of this study covers six different biomedical SRs: 1) Immune Checkpoint Inhibitors (ICI), 2) Hormone Replacement Therapy (HRT), 3) Cooking, 4) Accelerometer, 5) Acromegaly, and 6) COrona VIrus Disease (COVID). In this chapter, various aspects of dataset creation and statistics are discussed in detail. Here, the dataset creation procedure is explained for ICI data. In particular, this chapter talks about data sources and search strategies for ICI dataset to get overview of how articles are collected, and how manual annotations are done for all six datasets. Other datasets are also created in a similar way. At the end, statistical analysis and inclusion/exclusion criteria are presented for all six different datasets.

3.1 Data sources and Search Strategies

Devising correct search strategy is critical to ensure that review is not biased by easily accessible studies, and all relevant literature is retrieved (Tsafnat *et al.* (2014)). The purpose of this section is to describe detailed procedure of relevant article collection to conduct SR for ICI and increase understanding related to the data sources and search strategies. A comprehensive search of several databases from each database's inception to September 11th, 2018, any language was conducted. The databases included Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, and Daily, Ovid EMBASE, Ovid Cochrane Central Register

of Controlled Trials, Ovid Cochrane Database of Systematic Reviews, and Scopus. The search strategy was designed and conducted by an experienced librarian with input from the study's principle investigator. Controlled vocabulary supplemented with keywords was used to search for phase 2 or 3 clinical trials, systematic reviews, and meta-analyses of ICI drugs. After collecting the articles, the manual annotation (i.e., screening) is done based on title and abstract of each article since the purpose of this study is to automate abstract screening of SRs. Other five datasets are also created in similar way with different data sources and search strategies designed by experts in particular domain. Unfortunately, all datasets are created by Mayo Clinic and are not publicly available yet, hence, I am restricted to provide readers with more details about vocabulary and search strategies used for dataset creation.

3.2 Manual Screening Process

This section describes how these datasets are manually annotated by the experts in relevant domain for the use of this study. The decision for each article whether it's included in SR or not is based on inclusion/exclusion criteria. If article satisfies all the inclusion criteria, then the article is included in the final SR (i.e., Include). For abstract screening of SR, each article was annotated manually by two expert physicians from respective fields whether it's "Include" or "Exclude" by analyzing the title and abstract of an article. When there is a disagreement between two annotators, a positive class (i.e., "Include") is always preferred for the final label for the experiments conducted in this study. The title and abstract of each article are concatenated and used as the input data in our experiments. The articles rated as "Include" after the abstract screening process (i.e., after the 2^{nd} step of the SR process) are considered as the positive samples and the "Exclude" articles as the negative samples.

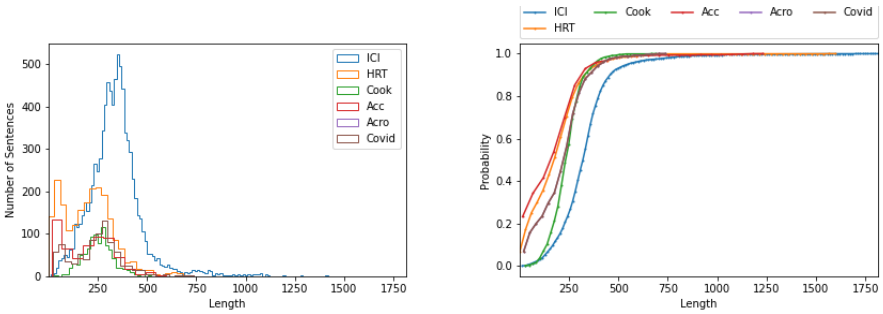
3.3 Statistical Analysis

Statistics about the datasets can be found in Table 3.1. Table 3.1 shows that all datasets are skewed - substantially fewer positive instances than the negative instances except the ICI dataset. Because of this, we consider all datasets as low resource data except ICI for this study. Few articles have the missing title or abstract, and we discarded them (see Table 3.1) if both are missing (NULL values). After abstract screening, included articles go in next stage (i.e., full-text screening). After full-text screening, candidate article will be included in the final SR. Hence, the table also shows the articles included in the SR after full-text reviews (i.e., at the end of the 3^{rd} step of the SR process). Reader can observe from the Table that number of article included after full-text screening reduced by large margin compared to abstract screening stage.

Statistics	**ICI**	**HRT**	**Cooking**	**Accelerometer**	**Acromegaly**	**COVID**
Total Articles	8817	2244	1005	717	1022	4310
Articles included after title and abstract screening	3978	323	136	164	185	683
Articles excluded after title and abstract screening	4839	1921	869	553	837	3627
Articles included after full-text screening	539	99	34	120	111	97
Abstract Missing from Included	2	34	5	0	7	242
Abstract Missing from Excluded	317	88	36	2	26	652
Title Missing	3	0	0	0	0	1

Table 3.1: Statistics of Datasets Developed from Six Different SRs

The concatenated input text of titles and abstracts varies significantly in length. Since it is necessary to fix the maximum input sequence length for SciBERT, I studied the input length distribution and used a data-driven approach for setting the maximum input sequence length. Figures 3.1a and 3.1b show the probability and cumulative distributions of the input sequence length, respectively. Considerable variation can be seen in sequence lengths across the datasets. ICI has longest sequences while



(a) (b)

Figure 3.1: (a) Probability and (b) Cumulative Distributions of Title and Abstract Sequences for the Datasets.

HRT and Accelerometer have the shortest sequences. Typically, 512 words is the 94^{th} percentile of the cumulative distribution for all datasets; however, the percentile for 256 words varied considerably among the datasets.

3.4 Inclusion and Exclusion Criteria

Inclusion and exclusion criteria set the boundaries for the SR. They are determined usually before the search is conducted. These are simply natural language queries. As a practice, SR developers clearly state the criteria at the beginning of the process and modify if necessary along the way. The inclusion and exclusion criteria for each dataset are shown in Appendix A.1. These criteria are important and can be used in automating articles screening. The criteria are used in the bibliographic searches as well as during the manual abstract screening and full article reviews. Moreover, example of one positive and negative sample of ICI dataset is presented in Appendix A.2 for readers understanding.

Chapter 4

EMPIRICAL STUDY OF HIGH-SENSITIVITY METHODS

The techniques needed to achieve high-sensitivity, a key requirement for screening published articles, have not been systematically studied before. In this chapter, I present an empirical study of high-sensitivity techniques using datasets developed from six systematic reviews in the clinical domain. Using SciBERT as a baseline supervised text classifier, five different techniques are studied for achieving high-sensitivity and it is observed that a combination of up-sampling and down-sampling of the training data achieves the best results. A new evaluation metric, the percentage of eventually-included articles that are missed at various high-sensitivity values along with corresponding precision, is proposed in this study. Our results show that 1% or fewer eventually-included articles were lost at $\sim 99\%$ sensitivity; typically 3% and 5% were lost at $\sim 95\%$ and $\sim 90\%$ sensitivity, respectively. However, a lower sensitivity may be used (with better precision) if losing about 5% or so eventually-included articles is acceptable.

4.1 Background

As stated in Chapter 1, a crucial step in SRs is the abstract screening of the article. As known, this step is one of the most time-consuming steps in the production of SR (Bagheri *et al.* (2018)). Abstract screening is often conducted by two individuals independently based on the list of inclusion and exclusion criteria followed by adjudication, thus doubling the effort. Automating this step would reduce the labor required and the time needed for a SR. Nowadays, ML and NLP are used extensively to create many tools to automate the SR process (Marshall *et al.* (2017)). As known,

the US Institute of Medicine explicitly favors high-sensitivity of literature searches and literature screening over high precision (Morton *et al.* (2011)). However, an important requirement of the SR development was not well studied, i.e., the need for high-sensitivity ($\sim 99\%$) for the "inclusion" (positive) class, even at the cost of precision. It can be colloquially stated as: *I don't want to miss any articles even if I have to review more full length papers.* In this Chapter, an empirical study of high-sensitivity techniques is presented using datasets developed from six SRs in the clinical domain. In this work, the important questions are addressed: what is the best way to achieve high-sensitivity with modern neural networks and how low precision would be at the high-sensitivity? Even more fundamentally, since precision reduces with increasing sensitivity (rather dramatically at high-sensitivity), how high the sensitivity needs to be in order to achieve effective automation for abstract screening?

An empirical study, addressing these questions is presented in this research. We (along with Mayo Clinic) developed a new dataset from six systematic reviews in the clinical domain, which will be released to the community (after this work is published). The state-of-the-art text classification BERT-based model (SciBERT) (Sun *et al.* (2019)) pre-trained on the relevant corpus, and the state-of-the-art information retrieval method (BM25) (Robertson and Walker (1994)), and their tandem assemblies, as the baseline methods for screening articles are implemented. Then, five different strategies for achieving high-sensitivity are explored in this work: (1) Lowering the probability threshold for inclusion; (2) Training with imbalanced data, favoring positive instances through up-sampling; (3) Training with imbalanced data, disfavoring negative instances through down-sampling; (4) A combination of up-sampling and down-sampling; (5) Cost-sensitive optimization. This study presents the highest sensitivity achievable with each method and corresponding precision, as well as sensitivity and precision for points prior to the maximum (i.e., plotting the P-R curves

in the high-sensitivity region). This research work also studies the *percentage loss of eventually-included* articles for the high-sensitivity achieving strategies. Eventually-included studies are the articles that were included in an SR *after* the third step of the study selection described in Chapter 3 (i.e., full-text screening).

The results showed that SciBERT achieves the best F-measure (harmonic mean of precision and sensitivity) across the datasets, and the strategy of simultaneously up-sampling and down-sampling achieves the highest sensitivity ($\sim 99\%$), with associated precision of about 20% for four data sets and about 50% for the remaining. The metric we introduced, percentage loss of eventually-included articles, was useful in understanding trade-offs in the sensitivity range from 90% to 99%: e.g., mostly 1% or fewer eventually-included articles were lost at 99% sensitivity; however, a lower sensitivity may be used (with better precision) if losing about 5% or so eventually-included articles is acceptable. This thesis work provides a new and useful data point for NLP in high-sensitivity applications and a new dataset for further research.

4.2 Methods

In this section, different methods used to develop the SR system are explained in detail, and the experiments are discussed for this empirical study.

4.2.1 Abstract-Screening Methods

Here, I discuss the two state-of-the-art text classification and IR methods used for abstract screening, and their combinations, as the baselines. It should be noted that for SRs, it is not necessary to rank the screened abstracts (unlike a search application) and text classification can be used as well as the IR techniques. Supervised Learning (SL) is used with SciBERT for text classification; IR was implemented using BM25, re-ranking methods were implemented by using them in combination (i.e., BM25 +

SciBERT and SciBERT + BM25).

Supervised Text Classification - SL

The SciBERT Model (base-uncased) is used with the linear layer along with the Softmax activation on the top of the model (Beltagy *et al.* (2019)). The schematic representation of the proposed model is presented in Figure 4.1.

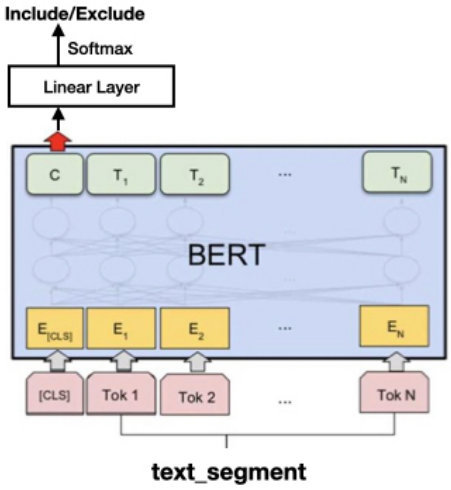


Figure 4.1: Schematic Representation of Supervised Text Classification (SL) Method.

The article text (title + abstract) is used as the input "sentence" for the SciBERT model. As the output, the model produces the probability for each class. To build a classifier, the model is fine-tuned on our data (10-fold cross-validation is used, details are described later). The embeddings are used corresponding to $[CLS]$ token for the classification purpose, as shown in Figure 4.1. At the inference time, the ($[CLS]$ + text segment + $[SEP]$) is given as input to this trained model and used the output $[CLS]$ representation in a linear layer with the Softmax activation function to predict probabilities for each class. As usual, the class with the highest probability is the final predict label.

Information Retrieval - BM25

Recently, BM25 is the most commonly used search and ranking model for the IR systems (Robertson and Zaragoza (2009)). Hence, this research work proposes to use BM25 ranking module to build the SR system. BM25 ranks articles returned by the bibliographic search for an SR based on the query terms corresponding to the SR. The BM25 query is formulated using the inclusion criteria for each dataset. The queries are presented in Appendix B.1.

Hybrid approaches

Two hybrid approach are used for abstract screening, i.e., BM25 + SciBERT and SciBERT + BM25, each of which comprise two modules working in tandem. In the first, the BM25 is the retriever and the SciBERT is next level filter (re-ranker). The retriever is responsible for selecting articles that contains the specific query terms, which are then passed to the SciBERT for second stage filtering, and vice-versa for SciBERT + BM25.

4.2.2 High-sensitivity Techniques

To achieve the high-sensitivity, five different techniques are proposed here. Experiments with the baseline abstract screening showed that text classification with SciBERT consistently achieved better results, and so the high-sensitivity techniques were targeted for SciBERT supervised learning.

Several techniques have been studied in past for training with imbalanced datasets to achieve optimum F-measure (Elkan (2001); Zhou and Liu (2005); Japkowicz *et al.* (2000)). Here, I propose to use these techniques for a different purpose - to destabilize the operating point for achieving high-sensitivity. In addition, this work also makes use of the standard approach of adjusting the probability threshold for the positive

class (commonly used with the feature engineered models such as Logistic Regression and SVM (Yu *et al.* (2015))).

Lowering the probability threshold for inclusion In this method, a custom (lower than normal) threshold is used for predicting the positive label (i.e., "Include") at output of the Softmax activation. In order to find the highest threshold that yields the highest sensitivity value, all the values from 0 to 1 with 10e-5 increment are used to calculate the sensitivity, and select the threshold with maximum sensitivity value.

Up-sampling positive In this method, positive instances are up-sampled k_u times by the replication. I varied k_u, and empirically found k_u value that yields the highest possible sensitivity for each dataset.

Down-sampling negative In a way similar to the above method, here negative instances are down-sampled by factor of k_d while keeping all the positive samples of the dataset (i.e., we randomly selected fewer negative samples compared to the positive samples). The value for k_d yielding the highest possible sensitivity for each dataset is determined empirically.

Hybrid sampling: Up-sampling positives and down-sampling negative Inspired by previous studies (Wang (2014); Padmaja *et al.* (2007); Seiffert *et al.* (2009)) for achieving optimum F-measure, a hybrid approach is proposed, i.e., a combination of up-sampling positives and down-sampling negatives to achieve a high-sensitivity. In this method, positive instances are up-sampled using duplication by a factor of k_1, and negatives instances are down-sampled by a factor of k_2. The combination of k_1 and k_2 is selected empirically in a way that yields the highest possible sensitivity for each dataset.

Cost-sensitive optimization In this method, a different weight is assigned to each class using the equation given below.

$$\text{weight (class)} = k_c * \frac{\text{\# of total samples}}{\text{\# of samples from class}} \tag{4.1}$$

The cost-sensitive factor k_c yielding the highest possible sensitivity is found empirically for each dataset. The intermediate values of precision and sensitivity are recorded for all the methods for plotting P-R curves.

4.2.3 Experiments

Various experiments are conducted to analyze the performance of the proposed systems. All the experiments were performed using Google Colab Pro. Note that the SciBERT standard (base) model allows sequence lengths up to 512. To determine ideal input sequence length, two experiments are performed on the SciBERT using: (1) maximum sequence length of 512 and batch size of 12; and (2) maximum sequence length 384 and a batch size of 16. Both the experiments resulted in the almost similar performance. Since increasing sequence length increases the computational (GPU memory and running time) requirements, maximum sequence length of 384 and a batch size of 16 are used along with AdamW optimizer. I used '0' padding for tokenized text less than 384, and removed tokens for tokenized text with length more than 384.

Next, experiments are conducted for the baseline abstract-screening methods to get optimum F-measure. 10-fold cross-validation is used to make best of the data we have as well as to calculate standard deviation. Positive and negative class data is randomly shuffled and divided into ten groups where each group contains 10% data of both classes (i.e., stratified sampling). Here, the simplest cross-validation method called "holdout method" is used (Raschka (2018)). For text classification - SL, the

model is trained for 10 epochs with learning rate 10e-4. For the IR method - BM25, 10-Fold cross-validation is used to learn the optimum threshold. The threshold is decided for each training fold so that the threshold should include all the positive samples. The test set threshold is the average of optimum thresholds across all training folds. The re-ranking experiments are also conducted using the combinations of BM25 and SciBERT. All the methods for optimum F-measure are trained using the balanced datasets (1:1 ratio of positive and negative samples).

Subsequently, the high-sensitivity experiments are performed with the text classification SL-based method because it achieved the highest optimum F-measure across the datasets. I started with 0 as a threshold for lowering the threshold for inclusion class and increasing with the margin of 10e-5 at the inference time. The sensitivity for the positive class is obtained for all possible values between 0 to 1 with 10e-5 increment and select the threshold with a maximum possible sensitivity for testing. For the up-sampling and down-sampling, I started with the 1:1 ratio of positive and negative samples. After that, I started improving the number of positive instances via replication or removing negative samples via randomly selecting fewer samples until the model gets biased towards positive samples (here, this means yielding sensitivity of 1.0 for every epoch). The value of factor k_u and k_d is selected just before the model gets biased towards positive samples. For the hybrid sampling, I started with the up-sampling factor and down-sampling factor obtained for the high-sensitivity individually; and the similar mechanism as up-sampling and down-sampling is used to estimate the factors k_1 and k_2 for hybrid sampling. In the case of cost-sensitive learning, the value of $k_c = 1$ is the starting point for the eq. 4.1, and increased the k by 0.2 for positive class and select value of k_c before model gets biased towards positive samples.

Lastly, the % of eventually-included articles is calculated for each dataset for

various sensitivity values varied from optimum sensitivity to maximum sensitivity achievable. I wrote code that counted the number of eventually included articles of an SR that were not in the list of articles that the model predicted as "include" at each operating point in the above range. The count was normalized as a percentage.

Metrics

Standard classification metrics, i.e., precision, sensitivity and F-measure are used as performance metrics. The performance is measured only for the positive class, since the goal of SR is to find included articles after the abstract screening process. For further analysis of high-sensitivity methods, Precision-Recall (P-R) curves are generated. To plot the P-R curve, precision and sensitivity values are used at different operating points. A new performance metric is also used in evaluation, the percentage of eventually-included articles that are missed at various high-sensitivity values along with precision for different methods. Eventually-included means that these are the studies that were in the SR after the full-text review.

4.3 Results

This section presents the results for the three experiments: (1) Optimum F-measure (Table 4.1), (2) high-sensitivity methods (Table 4.2), and (3) percentage of eventually included articles at various sensitivity (Table 4.3).

In Table 4.1, we can see that the supervised text classification method achieves better F-measure compared to the rest of the approaches. For ICI, HRT, and Accelerometer datasets, SL achieved F-measures of 0.892, 0.569 and 0.764, respectively. For the COVID, SL achieves similar performance in terms of F-measure as hybrid approaches. Although the F-measures of the hybrid approaches are better than SL for Cooking and Acromegaly, the difference is small. BM25 has better sensitivity for all

Method	Dataset	Precision	Sensitivity	F-measure
SL (SciBERT)	ICI	0.874 (0.014)	0.910 (0.018)	**0.892 (0.013)**
	HRT	**0.427 (0.061)**	0.873 (0.071)	**0.569 (0.047)**
	Cooking	0.453 (0.116)	0.838 (0.102)	0.577 (0.092)
	Accelerometer	0.675 (0.066)	0.895 (0.114)	**0.764 (0.051)**
	Acromegaly	0.383 (0.085)	0.796 (0.097)	0.509 (0.070)
	COVID	**0.482 (0.069)**	0.824 (0.065)	**0.604 (0.051)**
BM25	ICI	0.472 (0.009)	**0.957 (0.009)**	0.633 (0.007)
	HRT	0.144 (0.002)	**0.990 (0.015)**	0.252 (0.003)
	Cooking	0.154 (0.015)	**0.925 (0.102)**	0.264 (0.026)
	Accelerometer	0.229 (0.010)	**0.969 (0.044)**	0.371 (0.016)
	Acromegaly	0.184 (0.010)	**0.957 (0.056)**	0.309 (0.017)
	COVID	0.114 (0.007)	0.814 (0.055)	0.245 (0.013)
BM25 + SciBERT	ICI	**0.875 (0.014)**	0.911 (0.019)	0.892 (0.014)
	HRT	0.423 (0.065)	0.851 (0.058)	0.561 (0.052)
	Cooking	0.478 (0.111)	0.795 (0.112)	0.588 (0.092)
	Accelerometer	**0.678 (0.082)**	0.871 (0.103)	0.757 (0.056)
	Acromegaly	**0.391 (0.083)**	0.775 (0.117)	**0.511 (0.068)**
	COVID	**0.482 (0.069)**	**0.824 (0.064)**	**0.604 (0.051)**
SciBERT + BM25	ICI	**0.875 (0.014)**	0.911 (0.019)	0.892 (0.014)
	HRT	0.423 (0.067)	0.842 (0.052)	0.559 (0.054)
	Cooking	**0.544 (0.163)**	0.707 (0.147)	**0.595 (0.112)**
	Accelerometer	**0.678 (0.082)**	0.786 (0.135)	0.718 (0.064)
	Acromegaly	0.390 (0.086)	0.732 (0.123)	0.499 (0.072)
	COVID	**0.482 (0.069)**	0.824 (0.065)	**0.604 (0.051)**

Table 4.1: Optimum F-measure Results for Supervised Text Classification (SL), Information Retrieval - BM25, and Hybrid Approaches - BM25+SciBERT and SciBERT+BM25. All the Cross-validation Results Are Presented in the Table Has Mean (Standard Deviation) Format. **Bold** Results Indicate the Best Method for Each Dataset.

datasets except for COVID (Table 4.1) but has poor precision across the board. The SL and hybrid approaches have better precision compared to BM25. In summary, the SL method outperforms the IR method in terms of F-measure; the IR method

has superior sensitivity; and hybrid approaches perform similar to SL.

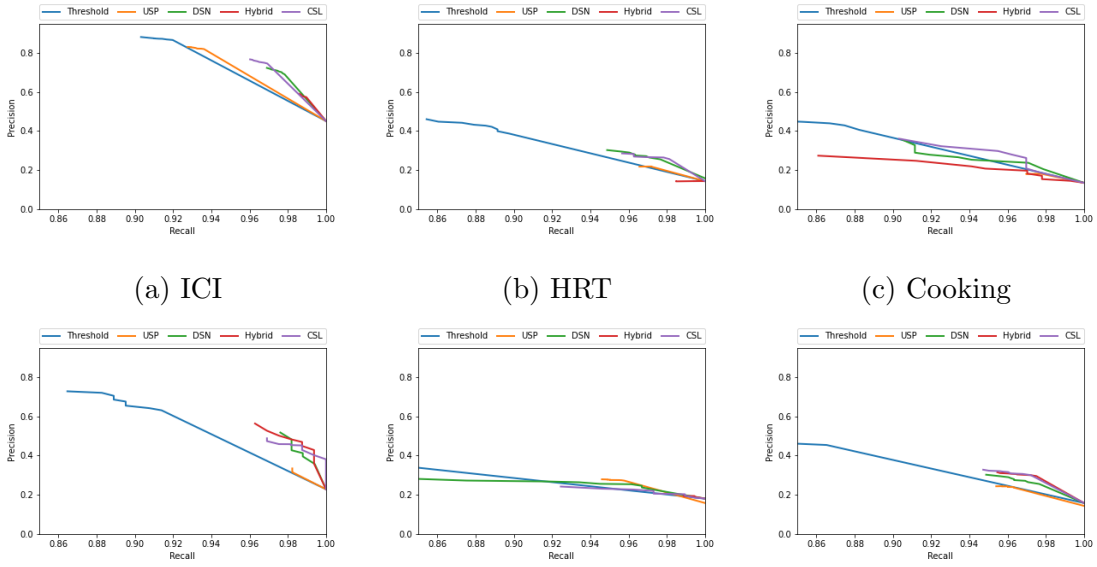
Method	Dataset	Precision	Sensitivity	F-measure
Th	ICI	0.576 (0.057)	0.990 (0.008)	0.726 (0.045)
	HRT	**0.249 (0.051)**	0.969 (0.0004)	**0.393 (0.064)**
	Cooking	0.252 (0.091)	0.926 (0.003)	0.389 (0.110)
	Accelerometer	0.556 (0.187)	0.939 (0.002)	0.680 (0.161)
	Acromegaly	**0.273 (0.065)**	0.946 (0.002)	0.420 (0.075)
	COVID	0.234 (0.065)	0.985 (0.0001)	0.374 (0.080)
USP	ICI	0.827 (0.015)	0.932 (0.010)	0.876 (0.011)
	HRT	0.219 (0.082)	0.972 (0.040)	0.350 (0.106)
	Cooking	0.180 (0.093)	0.970 (0.074)	0.291 (0.110)
	Accelerometer	0.325 (0.159)	0.982 (0.040)	0.468 (0.158)
	Acromegaly	0.276 (0.093)	0.950 (0.061)	**0.417 (0.108)**
	COVID	0.243 (0.099)	0.960 (0.059)	0.377 (0.123)
DSN	ICI	0.711 (0.063)	0.974 (0.008)	**0.820 (0.044)**
	HRT	0.221 (0.082)	**0.975 (0.040)**	0.360 (0.106)
	Cooking	0.279 (0.044)	0.919 (0.092)	0.426 (0.055)
	Accelerometer	0.443 (0.087)	0.982 (0.029)	0.606 (0.080)
	Acromegaly	0.256 (0.020)	0.945 (0.059)	0.402 (0.026)
	COVID	0.280 (0.018)	0.963 (0.035)	0.433 (0.022)
USP and DSN	ICI	0.584 (0.140)	**0.987 (0.013)**	0.724 (0.106)
	HRT	0.144 (0.002)	0.993 (0.019)	0.251 (0.003)
	Cooking	0.197 (0.047)	**0.970 (0.038)**	0.326 (0.067)
	Accelerometer	**0.469 (0.074)**	0.988 (0.026)	**0.633 (0.070)**
	Acromegaly	0.193 (0.026)	0.994 (0.018)	0.323 (0.035)
	COVID	**0.306 (0.057)**	**0.965 (0.031)**	**0.461 (0.071)**
CSL	ICI	**0.757 (0.109)**	0.964 (0.021)	0.843 (0.078)
	HRT	0.273 (0.082)	0.963 (0.029)	0.419 (0.102)
	Cooking	**0.262 (0.093)**	0.970 (0.039)	**0.403 (0.114)**
	Accelerometer	0.454 (0.201)	**0.982 (0.030)**	0.594 (0.195)
	Acromegaly	0.213 (0.027)	**0.973 (0.038)**	0.348 (0.037)
	COVID	0.315 (0.063)	0.960 (0.025)	0.472 (0.077)

Table 4.2: Results for High-sensitivity Strategies - Th: Lowering the Probability Threshold for Inclusion, USP: Up-Sampling Positive, DSN: Down-Sampling Negative, USP and DSN: A Combination of Up-Sampling Positive and Down-Sampling Negative, and CSL: Cost-Sensitive Learning. All the Cross-validation Results Are Presented in the Table Has a Mean (Standard Deviation) Format. Highlighted Results Indicate the Best Method and Bold Results Indicate the Second-best Method for Each Dataset.

All the high-sensitivity experiments in Table 4.2 are conducted using the SL (SciBERT) method. SL (SciBERT) achieves the best performance in terms of F-measure for ICI, HRT, and Accelerometer datasets; and achieves almost similar F-measure for Cooking, Acromegaly and COVID datasets compared to other methods from Table 4.1. Since our goal is to maintain reasonable precision at high-sensitivity, and high F-measure is an indication of both good precision and sensitivity, the high-sensitivity methods are analyzed only on the best performing SL (SciBERT) method. From Table 4.2, it can be observed that, as expected, precision drops (rather dramatically in some cases) at higher sensitivity values. The hybrid technique of up-sampling positive and down-sampling negative yields the highest sensitivity for HRT, Accelerometer, Acromegaly and the second-highest sensitivity for ICI, Cooking and COVID, ranging from 0.965-0.994.

Another observation is that cost-sensitive learning maintains best or second-best precision for ICI, HRT, Cooking and COVID with only 2% or less reduction in the sensitivity. Also, the threshold-based method achieves the highest-sensitivity (~ 0.99) for the ICI and COVID datasets; however, this method fails to achieve similar sensitivity for the remaining datasets. Up-sampling positives and down-sampling negatives individually perform better than the threshold-based method, for all the datasets except for ICI and COVID.

Figure 4.2 shows P-R curves for different high-sensitivity methods and for different datasets. In a P-R curve, ideal performance is when it shows horizontal flat line, which would mean that we do not lose precision while improving sensitivity. However, that is not usually the case in practice as we see here. In the high-sensitivity range (0.98-0.99), hybrid sampling achieves better precision compared to the other methods for three datasets but loses precision considerably for the remaining datasets. The threshold-based method is consistent across all the datasets but mostly has low



(a) ICI (b) HRT (c) Cooking

(d) Accelerometer (e) Acromegaly (f) COVID

Figure 4.2: P-R Curves for Different High-Sensitivity Methods Presented in Table 4.2. Threshold: Lowering the Probability Threshold for Inclusion, USP: Up-Sampling Positive, DSN: Down-Sampling Negative, Hybrid: Combination of Up-Sampling Positive and Down-Sampling Negative, and CSL: Cost-Sensitive Learning.

precision.

In Table 4.3, we can see that the eventually-included articles that are missed in abstract screening are 1% or fewer at the highest sensitivity for all datasets. As expected, the number of missed studies increases as sensitivity decreases. However, cooking and COVID datasets are an exception - only 1% or fewer missing studies for sensitivity at F-optimum.

At $\sim$ 90% sensitivity, we are getting reasonably high precision, but we are losing typically 5% or so eventually-included articles. But at $\sim$ 95% sensitivity, we are loosing typically 3% or fewer. The loss of eventually-included studies is much higher in Accelerometer and Acromegaly compared to the other datasets.

Dataset	Sensitivity	Precision	Missing/Total (%)
ICI	0.911	0.874	16/539 (2.97%)
	0.900	0.895	25/539 (4.63%)
	0.950	0.851	10/539 (1.86%)
	0.990	0.576	02/539 (0.37%)
HRT	0.873	0.427	04/99 (4.04%)
	0.900	0.364	03/99 (3.03%)
	0.950	0.294	02/99 (2.02%)
	0.970	0.249	02/99 (2.02%)
	0.993	0.144	01/99 (1.01%)
Cooking	0.838	0.453	00/34 (0%)
	0.870	0.462	01/34 (2.94%)
	0.900	0.392	01/34 (2.94%)
	0.926	0.252	00/34 (0%)
	0.970	0.180	00/34 (0%)
Accelerometer	0.850	0.675	13/120 (10.83%)
	0.895	0.733	06/120 (5.00%)
	0.900	0.625	07/120 (5.83%)
	0.939	0.556	05/120 (4.16%)
	0.988	0.469	02/120 (1.66%)
Acromegaly	0.796	0.383	22/111 (19.82%)
	0.850	0.351	14/111 (12.61%)
	0.900	0.309	10/111 (9.01%)
	0.946	0.234	6/111 (5.40%)
	0.994	0.193	3/111 (2.70%)
COVID	0.824	0.482	01/97 (1.03%)
	0.900	0.420	00/97 (0%)
	0.950	0.349	00/97 (0%)
	0.985	0.234	00/97 (0%)

Table 4.3: % Of Eventually Included a Study That Is Missed after Abstract-screening at Various High-Sensitivity along with Corresponding Precision. For Each Dataset, the First Row Presents the Sensitivity at Optimum F-measure, and the Last Row Represents the Maximum Achievable Sensitivity for Each Dataset.

4.4 Discussion

One of the reasons behind high-sensitivity but poor precision of BM25 may be its TF-IDF based approach for searching articles. An article should be included if it

matches all the inclusion criteria. However, the BM25 method includes articles that have not met all (but only some of) the criteria; hence more false positives. On the contrary, the SL method seems to learn better patterns from the positive samples. The improvement in precision for the hybrid approaches is because SciBERT reduces the false positives retrieved using BM25.

We hypothesize that the reason for better precision for ICI and COVID datasets at high-sensitivity is that these two datasets have more positive samples compared to the other datasets. A larger sample of positive instances is likely to lead to a more robust performance.

While the combination of up-sampling positive and down-sampling negative benefits from relatively larger number of positive samples, it also increases the false positives and thus reduces the precision at high-sensitivity. This can be seen in the considerable reduction in precision for this method from Figure 4.2.

Based on Table 4.3 results, this study suggests 90% sensitivity may good compromise where healthy precision can be achieved while losing typically on an average 5% or fewer eventually included articles. For example, for the COVID dataset, we are missing 0 eventually-included articles at 90% sensitivity with 40% precision. So, we are reducing the total articles by a factor of 2.5 for full-text analysis. Hence, out of 4310 (from Table 3.1), 1724 article can be filtered without manual effort for the full-text review. This operating point may be particularly interesting when used as a replacement for one of the human abstract-screeners.

This automation reduces the manual effort for SR. Let assume, we have N articles to be review for SR, and time for abstract review is t_A and time for full-text review is t_F ($t_F > t_A$). n_F articles are filtered for the full-text review from N articles, and $N - n_F$ articles are excluded. The time for this abstract review is $T_A = N * t_A$ and full-text review is $T_F = n_F * t_F$. Now, let assume that our system gives you $\hat{n}_F$ articles

for full-text review from N, the time for full-text review is $\hat{T}_F = \hat{n}_F * t_F$. and Time reduction in the manual effort can be calculated using the below equation:

$$T_{reduction(\%)} = \frac{T_A + T_F - \hat{T}_F}{T_A + T_F} * 100\% \tag{4.2}$$

As shown in eq. 4.2, percentage of time reduction is calculated w.r.t. to the manual effort required by one screener. This reduction may not make much of a difference for small datasets; however, this method is advantageous when you have a large datasets like ICI and COVID.

4.5 Chapter Summary

This is an empirical study of high-sensitivity methods using six datasets developed from actual SRs. From the results, we can conclude that the text classification system performs superior to IR using BM25 and re-ranking approaches involving both, in terms of optimum F-measure. We note that achieving high-sensitivity beyond 95% is still a challenging task, and precision reduces rather by a large margin beyond 95% sensitivity. A hybrid strategy of over-sampling positives and under-sampling negatives is a promising approach in achieving high-sensitivity ($\sim$99%). At that sensitivity level, only a few eventually-included articles would be missed but precision may be as low as 20%. A reasonable compromise might be to aim for 90% sensitivity and use the system as the alternate screener in the SR process.

Chapter 5

SYSTEMATIC REVIEW AS QUESTION ANSWERING

As stated in Chapter 1, a high-sensitivity system is a key requirement in clinical SRs. Most existing methods for SRs use binary classification systems trained on labeled data to predict inclusion/exclusion (proposed in Chapter 4). However, this training strategy has several limitations: (1) It ignores the inclusion/exclusion criteria, (2) lacks generalization ability, (3) suffers from low resource data, and (4) fails to achieve reasonable precision at high-sensitivity levels. To overcome these limitations, we reformulate SR as a Question Answering (QA) problem. The proposed QA model is compared with the binary classification model on datasets developed from six different clinical SRs. The experimental results show that our QA model achieves promising results in terms of precision and F-measure at the high-sensitivity levels ($\sim$0.99). More importantly, I train a general QA model and show that this model further improves the performance in low resource data compared to the data-specific model. To the best of author's knowledge, this is the first study to reformulate SR as a QA task and propose a *more general* abstract screening model for different SRs.

5.1 Background

With advances in NLP, many studies have proposed automating abstract screening as stated in Chapter 2 (Marshall *et al.* (2018); O'Mara-Eves *et al.* (2015); Miwa *et al.* (2014); Matwin *et al.* (2010); Gates *et al.* (2019); Del Fiol *et al.* (2018); Bian *et al.* (2019)). However, high-sensitivity requirements for the included articles even at the cost of precision for abstract screening are not well studied as discussed in Chapter 4. In other words, high-sensitivity systems are required in many SRs. How-

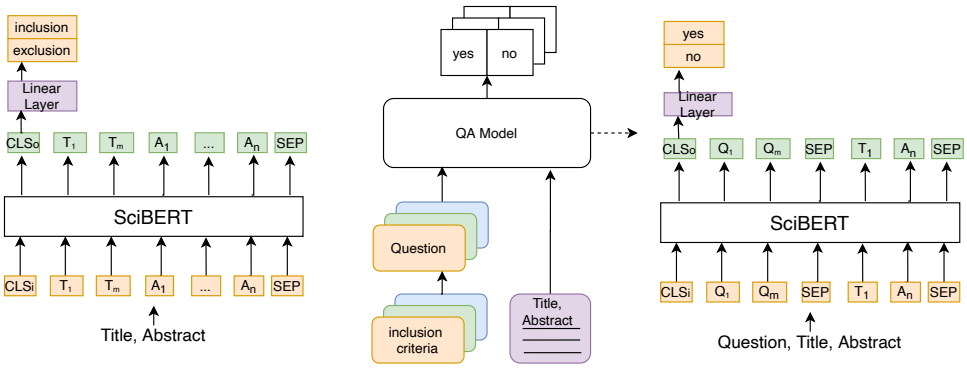
ever, it is challenging to maintain reasonably high precision at high-sensitivity levels. Most existing methods train classification models using labeled data for SR, i.e., a model predicts either *inclusion* or *exclusion* for a given article. However, this training strategy has several limitations: *limitation 1* - It ignores the inclusion/exclusion criteria, *limitation 2* - lack of generalization ability, *limitation 3* - suffers from low resource data since obtaining huge labeled data for SR is costly and time-consuming, and *limitation 4* - fails to achieve reasonable precision at high-sensitivity levels. To overcome these limitations, and inspired by recent works (Levy *et al.* (2017); McCann *et al.* (2018)), we reformulate SR as a QA problem.

The empirical study of the proposed QA approach is presented on datasets developed from six different clinical SRs. The proposed QA system utilizes inclusion criteria in an efficient way (addresses *limitation 1*). In particular, each inclusion criteria are converted to a boolean question, then the QA system takes the question as well as the title and abstract of an article as input and predicts the answer to the question, i.e., "Yes" or "No". If the answer to all inclusion criteria is *Yes*, then the article label is *inclusion*, otherwise *exclusion*. To train the proposed QA model, all six datasets are converted to QA format which gives us the advantage to train a *more general* abstract screening model (simply called general QA model in the rest of the Chapter) on combined data (addresses *limitation 2*). This idea of the general QA model gives an advantage for low resource data since it is trained on a combination of different datasets (addresses *limitation 3*). For six datasets, we also analyzed the effect of the pre-training model on five datasets and fine-tuning with the remaining dataset, i.e., a lifelong and continual learning scenario (Chen and Liu (2018)) (see Section 5.2.2).

The results show that the proposed QA model achieves on an average 7.6% higher precision compared to baseline at high-sensitivity levels ($\sim 99\%$) (addresses *limitation*

4). The general QA model achieves better performance than baseline and almost similar performance to the data-specific QA systems. However, the general model only uses 1/6 of parameters compared to data-specific models. In addition, the fine-tuning on a particular dataset after pre-training with the remaining datasets improves the performance. In summary, the main contributions of this work are a new high-sensitivity approach for screening for SRs, and a new QA based dataset for further research in SRs to be made available to the community.

5.2 Methods



(a) (b)

Figure 5.1: Schematic Representation of (a) Binary Classification Model (BCM), and (b) Proposed QA Approach.

5.2.1 Abstract Screening Systems

Baseline: Binary Classification Model

The SciBERT Model (base-uncased) is used with the linear layer along with the Softmax activation on the top of the model (Beltagy *et al.* (2019)) as shown in Figure 5.1a as Binary Classification Model (BCM). The article text (title + abstract) is

used as the input for the SciBERT model. As the output, the model produces the probability for each class (i.e., "inclusion" or "exclusion"). To build a classifier, the model is fine-tuned on our data. We used the embeddings corresponding to $[CLS_o]$ token for the classification purpose. At the inference time, we input the ($[CLS_i]$ + title and abstract + $[SEP]$) to the trained model and use the output $[CLS_o]$ representation in a linear layer with the Softmax activation function to predict probabilities for each class. As usual, the class with the highest probability is the final predicted label.

Proposed QA System for SRs

We creatively apply a QA-based approach to SRs. Each SR has inclusion criteria, and we convert them into boolean questions, i.e., the answer to the question is either "yes" or "no". In this way, a set of questions is obtained for SRs. Table 5.1 shows the corresponding questions for each inclusion criteria (shown in Table A.1) on six datasets. For a given article, title and abstract text is concatenated with all questions for each dataset and ask the QA model to predict answers for given questions. If all answers are "yes", it means this article satisfies all inclusion criteria thus the label is *inclusion*, otherwise, the label is *exclusion*. Figure 5.1b shows our proposed QA pipeline. Different from the baseline model described in Section 5.2.1, the QA model is given explicitly inclusion criteria as questions, which is the key component of the QA model.

Weak Negative Samples To train a QA model, the training set should consist of both positive samples (i.e., a tuple of a question, input text, a label "yes") and negative samples (i.e., a tuple of a question, input text, a label "no"). The positive samples can be constructed from the labeled data: the articles annotated as inclusion can be a positive sample. However, it is not easy to get negative samples since for an

exclusive article, we don't know precisely which inclusion criteria does it violate. To tackle this issue, a rule-based weak selection approach is applied. Specifically, some keywords are extracted manually for each criterion[1]. For example, the keyword of criteria "*Mentions ICI drug*" is "*ICI*". If an article does not include any keyword, then this article can be used as a weak negative sample.

Dataset	Questions
ICI	i) Is this article mentioned immune checkpoint inhibitors (ICI) drug?
	ii) Does this study report clinical trial data or systematic review or meta analysis?
	iii) Is this article about phase 2 or 3 randomized control trial (RCT)?
	iv) Does this article report trial results for at least one immune checkpoint inhibitors (ICI) and one non-ICI arm?
HRT	i) Does this study report randomized and non-randomized comparative hormone replacement therapy (HRT) studies?
	ii) Is the study related to postmenopausal women of any age?
	iii) Is the follow-up period more than 6 months?
	iv) Does this article compare study with non-HRT (hormone replacement therapy) studies?
Cooking	i) Is this study related to any person, adult or child, healthy or with comorbidity?
	ii) Is this article related to cooking class or culinary intervention delivered by anyone such as chef, dietitian, etc.?
	iii) Is this study compared with any control group?
Accelerometer	i) Is this study related to children ages 1-18 years?
	ii) Does this study use accelerometers in a validation experiment?
	iii) Are there any different accelerometers in the study?
	iv) Does this article report randomized and non-randomized comparative studies?
Acromegaly	i) Does this article include any type of observational and longitudinal studies?
	ii) Does this study report randomized and non-randomized, case-control, case series, and case reports?
	iii) Does this study include patient with any age who reported receiving medical treatment of acromegaly as a first line of treatment?
	iv) Is this article about medical or surgical treatment?
	v) Is this article about patients achieving biochemical control?
COVID	i) Is this study related to adults 18 years or older?
	ii) Does this article report patients with ARDS (acute respiratory distress syndrome)?
	iii) Does this article talk about cell therapy transplantation?
	iv) Is this study only about usual and supportive care only, no treatment?
	v) Is this any study designed that includes case reports?

Table 5.1: Set of Boolean Questions Created Manually from the Inclusion Criteria for Each Dataset

[1]Since criteria are few, it is easy to create rules manually

5.2.2 Experiments

Various experiments are conducted to analyze the performance of the proposed systems. All the experiments were performed using GTX1080 and V100 NVIDIA GPUs. For all the experiments, SciBERT (base-model) is used as a text encoder. Note that the SciBERT standard (base) model allows sequence lengths up to 512. Since increasing sequence length increases the computational (GPU memory and running time) requirements, a maximum sequence length of 384 and a batch size of 8 are used along with 4 gradient accumulation steps and AdamW optimizer. All the models are optimized using Binary Cross Entropy (BCE) loss. All models are trained for 10 epochs with a learning rate of 1e-4. We used '0' padding for tokenized text less than 384, and removed tokens from tokenized text with a length more than 384. Proposed experiments are divided into four categories: (1) data-specific experiments, (2) general QA model, (3) high-sensitivity experiments, and (4) pre-training and fine-tuning. To measure the performance, standard classification metrics, i.e., precision, sensitivity and F-measure are used as performance metrics. Here, the performance is measured only for the positive class since the goal of SR is to find included articles after the abstract screening process.

Data-Specific Experiments

The data-specific experiments are performed for the baseline and QA model where model is trained on each dataset separately to get the optimum F-measure. For these experiments, we use the simplest "set aside test set" method. In order to make a training and testing set, positive (i.e., inclusion) and negative (i.e., exclusion) class data is randomly shuffled, and divide it into 70% of training and 30% of testing. For BCM, six different models are trained on training data of each dataset. For the

proposed QA model, all the training data of each dataset is converted in QA format as suggested in Section 5.2.1. After that, the QA model is trained similarly to the BCM model as described in Section 5.2.1. All the data-specific models are tested on the testing data from the respective dataset.

General QA Model

To analyze the effect of the generalization ability of the proposed QA model on all six SRs, the experiment is conducted to train the general QA model. Since modifying SR datasets in QA format (boolean questions for inclusion criteria) gives us the advantage in terms of training the *more general* model. For the training, training data from all six SR is combined and converted into a QA dataset. After that, the text classification SciBERT model is trained as suggested in Section 5.2.1. In the end, the general QA model is tested on the testing data of each dataset.

High-Sensitivity Experiments

High-sensitivity experiments are conducted for data-specific and general QA model. Here, this work makes use of the standard approach of lowering the probability threshold for the positive class (commonly used with the feature engineered models such as logistic regression and SVM (Yu *et al.* (2015))). In this method, a custom threshold is used for predicting the positive label (i.e., "Inclusion") in the case of the BCM model and positive label (i.e., "Yes") in the case of the QA model at the output of the Softmax activation. In order to find the highest threshold that yields the highest sensitivity value, I tried all the values from 0 to 1 with 1e-4 increment to calculate the sensitivity and select the threshold with maximum sensitivity value.

Pre-training and Fine-tuning

To show that the proposed QA approach can be generalized for new SR (not from these six SRs), this experiment is simulated. This is a scenario of lifelong or continual learning, where the model needs to keep learning when a new dataset (or task) comes (Chen and Liu (2018)). In our case, we assume that five datasets come at first, on which we pre-train a QA model, then a new dataset comes, we further fine-tune the pre-trained QA model on this dataset.

5.3 Results and Discussion

In this section, results are presented for (1) comparison of optimum F-measure and high-sensitivity results with baseline, data-specific QA model and general QA model (Table 5.2); and (2) optimum F-measure for pre-training and fine-tuning (Table 5.3).

Model	High Sensitivity	ICI			HRT			Cooking			Accelerometer			Acromegaly			COVID		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
BCM		**0.868**	0.894	**0.881**	**0.434**	0.858	**0.577**	**0.603**	0.780	**0.681**	**0.690**	0.800	**0.741**	**0.486**	0.625	**0.547**	**0.603**	0.780	**0.681**
QA		0.747	0.939	0.832	0.299	**0.979**	0.458	0.194	**0.976**	0.324	0.385	**0.990**	0.556	0.224	**0.982**	0.365	0.330	**0.922**	0.486
General-QA		0.738	**0.953**	0.832	0.335	0.928	0.503	0.172	0.927	0.463	0.371	**0.990**	0.602	0.242	0.929	0.418	0.374	0.892	0.526
BCM	✓	0.553	**0.995**	0.711	0.158	**0.990**	0.273	0.167	0.976	0.285	0.243	0.980	0.389	0.202	**0.982**	0.335	0.170	**0.995**	0.290
QA	✓	**0.662**	0.983	**0.791**	**0.214**	**0.990**	**0.353**	**0.194**	**0.978**	**0.324**	0.412	**0.983**	0.580	**0.225**	**0.982**	**0.365**	0.242	0.985	0.389
General-QA	✓	0.626	0.981	0.764	**0.214**	**0.990**	**0.353**	0.180	0.976	0.303	**0.516**	0.980	**0.676**	0.188	**0.982**	0.314	**0.279**	0.975	**0.433**
QA (RP)	✓	10.90%	-1.20%	8.00%	5.60%	0%	8.00%	2.70%	0.2%	3.90%	16.90%	0.3%	19.10%	2.30%	0%	3.00%	7.20%	-1.00%	9.90%
General-QA (RP)	✓	7.30%	-1.41%	5.30%	5.60%	0%	8.00%	1.30%	0%	1.80%	27.3%	0%	28.70%	-1.40%	0%	-2.10%	10.90%	-2.00%	14.30%

Table 5.2: The Comparison of Optimum F-measure (First Block) and High-sensitivity (Second Block) Results Between Baseline BCM and Proposed Models (QA and General-QA). The Third Block Presents Relative Performance (RP) of Proposed Models Compared to Baseline at High-sensitivity (Green Highlighted % Indicates Improvement and Red Highlighted % Indicates Degradation). P: Precision, R: Sensitivity, F: F-measure, BCM: Binary Classification Model, QA: Question-Answering Model and General-QA: QA Model Trained on Six Datasets. **Bold** Results Indicate the Best Method for Each Dataset

As shown, the BCM model achieves higher precision at optimum F-measure for all datasets as shown in Table 5.2 (first block). However, the BCM model fails to achieve high-sensitivity at optimum F-measure. Since high-sensitivity is a key requirement in clinical SRs, the results are compared at high-sensitivity levels. Table 5.2 (second block) represents high-sensitivity results for the baseline and proposed QA models. It can be observed that BCM fails to achieve higher precision compared to QA models at high-sensitivity levels ($\sim$0.99). Relative Performance (RP) in % is analyzed for QA and general QA model compared to BCM as shown in Table 5.2 (third block). RP (%) is calculated via simply subtracting performance of one method from other. We observe that the QA model outperforms the baseline model for all six datasets in terms of precision and F-measure by on an average 7.6% and 8.65%, respectively. The general QA model yields better precision and F-measure compared to the baseline model for all six datasets except Acromegaly. In particular, the general QA model outperforms the baseline model in terms of precision and F-measure by on an average 8.5% and 9.33%, respectively. From Table 5.2, we also observe that the data-specific QA model achieved slightly better precision and F-measure than the general QA model for ICI, Cooking and Acromegaly datasets. However, the general QA model performs better or similar compared to QA model for remaining datasets. The advantage of general QA model over QA model is that a it (single model) can be applied to all six datasets, hence, less complex model for a deployment.

Table 5.3 represents the optimum F-measure results for pre-training and fine-tuning. Moreover, Table 5.4 indicates the RP (%) improvement between pre-trained and fine-tuned models. From Table 5.3 and Table 5.4, we observe that fine-tuning helps to improve precision and F-measure for all low resource datasets (i.e., all datasets except the ICI). However, fine-tuning hampers the performance of the model on other datasets used for pre-training. In particular, fine-tuning improves precision

		ICI			HRT			Cooking			Accelerometer			Acromegaly			COVID		
Pre-trained with	*Fine-tune*	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
H+C+ACC+ACR+COV		0.0	0.0	0.0	0.397	0.856	0.542	0.630	0.829	0.716	0.661	0.820	0.732	0.378	0.857	0.525	0.467	0.820	0.595
H+C+ACC+ACR+COV	I	0.743	0.958	0.837	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I+C+ACC+ACR+COV		0.748	0.951	0.837	0.0	0.0	0.0	0.175	0.927	0.295	0.368	0.980	0.536	0.218	0.964	0.355	0.333	0.946	0.493
I+C+ACC+ACR+COV	H	0.728	0.948	0.823	0.365	0.918	0.522	0.248*	0.902	0.389*	0.416*	0.940	0.577*	0.299*	0.679	0.415*	0.339*	0.849	0.485
I+H+ACC+ACR+COV		0.756	0.941	0.838	0.291	0.969	0.448	0.0	0.0	0.0	0.446	0.999	0.617	0.202	0.982	0.335	0.367	0.898	0.521
I+H+ACC+ACR+COV	C	0.795*	0.879	0.835	0.401*	0.876	0.550*	0.275	0.951	0.426	0.333	0.460	0.387	0.231*	0.589	0.332	0.418*	0.751	0.538
I+H+C+ACR+COV		0.687	0.976	0.806	0.288	0.979	0.445	0.305	0.951	0.462	0.0	0.0	0.0	0.225	0.964	0.365	0.350	0.902	0.504
I+H+C+ACR+COV	ACC	0.768*	0.889	0.824*	0.426*	0.835	0.564*	0.371*	0.951	0.534*	0.467	0.980	0.632	0.331*	0.804	0.469*	0.397*	0.780	0.526*
I+H+C+ACC+COV		0.748	0.958	0.840	0.394	0.918	0.551	0.343	0.878	0.493	0.462	0.960	0.623	0.0	0.0	0.0	0.396	0.834	0.537
I+H+C+ACC+COV	ACR	0.795*	0.941	0.862*	0.379	0.804	0.515	0.818*	0.439	0.571*	0.714*	0.200	0.312	0.242	0.964	0.387	0.697*	0.112	0.193
I+H+C+ACC+ACR		0.817	0.904	0.858	0.378	0.928	0.537	0.342	0.927	0.500	0.412	0.980	0.580	0.296	0.893	0.444	0.0	0.0	0.0
I+H+C+ACC+ACR	COV	0.750	0.892	0.815	0.476*	0.608	0.534	0.778*	0.512	0.618*	0.999*	0.060	0.113	0.0	0.0	0.0	0.353	0.922	0.511

Table 5.3: Optimum F-measure Results for Pre-training and Fine-tuning Experiments. I: ICT, H: HRT, C: Cooking, ACC: Accelerometer, ACR: Acromegaly, COV: COVID. * Means Fine-tuning a Model on a New Dataset Improves the Performance on Other Datasets

and F-measure by on an average 3.93% and 4.9%, respectively. One interesting finding from Table 5.3 is that pre-trained models can not be used directly (without fine-tuning) for particular datasets since they are performing poorly (∼0.0). We also observe that fine-tuning with large datasets like ICI degrades the performance of other datasets dramatically (∼0.0).

5.3.1 P-R Curves

To analyze the performance of the proposed systems at a different level of sensitivity (especially high-sensitivity range), we plot Precision-Recall (P-R) curves for all datasets as shown in Figure 5.2. In a P-R curve, ideally, we want a horizontal flat line, which would mean that we do not lose precision while improving sensitivity. However, that is not usually the case in practice as we see here. At high-sensitivity (∼0.98-0.99), our proposed QA models achieve better precision compared to base-

Dataset	Pretrained on other five datasets		
	P	R	F
ICI	-0.4%	+1.9%	+0.5%
HRT	+6.6%	-6.1%	+6.4%
Cooking	+8.1%	-2.5%	+10.2%
Accelerometer	+8.2%	-0.1%	+7.6%
Acromegaly	+1.8%	-1.8%	+2.2%
COVID	+2.3%	0.0%	+2.5%

Table 5.4: Relative Performance (in %) of Optimum F-measure Results Between Fine-tuning Model Pre-trained on Other Five Datasets and Model Trained Only Using Particular Dataset (Data-Specific QA Model)

line for ICI, HRT, Accelerometer, Acromegaly, and COVID; and show almost similar performance for the Cooking dataset.

5.3.2 Analysis

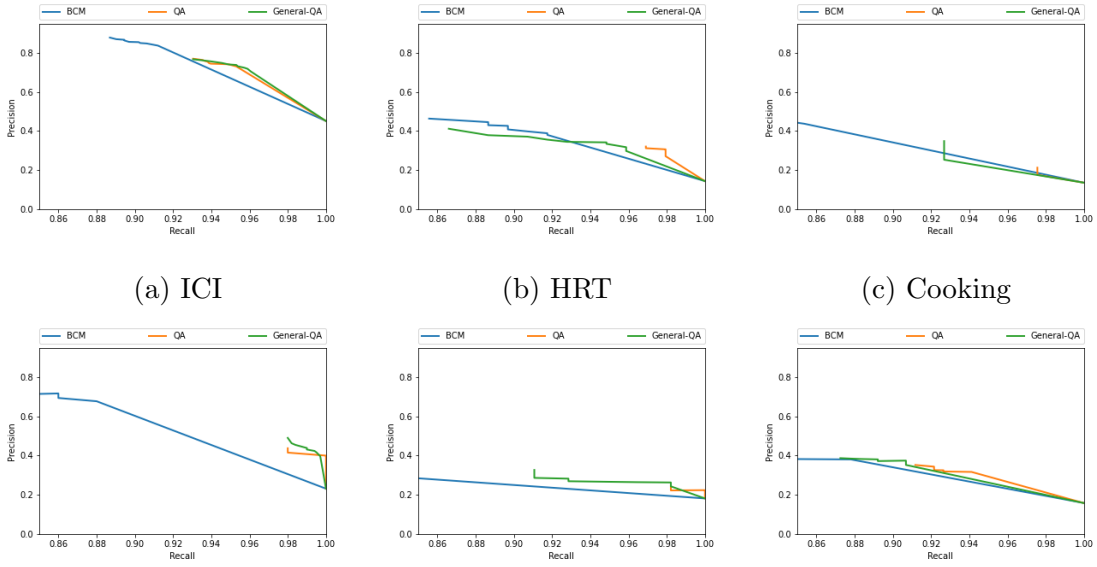
Here, this section discusses the strengths of the proposed methods in terms of explainability and robustness in detail.

Explainability

Inclusion criteria satisfied and unsatisfied by any article can be identified easily using our QA model since it is giving us predictions corresponding to each inclusion criteria. To achieve this kind of similar explainability using BCM is very challenging.

Robustness

Here, two different questions are answered: (1) *can the model understand two semantic equivalent questions at inference time?*; and (2) *is the model sensitive to*



(a) ICI (b) HRT (c) Cooking

(d) Accelerometer (e) Acromegaly (f) COVID

Figure 5.2: P-R Curves for Different Methods Presented in Table 5.2. BCM: Binary Classification Model, QA: Question Answering Model and General-QA: General QA Model Trained on Combined Data

Model	Training	Testing	ICI			HRT			Cooking			Accelerometer			Acromegaly			COVID		
			P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
QA	Q1	Q1	0.747	0.939	0.832	0.299	0.979	0.458	0.194	0.976	0.324	0.385	0.999	0.556	0.224	0.982	0.365	0.330	0.922	0.486
QA	Q1	Q2	0.763	0.940	0.842	0.281	0.969	0.436	0.196	0.976	0.327	0.394	0.999	0.565	0.225	0.999	0.367	0.336	0.922	0.493
QA	Q1	Q1	0.747	0.939	0.832	0.299	0.979	0.458	0.194	0.976	0.324	0.385	0.990	0.556	0.224	0.982	0.365	0.330	0.922	0.486
QA	Q2	Q1	0.766	0.947	0.847	0.306	0.948	0.462	0.198	0.976	0.329	0.427	0.999	0.599	0.229	0.964	0.370	0.356	0.946	0.517

Table 5.5: Optimum F-measure Results for the Data-specific QA Model to Analyze the Robustness of the Model Towards Different Question Formats. Q1 Indicates Set of Questions given in Table 5.1, and Q2 Indicates Set of New Semantically Equivalent but Syntactically Different Questions Prepared from Q1

questions?. To answer the first question, data-specific QA model is trained on the QA dataset prepared using questions mentioned in Table 5.1 (Q1). At the inference time, semantically identical questions are provided but with different syntactic

formulation (Q2), and yet the QA model can maintain its performance as shown in Table 5.5 (first block). Hence, we can say that the QA model can understand two semantic equivalent questions at inference time. To answer the second question, two different data-specific QA models are trained, one on the QA dataset generated using questions mentioned in Table 5.1 (Q1) and the other on the QA dataset generated using semantically similar questions (Q2). However, both models train on different questions show almost similar performance at inference time as shown in Table 5.5 (second block). Hence, we can say that the model is not sensitive to questions at training time.

Pre-training and Fine-tuning

Here, two different questions are answered: (1) *can a pre-train model generalize well on unseen data?*; and (2) *does a pre-training model help?*. From Table 5.3, we observe that the only pre-trained model is performing poorly ($\sim$0.0) if the dataset is not included at the training time (i.e., unseen data). Hence, a pre-train model can not be generalized well on unseen data. However, fine-tuning on a particular dataset after pre-training on other remaining datasets has improved precision and F-measure compared to QA model trained only on the particular dataset (RP(%) is shown in Table 5.4 after fine-tuning).

5.4 Chapter Summary

In summary, this research study presented clinical SRs as a QA task for the first time to overcome several limitations of BCM. This QA approach is promising in various aspects such as utilizing inclusion criteria in effective ways, providing generalization ability, achieving reasonable precision at high-sensitivity levels, and an efficient way to utilize low resource datasets. More importantly, this work studied

the high-sensitivity systems for clinical SRs and achieved promising results in terms of precision. Besides, this study provides a new QA-based dataset for further research in clinical SRs.

Chapter 6

CONCLUSIONS

This chapter presents a summary of the thesis work, limitations of this current work, and future research scopes in the field of SRs.

6.1 Summary

The aim of this work is to develop low manual efforts NLP-based classifiers to automate abstract screening of SR. The major contribution of this work is that it presents an empirical study of different techniques to achieve high sensitivity for six different SRs. To the best of the author's knowledge, this work presents SR as a QA system for the first time to overcome the limitations of binary classification training. The proposed QA system outperforms existing binary classification models in terms of precision, sensitivity, and F-measure. The QA formulation of SR gives an opportunity to develop *more general* SR system so that a single system can work on various datasets. This approach can be helpful for low resource datasets. This work analyzes the effectiveness of this approach on five low resource SR datasets. All the proposed approaches are evaluated on datasets developed from six different clinical SRs. In addition, this work provides a novel QA dataset to the community for continuing research in the domain of clinical SR.

6.2 Limitations and Future Research

This research work has various limitations, and they provide future research directions. This study provides a scope of further research in the domain of clinical SRs.

- This thesis work is conducted only on six datasets. Extending proposed methodologies to a larger number of datasets would improve our understanding. This work opens a future research direction in terms of new datasets creation and the use of proposed methodologies on these clinical SR datasets.

- This study is limited to title and abstract screening only, and so applying NLP-based classifiers to full-text reviews would enable end-to-end automation for SR development. Automated full-text reviews may eliminate the need for abstract screening or act as a countermeasure to the low precision and loss of eventually-included articles in the automated abstract-screening step. A detailed study of this aspect would be a very promising future research score in clinical SRs.

- We have not leveraged the exclusion criteria in BM25 queries and creating questions for the proposed QA approach. We have not explored text classification methods that directly leverage both inclusion and exclusion criteria in addition to the context language modelling of SciBERT. Both are suitable topics for future study.

- The data creation for the clinical SRs is very time-consuming and costly. Moreover, getting new datasets and waiting for the model to be trained on this new data limits scalability and adoption in real world systems. The proposed framework in this thesis can be restructured to the "instruction paradigm" (Le Scao and Rush (2021); Mishra *et al.* (2021)) where instructions or prompts written in natural language can replace a lot of data samples.

- Several recent works have shown that spurious dataset biases provide unintended shortcuts for models to solve the task without truly understanding its underlying features, this overestimates performance of model and prevent it

from generalizing. Removing these biases have shown to significantly improve model performance (Le Bras *et al.* (2020); Mishra *et al.* (2020)). Similarly for the SRs, filtering can be performed in the pre-screening phase to remove redundant or potentially biased documents.

- Unlike general ML paradigm, incorrect answering in health care domain have serious consequences. Hence, model should have an option to skip answering whenever model is not confident and seek human intervention. We can incorporate selective answering methodology (Kamath *et al.* (2020); Varshney *et al.* (2020)) in our proposed framework to build a more reliable system for real world application.

REFERENCES

Adams, C. E., S. Polzmacher and A. Wolff, "Systematic reviews: work that needs to be done and not to be done", Journal of Evidence-Based Medicine **6**, 4, 232–235 (2013).

Allen, I. E. and I. Olkin, "Estimating time to conduct a meta-analysis from number of citations retrieved", Jama **282**, 7, 634–635 (1999).

Bagheri, E., P. Rios, A. Pourmasoumi, R. C. Robson, J. Hwee, W. Isaranuwatchai, N. Darvesh, M. J. Page, A. C. Tricco *et al.*, "Improving the conduct of systematic reviews: a process mining perspective", Journal of clinical epidemiology **103**, 101–111 (2018).

Banerjee, P., K. K. Pal, M. Devarakonda and C. Baral, "Knowledge guided named entity recognition for biomedical text", arXiv: Computation and Language (2019).

Bannach-Brown, A., P. Przybyła, J. Thomas, A. S. Rice, S. Ananiadou, J. Liao and M. R. Macleod, "Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error", Systematic reviews **8**, 1, 1–12 (2019).

Begert, D., J. Granek, B. Irwin and C. Brogly, "Towards automating systematic reviews on immunization using an advanced natural language processing–based extraction system", Canada Communicable Disease Report **46**, 6, 174–179 (2020).

Bekhuis, T. and D. Demner-Fushman, "Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers", Artificial intelligence in medicine **55**, 3, 197–207 (2012).

Beltagy, I., K. Lo and A. Cohan, "Scibert: A pretrained language model for scientific text", arXiv preprint arXiv:1903.10676 (2019).

Bian, J., S. Abdelrahman, J. Shi and G. Del Fiol, "Automatic identification of recent high impact clinical articles in pubmed to support clinical decision making using time-agnostic features", Journal of biomedical informatics **89**, 1–10 (2019).

Brockmeier, A. J., M. Ju, P. Przybyła and S. Ananiadou, "Improving reference prioritisation with pico recognition", BMC medical informatics and decision making **19**, 1, 1–14 (2019).

Chalmers, I., L. V. Hedges and H. Cooper, "A brief history of research synthesis", Evaluation & the health professions **25**, 1, 12–37 (2002).

Chen, Z. and B. Liu, "Lifelong machine learning", Synthesis Lectures on Artificial Intelligence and Machine Learning **12**, 3, 1–207 (2018).

Cohen, A. M., "An effective general purpose approach for automated biomedical document classification", in "AMIA annual symposium proceedings", vol. 2006, p. 161 (American Medical Informatics Association, 2006).

Cohen, A. M., K. Ambert and M. McDonagh, "Studying the potential impact of automated document classification on scheduling a systematic review update", BMC medical informatics and decision making **12**, 1, 1–11 (2012).

Cohen, A. M., W. R. Hersh, K. Peterson and P.-Y. Yen, "Reducing workload in systematic review preparation using automated citation classification", Journal of the American Medical Informatics Association **13**, 2, 206–219 (2006).

Del Fiol, G., M. Michelson, A. Iorio, C. Cotoi and R. B. Haynes, "A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study", Journal of medical Internet research **20**, 6, e10281 (2018).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805 (2018).

Elkan, C., "The foundations of cost-sensitive learning", in "International joint conference on artificial intelligence", vol. 17, pp. 973–978 (Lawrence Erlbaum Associates Ltd, 2001).

Fiszman, M., E. Ortiz, B. E. Bray and T. C. Rindflesch, "Semantic processing to support clinical guideline development", in "AMIA Annual Symposium Proceedings", vol. 2008, p. 187 (American Medical Informatics Association, 2008).

Gates, A., M. Gates, D. DaRosa, S. A. Elliott, J. Pillay, S. Rahman, B. Vandermeer and L. Hartling, "Decoding semi-automated title-abstract screening: findings from a convenience sample of reviews", Systematic reviews **9**, 1, 1–12 (2020).

Gates, A., S. Guitard, J. Pillay, S. A. Elliott, M. P. Dyson, A. S. Newton and L. Hartling, "Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools", Systematic reviews **8**, 1, 1–11 (2019).

Gates, A., C. Johnson and L. Hartling, "Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the abstrackr machine learning tool", Systematic reviews **7**, 1, 45 (2018).

Glujovsky, D., A. Bardach, S. G. Martí, D. Comandé and A. Ciapponi, "Prm2 eros: a new software for early stage of systematic reviews", Value in Health **14**, 7, A564 (2011).

Gough, D., S. Oliver and J. Thomas, *An introduction to systematic reviews* (Sage, 2017).

Greenhalgh, T., "Effectiveness and efficiency: Random reflections on health services", Bmj **328**, 7438, 529 (2004).

Howard, B. E., J. Phillips, K. Miller, A. Tandon, D. Mav, M. R. Shah, S. Holmgren, K. E. Pelch, V. Walker, A. A. Rooney *et al.*, "Swift-review: a text-mining workbench for systematic review", Systematic reviews **5**, 1, 1–16 (2016).

Howard, B. E., J. Phillips, A. Tandon, A. Maharana, R. Elmore, D. Mav, A. Sedykh, K. Thayer, B. A. Merrick, V. Walker *et al.*, "Swift-active screener: accelerated document screening through active learning and integrated recall estimation", Environment International **138**, 105623 (2020).

Japkowicz, N. *et al.*, "Learning from imbalanced data sets: a comparison of various strategies", in "AAAI workshop on learning from imbalanced data sets", vol. 68, pp. 10–15 (AAAI Press Menlo Park, CA, 2000).

Jaspers, S., E. De Troyer and M. Aerts, "Machine learning techniques for the automation of literature reviews and systematic reviews in efsa", EFSA Supporting Publications **15**, 6, 1427E (2018).

Kamath, A., R. Jia and P. Liang, "Selective question answering under domain shift", arXiv preprint arXiv:2006.09462 (2020).

Kanoulas, E., D. Li, L. Azzopardi and R. Spijker, "Clef 2019 technology assisted reviews in empirical medicine overview", in "CEUR workshop proceedings", vol. 2380 (2019).

Le Bras, R., S. Swayamdipta, C. Bhagavatula, R. Zellers, M. Peters, A. Sabharwal and Y. Choi, "Adversarial filters of dataset biases", in "International Conference on Machine Learning", pp. 1078–1088 (PMLR, 2020).

Le Scao, T. and A. M. Rush, "How many data points is a prompt worth?", in "Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", pp. 2627–2636 (2021).

Levy, O., M. Seo, E. Choi and L. Zettlemoyer, "Zero-shot relation extraction via reading comprehension", arXiv preprint arXiv:1706.04115 (2017).

Li, X., J. Feng, Y. Meng, Q. Han, F. Wu and J. Li, "A unified mrc framework for named entity recognition", arXiv preprint arXiv:1910.11476 (2019).

Liao, J., S. Ananiadou, L. Currie, B. E. Howard, A. Rice, S. Sena, J. Thomas, A. Varghese and M. R. Macleod, "Automation of citation screening in pre-clinical systematic reviews", bioRxiv p. 280131 (2018).

Liberati, A., D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen and D. Moher, "The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration", Journal of clinical epidemiology **62**, 10, e1–e34 (2009).

Lind, J., *A treatise of the scurvy, in three parts* (Cambridge University Press, 2014).

Liu, J., P. Timsina and O. El-Gayar, "A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews", Information Systems Frontiers **20**, 2, 195–207 (2018).

Ma, Y., *Text classification on imbalanced data: Application to Systematic Reviews Automation*, Ph.D. thesis, University of Ottawa (Canada) (2007).

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

Marshall, I. J., A. Noel-Storr, J. Kuiper, J. Thomas and B. C. Wallace, "Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide", Research synthesis methods **9**, 4, 602–614 (2018).

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

Matwin, S., A. Kouznetsov, D. Inkpen, O. Frunza and P. O'Blenis, "A new algorithm for reducing the workload of experts in performing systematic reviews", Journal of the American Medical Informatics Association **17**, 4, 446–453 (2010).

McCann, B., N. Keskar, C. Xiong and R. Socher, "The natural language decathlon: Multitask learning as question answering", ArXiv **abs/1806.08730** (2018).

Michelson, M., M. Ross and S. Minton, "Ai2 leveraging machine-assistance to replicate a systematic review", Value in Health **22**, S34 (2019).

Miller, S. A. and J. L. Forrest, "Enhancing your practice through evidence-based decision making: Pico, learning how to ask good questions", Journal of Evidence Based Dental Practice **1**, 2, 136–141 (2001).

Mishra, S., A. Arunkumar, B. Sachdeva, C. Bryan and C. Baral, "Dqi: Measuring data quality in nlp", arXiv preprint arXiv:2005.00816 (2020).

Mishra, S., D. Khashabi, C. Baral and H. Hajishirzi, "Natural instructions: Benchmarking generalization to new tasks from natural language instructions", arXiv preprint arXiv:2104.08773 (2021).

Miwa, M., J. Thomas, A. O'Mara-Eves and S. Ananiadou, "Reducing systematic review workload through certainty-based screening", Journal of biomedical informatics **51**, 242–253 (2014).

Moher, D., D. J. Cook, S. Eastwood, I. Olkin, D. Rennie and D. F. Stroup, "Improving the quality of reports of meta-analyses of randomised controlled trials: the quorom statement", Oncology Research and Treatment **23**, 6, 597–602 (2000).

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman and P. Group, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement", PLoS medicine **6**, 7, e1000097 (2009).

Morton, S., A. Berg, L. Levit, J. Eden *et al.*, *Finding what works in health care: standards for systematic reviews* (National Academies Press, 2011).

Nguyen, V., "Question answering in the biomedical domain", in "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop", pp. 54–63 (2019).

Ouzzani, M., H. Hammady, Z. Fedorowicz and A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews", Systematic reviews **5**, 1, 1–10 (2016).

O'Mara-Eves, A., J. Thomas, J. McNaught, M. Miwa and S. Ananiadou, "Using text mining for study identification in systematic reviews: a systematic review of current approaches", Systematic reviews **4**, 1, 5 (2015).

Padmaja, T. M., N. Dhulipalla, P. R. Krishna, R. S. Bapi and A. Laha, "An unbalanced data classification model using hybrid sampling technique for fraud detection", in "International Conference on Pattern Recognition and Machine Intelligence", pp. 341–348 (Springer, 2007).

Poon, H.-K., W.-S. Yap, Y.-K. Tee, W.-K. Lee and B.-M. Goi, "Hierarchical gated recurrent neural network with adversarial and virtual adversarial training on text classification", Neural Networks **119**, 299–312 (2019).

Qin, X., J. Liu, Y. Wang, Y. Liu, K. Deng, Y. Ma, K. Zou, L. Li and X. Sun, "Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews", Journal of Clinical Epidemiology **133**, 121–129 (2021).

Raschka, S., "Model evaluation, model selection, and algorithm selection in machine learning", arXiv preprint arXiv:1811.12808 (2018).

Rathbone, J., T. Hoffmann and P. Glasziou, "Faster title and abstract screening? evaluating abstrackr, a semi-automated online screening program for systematic reviewers", Systematic reviews **4**, 1, 1–7 (2015).

Reddy, S. M., S. Patel, M. Weyrich, J. Fenton and M. Viswanathan, "Comparison of a traditional systematic review approach with review-of-reviews and semi-automation as strategies to update the evidence", Systematic reviews **9**, 1, 1–13 (2020).

Robertson, S. and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond* (Now Publishers Inc, 2009).

Robertson, S. E. and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval", in "SIGIR'94", pp. 232–241 (Springer, 1994).

Ros, R., E. Bjarnason and P. Runeson, "A machine learning approach for semi-automated search and selection in literature studies", in "Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering", pp. 118–127 (2017).

Schmidt, L., J. Weeds and J. Higgins, "Data mining in clinical trial text: Transformers for classification and question answering tasks", arXiv preprint arXiv:2001.11268 (2020).

Seiffert, C., T. M. Khoshgoftaar and J. Van Hulse, "Hybrid sampling for imbalanced data", Integrated Computer-Aided Engineering **16**, 3, 193–210 (2009).

Shemilt, I., A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. O'Mara-Eves, M. P. Kelly and J. Thomas, "Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews", Research Synthesis Methods **5**, 1, 31–49 (2014).

Sun, C., X. Qiu, Y. Xu and X. Huang, "How to fine-tune bert for text classification?", in "China National Conference on Chinese Computational Linguistics", pp. 194–206 (Springer, 2019).

Tang, D., B. Qin, X. Feng and T. Liu, "Effective lstms for target-dependent sentiment classification", arXiv preprint arXiv:1512.01100 (2015).

Tawfik, G. M., K. A. S. Dila, M. Y. F. Mohamed, D. N. H. Tam, N. D. Kien, A. M. Ahmed and N. T. Huy, "A step by step guide for conducting a systematic review and meta-analysis with simulation data", Tropical medicine and health **47**, 1, 1–9 (2019).

Thakur, A., A. P. Mishra, B. Panda, D. Rodríguez, I. Gaurav and B. Majhi, "Application of artificial intelligence in pharmaceutical and biomedical studies", Current pharmaceutical design **26**, 29, 3569–3578 (2020).

Tsafnat, G., A. Dunn, P. Glasziou and E. Coiera, "The automation of systematic reviews", (2013).

Tsafnat, G., P. Glasziou, M. K. Choong, A. Dunn, F. Galgani and E. Coiera, "Systematic review automation technologies", Systematic reviews **3**, 1, 1–15 (2014).

Tsafnat, G., P. Glasziou, G. Karystianis and E. Coiera, "Automated screening of research studies for systematic reviews using study characteristics", Systematic reviews **7**, 1, 1–9 (2018).

Varshney, N., S. Mishra and C. Baral, "It's better to say" i can't answer" than answering incorrectly: Towards safety critical nlp systems", arXiv preprint arXiv:2008.09371 (2020).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", arXiv preprint arXiv:1706.03762 (2017).

Wallace, B. C., K. Small, C. E. Brodley, J. Lau, C. H. Schmid, L. Bertram, C. M. Lill, J. T. Cohen and T. A. Trikalinos, "Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining", Genetics in medicine **14**, 7, 663–669 (2012).

Wallace, B. C., T. A. Trikalinos, J. Lau, C. Brodley and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews", BMC bioinformatics **11**, 1, 1–11 (2010).

Wang, L. L. and K. Lo, "Text mining approaches for dealing with the rapidly expanding literature on covid-19", Briefings in Bioinformatics **22**, 2, 781–799 (2021).

Wang, Q., "A hybrid sampling svm approach to imbalanced data classification", in "Abstract and Applied Analysis", vol. 2014 (Hindawi, 2014).

Wang, X. D., L. Weber and U. Leser, "Biomedical event extraction as multi-turn question answering", in "Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis", pp. 88–96 (2020).

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing", arXiv preprint arXiv:1910.03771 (2019).

Yu, H., C. Mu, C. Sun, W. Yang, X. Yang and X. Zuo, "Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data", Knowledge-Based Systems **76**, 67–78 (2015).

Yu, W., M. Clyne, S. M. Dolan, A. Yesupriya, A. Wulf, T. Liu, M. J. Khoury and M. Gwinn, "Gapscreener: an automatic tool for screening human genetic association literature in pubmed using the support vector machine technique", BMC bioinformatics **9**, 1, 1–9 (2008).

Zhou, Z.-H. and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem", IEEE Transactions on knowledge and data engineering **18**, 1, 63–77 (2005).

APPENDIX A

DATASETS

A.1 INCLUSION/EXCLUSION CRITERIA

Dataset	Inclusion Criteria	Exclusion Criteria
ICI	Mentions ICI drug, Reports clinical trial data or is SRMA, Phase 2 or 3 RCT having at least one ICI arm and one non-ICI arm that reports trial results	Focus is on drugs other than ICI, phase 1 trials; non-randomized/single-arm trials; trial protocols with no results reported; trials that include ICI drugs in both arms; non-English language articles; conference abstracts; or are article types other than RCT and SRMA
HRT	Randomized and non-randomized comparative HRT studies, Study-related to postmenopausal women of any age, The follow-up period of more than 6 months, Comparison with non-HRT studies	Systematic review; case series or uncontrolled studies; different population; different intervention; follow-up period less than 6 months; mortality is not one if the outcomes
Cooking	Study-related to any person, adult or child, healthy or with comorbidity, Cooking class/culinary intervention delivered by anyone (chef, dietitian, etc.), Comparison with any control group	-
Accelerometer	Study-related to children ages 1-18 years, Use of accelerometers in a validation experiment, Different accelerometers in the study, Randomized and non-randomized comparative studies	Studies with specific populations in which physical activities are limited (ex, asthma, cerebral palsy) Reviews, conference abstracts and any non-original research.
Acromegaly	Any type of observational and longitudinal studies; randomized and non-randomized; case-control; case series; and case reports, any age who reported receiving medical treatment of acromegaly as a first line of treatment, medical or surgical treatment, patients achieving biochemical control	Non-original research (review, guidelines, meta-analysis, etc.); non-surgical or non-medical groups included or compared; other intervention of interest is used; outcomes of interest are not included; ascertainment of IGF-I is not included; follow-up period <2 weeks; non-acromegaly population; patients received another treatment prior to or during enrollment
COVID	Study-related to adults 18 years or older, Patients with ARDS (acute respiratory distress syndrome), Cell therapy transplantation, usual and supportive care only, no treatment, Any study design including case reports	Non-original data (e.g., narrative reviews, editorials, letters or erratum); qualitative studies; cost-benefit analysis; cross-sectional (i.e., non-longitudinal) studies; animal studies; children

Table A.1: Inclusion and Exclusion Criteria for Each Dataset. RCT: Randomized Control Trial, SRMA: Systematic Review and Meta Analysis

A.2 DATA SAMPLES

Data Samples	Label
Antitumor Effect of Nivolumab on Subsequent Chemotherapy for Platinum-Resistant Ovarian Cancer. Platinum-resistant recurrent ovarian cancer is generally refractory to chemotherapy. Programmed cell death-1 (PD-1) signaling is a new target for antitumor therapy. The anti-PD-1 antibody nivolumab had a 10% durable complete response rate in our phase II clinical trial. However, how nivolumab affects sensitivity to subsequent chemotherapy remains unclear. We encountered several cases of unexpected antitumor response among patients who underwent palliative chemotherapy in the follow-up study of our phase II nivolumab trial (UMIN000005714). Several agents had an unexpected antitumor response in patients who were resistant or refractory to standard chemotherapeutic agents. In one patient, both pegylated liposomal doxorubicin (PLD) and nedaplatin (CDGP) resulted in partial response. In another patient, PLD and CDGP resulted in partial response and stable disease, respectively. These two patients remained alive on the cutoff date. These two cases raise the possibility that nivolumab might improve sensitivity to adequate chemotherapy for ovarian cancer.	Include
Anti-HER agents in gastric cancer: from bench to bedside. Despite some advances in the past few years, the search for effective treatment modalities for advanced gastric and gastro-esophageal junction cancer is far from over. Available data clearly demonstrate that the development of new drugs will have little, if any, chance of success if it is not guided by in-depth knowledge of disease biology. However, using biologic agents to target key molecular pathways, such as those regulated by human epidermal growth factor receptor (HER) family members, may be effective. Indeed, the positive results achieved by the anti-HER2 agent trastuzumab in a phase III trial in HER2-positive patients support this approach. Many new anti-HER molecules are now under evaluation for the treatment of gastric and gastro-esophageal junction cancer, but so far attempts to identify reliable predictive factors from phase I and II trials have produced inconclusive results. In addition, large phase III trials are still being conducted in molecularly unselected populations. Refining patient selection is essential to maximize the benefit of targeted agents, to avoid significant toxicities and for the development of alternative therapeutic approaches in patients who have nonresponsive disease.	Exclude

Table A.2: Manually Annotated Positive (i.e., Include) and Negative (i.e., Exclude) Samples from ICI Dataset. Here, the Data Sample is Concatenation of Title and Abstract from the Candidate Article

APPENDIX B

EMPIRICAL STUDY OF HIGH-SENSITIVITY METHODS

B.1 QUERIES FOR BM25

Here are the queries formulated from inclusion criteria for each dataset.

- **ICI:** clinical trials RCT randomized control trials randomly assigned immunotherapy IO drugs ICI PD PDL1 checkpoint inhibitors Pembrolizumab Nivolumab Cemiplimab Atezolizumab Avelumab Durvalumab CTLA-4 MPDL3280A MSB0010718C MEDI4736 BMS-734016 MDX-010 MDX-101 ONO-4538 BMS-936558 MDX1106 MK-3475 REGN-2810 placebo two arm
- **HRT:** randomized and non-randomized comperative studies related to Hormone replacement therapy HRT for postmenopausal women of any age with follow up period of more than 6 months and comparision with no HRT
- **Cooking:** comparison with any control group cooking class culinary intervention delivered by anyone chef dietitian adult child healthy with comorbidity
- **Accelerometer:** randomized and non-randomized comparative studies related to use of different techniques or accelerometers in validation experiments on children ages one to eighteen 1-18
- **Acromegaly:** observational longitudinal randomized non-randomized case control series report comparative study patients receiving medical treatment of acromegaly microadenoma macroadenoma surgical treatment as first line of treatment somatostatin analogues octreotide lanreotide dopamine agonists cabergoline biochemical control igf level morning gh hypopituitarism ha mortality
- **COVID:** patient with ARDS acute respiratory distress syndrome failure adults 18 years and older cell therapy transplantation stem cell mesenchymal stromal msc progenitor cell ips ipsc exosome extracellular vesicle secretome supportive care publications 1990