The Academic College of Tel-Aviv

THE SCHOOL OF COMPUTER SCIENCE

Automatic Extraction of Disease Risk Factors from Medical Publications

Thesis submitted in partial fulfillment of the requirements for the M.Sc. degree in the School of Computer Science of the Academic College of Tel Aviv University

By

Maxim Rubchinsky

The research work for the thesis has been carried out under the supervision of

Dr. Adi Shraibman

Dr. Dorit Shweiki

Dr. Ella Rabinovich

July 2024

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisors, Dr. Adi Shraibman, Dr. Dorit Shweiki, and Dr. Ella Rabinovich. Your unwavering support, invaluable guidance, and profound knowledge have been instrumental in shaping this research. Your encouragement and insightful feedback have not only enhanced the quality of this work but also significantly contributed to my personal and professional growth.

A special thank you goes to Netanel Golan, the medical student, and Tali Sahar, who have diligently performed the manual annotation and verification of data. Your meticulous work and dedication have been essential in ensuring the accuracy and reliability of the datasets used in this study.

Thank you all for your invaluable support.

Table of Contents

1 Introduction

Research Topic

This thesis presents an approach to automating the identification of disease risk factors from medical literature. Leveraging advances in machine learning and natural language processing, and specifically leveraging pre-trained models in the bio-medical domain, while tuning them for the specific task—this research aims to develop a comprehensive pipeline that automates the extraction of risk factors from a vast array of medical articles. This system identifies relevant articles, classifies them based on the presence of risk factor discussions, and extracts specific risk factor information through a sophisticated question-answering model.

Justification

The identification of risk factors for diseases is crucial in preventive medicine, enabling healthcare professionals to devise effective prevention strategies and enhance patient outcomes. Traditional methods heavily rely on the manual review of extensive medical literature—a process that is both time-consuming and labor-intensive. Automated tools can significantly streamline this process, improve knowledge accessibility, and facilitate the effective use of information. For instance, recent compelling evidence has emerged linking Lipoprotein A (Lp(a)) --- a particle operating similarly to the more familiar LDL molecule --- to the pathogenesis of atherosclerosis and subsequent coronary artery disease, commonly referred to as Myocardial Infarction (MI). Despite the established role of Lp(a) as a risk factor [1],many primary care clinicians remain inadequately informed, occasionally lacking knowledge regarding its testing procedures. Moreover, in a conversation with a board-certified professor of interventional cardiology, he disclosed receiving frequent inquiries from other clinicians questioning the necessity of referrals for Lp(a) testing. This highlights the pressing need for an automated tool capable of screening vast amounts of scientific literature and identifying prominent risk factors for various diseases.

Scope

This research focuses on the extraction of risk factors specifically from scientific medical literature, as opposed to the analysis of electronic health records. Here, the primary challenge lies in the diverse and unstructured nature of medical publications, where risk factors are described in various contexts and formats. Moreover, the continuous discovery of new risk factors necessitates a dynamic approach that can adapt to the evolving body of medical knowledge.

The scope of this thesis is distinctly focused on the extraction of risk factors from scientific medical literature. Utilizing pre-trained large language models, based on BioBERT [2], the research builds a specialized multi-step system that first identifies relevant medical articles, classifies them based on the presence of risk factor discussions, and then extracts specific risk factor information through a question-answering (QA) model.

Additionally, our work has been recognized by the academic community. Our paper submission, titled "Automatic Extraction of Disease Risk Factors from Medical Publications" has been accepted to appear at the BioNLP 2024 workshop as a poster and it will be published in the proceedings of BioNLP@ACL2024.

All our code, trained models and data are available at the following repositories:

GitHub: https://github.com/maximrub/diseases-risk-factors

Hugging Face: https://huggingface.co/diseases-risk-factors

Our approach to the extraction of disease risk factors is illustrated in the following figure that describes the pipeline for extraction of disease's risk factors:

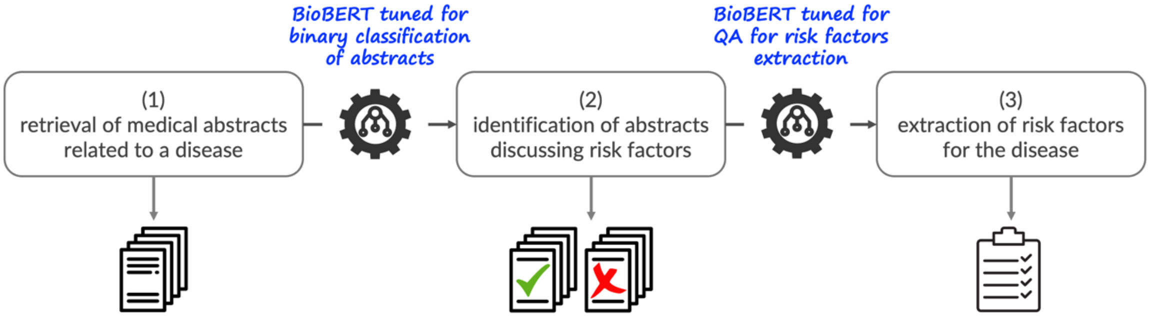

BioBERT tuned for binary classification of abstracts
BioBERT tuned for QA for risk factors extraction
(1) retrieval of medical abstracts related to a disease
(2) identification of abstracts discussing risk factors
(3) extraction of risk factors for the disease

Figure 1 pipeline for extraction of disease's risk factors

1. Retrieval of relevant medical abstracts from extensive databases like PubMed.
2. Employment of a specifically fine-tuned binary classifier to discern articles that discuss risk factors.
3. Application of a fine-tuned question-answering model on manually annotated QA items to precisely extract textual spans containing the risk factors.

A significant component of this thesis is the development and utilization of comprehensive datasets. These include a meticulously curated set of over 1,700 risk factors linked to 15 different diseases and a larger dataset comprising over 160,000[1] automatically extracted risk factors, nearly 1,500 of which have been manually evaluated to refine the system's accuracy and reliability.

Limitations

Our study, while contributing valuable insights into the automation of risk factor identification from medical publications, is subject to several limitations that merit a thorough discussion.

One of the primary limitations is the challenge of accurately distinguishing risk factors specifically associated with the disease in question (type 1) from valid risk factors that are not directly related to the disease under investigation (type 2). While our models demonstrated a high capacity for identifying potential risk factors, the precision in contextualizing these factors to specific diseases varied. This aspect highlights a critical area for future research, emphasizing the need for enhanced specificity in the models to improve their utility in targeted medical research and practice.

[1] We note that the set of over 160,000 automatically extracted risk factors are of admittedly mixed quality (see Human Evaluation and Table 6 Distribution of manual evaluation annotations by disease family for details), yet, we thought this data can serve the community for further research in the field.

Moreover, the study's reliance on free-text medical articles introduces variability in the data quality and representation. The unstructured nature of these texts and the diversity in how risk factors are described pose significant challenges for both the binary classification and question-answering models. Efforts to standardize data representation and improve model robustness against such variability are essential steps forward.

The datasets used in this study, while extensive, are not exhaustive. The landscape of medical research is continuously evolving, with new findings emerging regularly. The datasets, therefore, represent a snapshot in time, and ongoing efforts to update and expand these resources are necessary to maintain their relevance and utility.

Finally, the study's scope was constrained by the computational resources available. Future work could explore more complex models or ensemble approaches that might offer improved accuracy but require more substantial computational power.

Despite these limitations, this study represents a significant step toward automating the identification of disease risk factors from medical literature. Acknowledging and addressing these limitations in future research will be crucial for advancing the field and enhancing the practical applicability of these technologies in healthcare.

2 Related Work

The automatic identification of disease risk factors through the analysis of medical texts has increasingly become a focal point across various research domains, particularly in applying natural language processing (NLP) and machine learning techniques to electronic health records (EHRs) and electronic medical records (EMRs), emphasizing the intersection of natural language processing (NLP) and machine learning with healthcare informatics. This section reviews seminal contributions that either parallel or diverge from our approach, which uniquely focuses on free-text medical articles.

Deep learning's integration into healthcare has predominantly centered around structured electronic health records (EHRs) and electronic medical records (EMRs). For instance, Chokwijitkul et al. (2018) [3] utilize deep learning techniques to extract heart disease risk factors from EHRs. Their methodology, which is rooted in the analysis of structured data, contrasts sharply with our focus on unstructured textual data from medical literature, highlighting the variability in data sources utilized for risk factor identification.

Boytcheva (2017) [4] employs association rules to mine clinical texts for risk factors, particularly using data from the Diabetes Register formatted in XML. This structured approach to data analysis starkly differs from our application of pre-trained large language models to unstructured, free-text medical articles, underscoring the diversity in methodologies for text understanding.

A comprehensive work on identifying risk factors for heart disease (from clinical data) over time was done in a shared task organized by UTHealth[2].

(Stubbs et al., 2015) [5], Sheikhalishahi et al. (2019) [6] offer an overview of NLP applications in analyzing clinical notes for chronic disease management, highlighting the increasingly significant contribution of language models to healthcare applications.

In the domain of precision medicine, Sabra et al (2017) [7] focus on extracting semantic information and assessing sentiments in clinical notes.

Various works have employed data mining and machine learning (ML) techniques for identifying risk factors from patient data (Abdelhamid et al 2023) [8], or clinical outcome

[2] The University of Texas Health Science Center.

prediction (Kavakiotis et al 2017 [9]; Mehmood et al 2021 [10]; Naik et al 2021 [11]). Recently, the identification of risk factors for delirium prediction, a rare adverse reaction observed in COVID-19 patients, was developed utilizing ML applied to nursing records (Miyazawa et al 2024 [12]).

Additional line of studies focuses on building language models specifically-tailored for medical literature related tasks (Roitero et al 2021 [13]; Yang et al 2022 [14]; Singhal et al 2023 [15]).

Several significant contributions have been made in the field of biomedical relation extraction, which includes identifying factors that predispose individuals to diseases. The SemRep [16] tool extracts semantic predications from biomedical texts, including relationships such as "predisposes". The outputs of SemRep have been used to create SemMedDB [17], a large-scale repository of semantic predications from PubMed. Building on these resources, BioPREP [18] employs deep learning techniques for predicate classification. The BioRED [19] dataset includes a "positive correlation" relation between diseases and other biomedical entities like genes and chemicals.

Conclusion

While the majority of existing research focuses on analyzing structured electronic health records and electronic medical records to identify disease risk factors, our study pushes beyond these confines by examining free-text medical literature. Processing unstructured medical text introduces distinct challenges, especially due to language complexity, variation, and the potential for nuanced double meanings, and even worse, due to the necessity to discern context accurately. Consequently, it opens broad opportunities for subtle understandings of disease risk factors, facilitating both research and practical applications.

3 Methodology

Datasets

The data collection process for this research can be viewed as a three-step process:

(i) Collection of the set of disease names spanning multiple disease families.

(ii) Manual annotation of scientific article abstracts containing explicit mention of risk factors of a subset of diseases -- "abstracts seed".

(iii) Manual annotation of risk factors description (span) in abstract texts found in (ii) -- "risk factors seed".

We detail on each step in this multi-phase procedure.

Disease Dataset Collection

Aiming to assemble a comprehensive list of diseases, we made use of the KEGG Disease Database API[3]; specifically, we used its REST API service at https://www.kegg.jp/kegg/rest to retrieve disease-related information, including names, description and relevant medical codes such as MeSH (Medical Subject Headings), ICD-10 and ICD-11[4].

This process resulted in 2,624 distinct disease names, comprising the foundation for further retrieval of scientific abstracts and, ultimately, automatic extraction of risk factors, from scientific medical literature.

Seed Dataset with Relevant Abstracts

Retrieval of Abstracts Discussing Risks

Using the list of disease names retrieved from KEGG, we next queried PubMed[5], a large, reliable, and authoritative resource of biomedical literature, for article abstracts containing the disease names. Specifically, we used the Entrez Programming Utilities[6] via the biogo package[7].

The inherent limitation of this study is related to the fact that only abstracts are freely available through the PubMed interface. However, paper abstracts typically contain a concise summary and main findings of the work, hence constitute a sufficient input for the task at hand. Similarly, prior studies analyzed abstracts retrieved from PubMed for

[3] KEGG database: https://www.kegg.jp/kegg/disease

[4] As of April 2024, ICD-11 (International Classification of Diseases, v11) is the most up-to-date code collection.

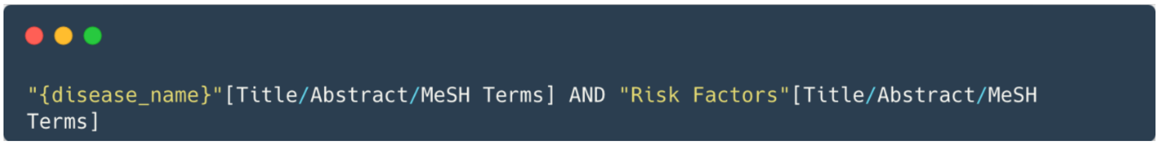
[5] https://pubmed.ncbi.nlm.nih.gov

[6] https://www.ncbi.nlm.nih.gov/books/NBK25501

[7] https://github.com/biogo/ncbi

building a biological network [20], topical clustering [21], and identification of negative and positive domain-specific medical terms [22].

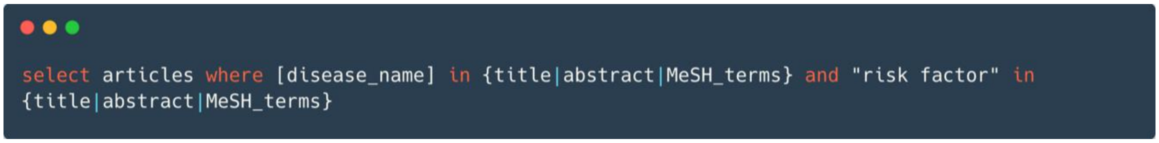
Aiming at retrieval of abstracts discussing findings related to risk factors, we queried PubMed for containment of the phrase "risk factor" and the disease name in a paper's information: title, abstract or MeSH terms using the following search term:



```
"{disease_name}"[Title/Abstract/MeSH Terms] AND "Risk Factors"[Title/Abstract/MeSH
Terms]
```

Equation 1 Articles search term

This search term is equivalent to the following pseudo-code:



```
select articles where [disease_name] in {title|abstract|MeSH_terms} and "risk factor" in
{title|abstract|MeSH_terms}
```

where {disease_name} refers to the disease we are seeking risk factors for, and the exact search term "risk factor" (surfacing also the plural "risk factor**(s)**) can appear in abstract, title or MeSH terms.

Annotation of Abstracts for Risk Factors

Despite the evident potential, not every abstract with explicit mention of "risk factor" or marked with a "risk factor" MeSH term contains risk factors for a pre-defined disease. As a concrete example, a medical study can mention a list of potential risk factors tested, without any of them showing as significant. We, therefore, define our first (pre-processing) task as automatic classification of a retrieved abstract for spelling out an artifact, found to be a risk factor for the disease in the study.

A qualified annotator with medical background (one of the authors of this paper) annotated a random set of 182 abstracts. The procedure resulted in 87 positive abstracts (explicitly mentioning a risk factor) and 95 negatives, thereby comprising a sufficient training set for the binary classifier - step (2) in Figure 1 pipeline for extraction of disease’s risk factors.

The following table shows two examples of relevant abstract parts containing risk-related phrases retrieved per our Equation 1 Articles search term, which do or do not qualify as risk factors, as identified by the annotator. Evidently, the nuanced language used to discuss risks in various contexts renders the task as non-trivial for both humans and automatic tools.

article title: **Risk Factors for Pediatric Human Immunodeficiency Virus-related Malignancy** (2003)
Context: Although cancers occur with increased frequency in children with human immunodeficiency virus (HIV) infection, the specific clinical, immunological, and viral risk factors for malignancy have not been identified. Objective: To identify risk factors for malignancy among HIV-infected children. [...] Epstein-Barr virus viral load of more than 50 viral genome copies per 105 peripheral blood mononuclear cells was strongly associated with cancer risk but only for children with CD4 cell counts of at least 200/ microL (odds ratio [OR], 11.33; 95\% confidence interval [CI], 2.09-65.66, P<.001). [...] High viral burden with EBV was associated with the development of malignancy in HIV-infected children although the effect was modified by CD4 cell count. The pathogenesis of HIV-related pediatric malignancies remains unclear and other contributing risk factors can be elucidated only through further study.
article title: **Profound Hypoglycemia and High Anion Gap Metabolic Acidosis in a Pediatric Leukemic Patient Receiving 6-Mercaptopurine** (2024)
A 13-year-old male undergoing maintenance chemotherapy with methotrexate and 6-mercaptopurine (6MP), for very high-risk B-cell acute lymphoblastic leukemia (ALL), presented with vomiting due to severe hypoglycemia with metabolic acidosis. While his laboratory values were concerning for a critically ill child, the patient was relatively well appearing. Hypoglycemia is a rare but serious side effect of 6MP with an unexpectedly variable presentation; therefore, a high index of suspicion is needed for its prompt detection and treatment. [...] 6MP-induced hypoglycemia can be ameliorated with the addition of allopurinol to shunt metabolism in favor of the production of therapeutic metabolites over hepatotoxic metabolites. Additionally, a morning administration of 6MP and frequent snacks may also help to prevent hypoglycemia. Overall, this case adds to the literature of unusual reactions to 6MP including hypoglycemia in an older child without traditional risk factors.

Table 1 Top -- abstract identified as relevant for risk factors extraction by the annotator, where the highlighted part refers to the discussed factor. Bottom -- abstract mentioning "risk factors" yet annotated as irrelevant.

QA Seed Dataset with Risk Factors

Given an article abstract specifying a risk factor(s) for a certain disease, we cast the risk factor identification problem as *extractive question answering* scenario, where given the abstract and the question "What are the risk factors for {disease name}?", a textual span, containing the answer, will be identified.

In the Section Identification of Disease Risk Factors we make use of the established and popular BERT-based QA model – BioBERT [2], and fine-tune it for the task at hand using a manually annotated set of QA items: context (article abstract), a targeted question of the form mentioned above, and a set of manually marked answers in the form *span_start* and *answer_text* (implying *span_end*).

In the absence of suitable annotated datasets for this nuanced task, we have developed a web interface for medical students to manually annotate article abstracts. This interface is used for (manual) identification of text segments within abstracts, given the disease discussed in the article.

We will review the annotating tool in the Web UI section and release the tool for the community.

The annotator with medical background marked text spans containing risk factors in a random set of 668 abstracts identified to contain explicit mention of a risk factor[8], resulting in the total of 1,712 QA items, spanning 15 diverse diseases listed in the Diseases with Annotated Risk Factors in the QA dataset (the training set) section, where each QA item reflects a single risk factor in an abstract that (possibly) encompasses multiple valid risks. Sentences suggesting risk factors significant only within specific population subgroups were denoted as such.

[8] The abstracts were sampled from the set automatically classified as "positive"

We present two examples of QA items: disease name, abstract, and the highlighted risk factor span, as marked by the annotator.

disease: **Diabetes in Men**
OBJECTIVE: To examine the association between smoking, alcohol consumption, and the incidence of non-insulin dependent diabetes mellitus in men of middle years and older. [...] RESULTS: During 230,769 person years of follow up 509 men were newly diagnosed with diabetes. After controlling for known risk factors men who smoked 25 or more cigarettes daily had a relative risk of diabetes of 1.94 (95\% confidence interval 1.25 to 3.03) compared with non-smokers. Men who consumed higher amounts of alcohol had a reduced risk of diabetes (P for trend < 0.001). Compared with abstainers men who drank 30.0-49.9 g of alcohol daily had a relative risk of diabetes of 0.61 (95\% confidence interval 0.44 to 0.91). CONCLUSIONS: Cigarette smoking may be an independent, modifiable risk factor for non-insulin dependent diabetes mellitus. Moderate alcohol consumption among healthy people may be associated with increased insulin sensitivity and a reduced risk of diabetes.
disease: **Breast and Colorectal Cancer**
BACKGROUND: Increasing evidence suggests that diabetes mellitus (DM) is associated with increased cancer incidence and mortality. Several mechanisms involved in diabetes, such as promotion of cell proliferation and decreased apoptosis, may foster carcinogenesis. This study investigated the association between DM and cancer incidence and cancer-specific mortality in patients with breast and colorectal carcinoma. [...] The overall HR for breast cancer incidence was 1.23 (95 per cent confidence interval 1.12 to 1.34) and that for colorectal cancer was 1·26 (1·14 to 1·40) in patients with DM compared with those without diabetes. The overall HR was 1.38 (1.20 to 1.58) for breast cancer- and 1.30 (1.15 to 1.47) for colorectal cancer-specific mortality in patients with DM compared with those without diabetes. CONCLUSION: This meta-analysis indicated that DM is a risk factor for breast and colorectal cancer, and for cancer-specific mortality.

Table 2 Example of two paper abstracts manually annotated for risk factors

Note that in some cases the precise name of the risk factor (e.g., "cigarette smoking") for a disease (e.g., "diabetes in men") is annotated in its broader context, to ensure the model is trained to extract risk factors tied to the disease, and not other, unrelated, artifacts.

Collectively the carefully curated and annotated set of abstracts for binary classification of medical articles, and the set of QA items, comprise a high-quality collection for tuning pre-trained language models for the purpose of this study.

Diseases with Annotated Risk Factors in the QA dataset (the training set)

family	disease
Autoimmune disease	Celiac disease
Autoimmune disease	Rheumatoid arthritis
Autoimmune disease	Type 1 diabetes mellitus
Carcinomas	Bladder cancer
Carcinomas (to the most part)	Breast cancer
Carcinomas (to the most part)	Colorectal cancer
Chronic lung disease	Chronic obstructive pulmonary disease
Chronic lung disease	Asthma
Circulatory disorder	High blood pressure
Heart disease	Myocardial infarction
Melanoma/Skin cancer	Melanoma
Metabolic disease	Metabolic syndrome
Metabolic disease	Type 2 diabetes mellitus
Neurodegenerative disorder	Alzheimer disease
Neurologic disorder	Migraine

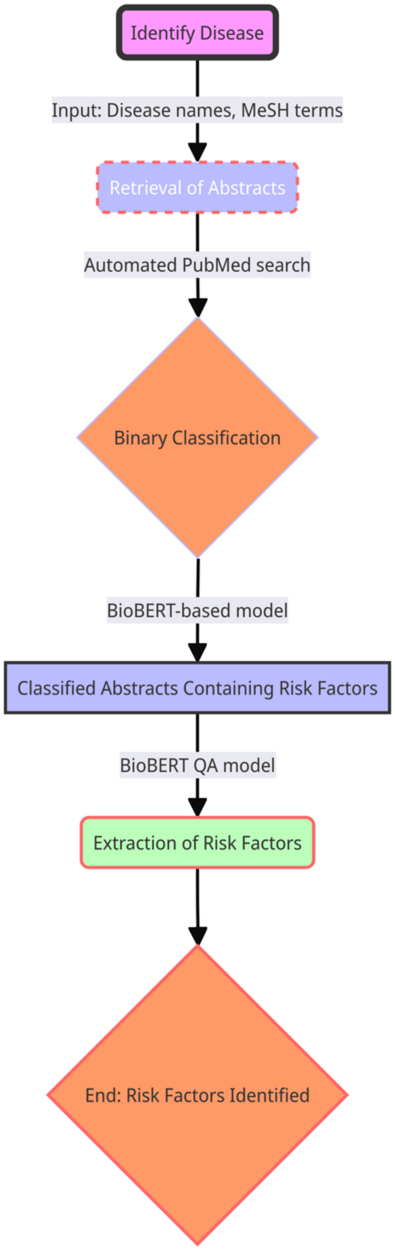
Table 3 Disease distribution by disease family in the manually annotated set of 1,712 risk factors used for BioBERT QA fine-tuning

Methodology and Experiments

We further describe in detail our methodological approach, experimental setup and results.

Methodology

We apply a multi-step approach to automate the identification of disease risk factors from medical literature. Central to our methodology is the use of BioBERT, a variant of BERT pre-trained on biomedical texts, enabling nuanced understanding of complex medical language [2].


Identify Disease
Input: Disease names, MeSH terms
Retrieval of Abstracts
Automated PubMed search
Binary Classification
BioBERT-based model
Classified Abstracts Containing Risk Factors
BioBERT QA model
Extraction of Risk Factors
End: Risk Factors Identified

We next provide details on each step in the process. This model was chosen due to its proven benefits in the biological domain, and its encoder-based architecture -- (arguably) the most appropriate choice for both the classification and extractive question answering tasks at hand[9].

Detection of Abstracts with Risk Factors

The pre-trained BioBERT-based classifier[10] was tuned for abstracts classification using the training part (80%) out of over 182 manually annotated abstracts (see Section Annotation of Abstracts for Risk Factors) and tested on the held-out part (20%), achieving the accuracy of 92%. The following table reports the per-class classification results.

class	Precision	Recall	F1-score
POS (with risk factor)	0.89	0.94	0.92
NEG (w/o risk factor)	0.94	0.89	0.92

Table 4 Classification results reported on the test set (20%) of the manually annotated 182 abstracts

This encouraging result facilitated our efforts of analyzing content that is most likely to yield valuable insights into disease-risk factor associations.

We collected a substantial dataset of abstracts, by querying PubMed for each one of over 2400 diseases, as detailed in Section Seed Dataset with Relevant Abstracts; this step resulted in 137,740 abstracts. We next apply the fine-tuned classifier to identify abstract potentially containing risk factors for a disease. Out of the total number of 137,740 abstracts, 89,834 were classified as positive -- containing explicit mentions of risk factors for diseases. Naturally, some diseases (and disease families) resulted in more prolific retrieval, due to their higher coverage in the medical literature: while various cancer types (e.g., Carcinoma, Leukemia) have large body of related articles, genetic disorders are surveyed less frequently in the context of risk factor discussion.

Identification of Disease Risk Factors

The collection of abstracts classified positively to contain a risk factor, was then subject to the task of risk factor extraction -- step (3) in Figure 1 pipeline for extraction of disease's risk factors.

We cast the task as extractive QA, where the medical abstract represents the context, and the question template is formulated as "What are the risk factors for {disease name}?". We anticipate the BioBERT QA model [2] to identify span(s) in the abstract containing the answer (or answers, in case multiple risk factors are mentioned in the

[9] Our future work includes investigation of decoder-based models (e.g., GPT), casting the QA part as an abstractive task.

[10] https://huggingface.co/dmis-lab/biobert-v1.1

same abstract), similarly to examples presented in Table 2 Example of two paper abstracts manually annotated for risk factors. We fine-tune the model for the specific task, as described below.

Fine-tuning the QA Model

We tuned the BioBERT model for our use case using the training part (80%) of the 1,712 QA items annotated manually by the author with medical background (see Section Seed Dataset with Relevant Abstracts); the remaining 20% were used for testing. Notably, the set of 15 diseases in the 668 abstracts was carefully split into training and test sets, so that the same disease does not appear in both sets, facilitating the assessment of the model's generalizability and performance across a variety of disease contexts.

The model tuning was done using the maximum context length of 384 tokens, learning rate of 2e-5, and 25 epochs.

We use two common metrics for automatic evaluation of extractive question answering: *exact-match* and *F1-score*. Applied on the test set (342 QA items), the metrics obtained 61.76% for exact-match, and 88.23% for *F1-score* highlighting the potential of the approach.

Determining the Maximum Answer Length

We have determined the maximum length for answers in our QA model by analyzing the lengths of all answers within our training dataset. We calculated the length of each answer (in characters) and studied their distribution. The maximum answer length was set at the 95th percentile of these lengths to encompass the majority of real-world answers while excluding outliers. This threshold is crucial for maintaining focus on concise and relevant answer segments, thereby enhancing the model's training and operational effectiveness. In practice, when the model evaluates potential answers, it only considers text segments whose length does not exceed this predefined limit. Specifically, the text extracted between the predicted start and end indices is compared against the maximum length, and any text exceeding this threshold is disregarded.

Identification of Risk Factors at Scale

Utilizing the fine-tuned QA model, we then processed the collected abstracts to identify and validate risk factors for a wide range of diseases, culminating in a dataset that catalogs these findings in much detail. As a concrete example, the entry for the "B-cell acute lymphoblastic leukemia" includes 16 (not necessarily unique) automatically extracted risk factors. Along with the extracted span, the BioBERT QA model provides its probability (confidence, in the 0-1 range) for the identified answer. For a given disease, we only considered answers exceeding the confidence of $0.6 * \max\ \_answer\_probability$, where the *max_answer_probability* is the maximum probability assigned to an answer for the disease.

The final dataset encompasses the total of 162,409 identified risk factors spanning 744 diseases, extracted from 54,820 PubMed abstracts.

Due to the inherently strict nature of the *exact-match* metric, we could observe multiple cases where the extracted answer was largely correct, but didn't represent a precise overlap with the "gold" answer due to a single missing or redundant word. In particular, while some cases surface useful information about a disease risk factors, they are marked as inaccurate by the automatic metric. We complement the evaluation pipeline by sampling a large amount of (automatically identified) risk factors for diseases, and performing fine-grained human assessment of the results' quality.

Risk Factor Annotation System

Overview

The risk factor annotation system comprises three main components designed to streamline the process of annotating risk factors in medical articles. This system was instrumental in creating the datasets used in our research.

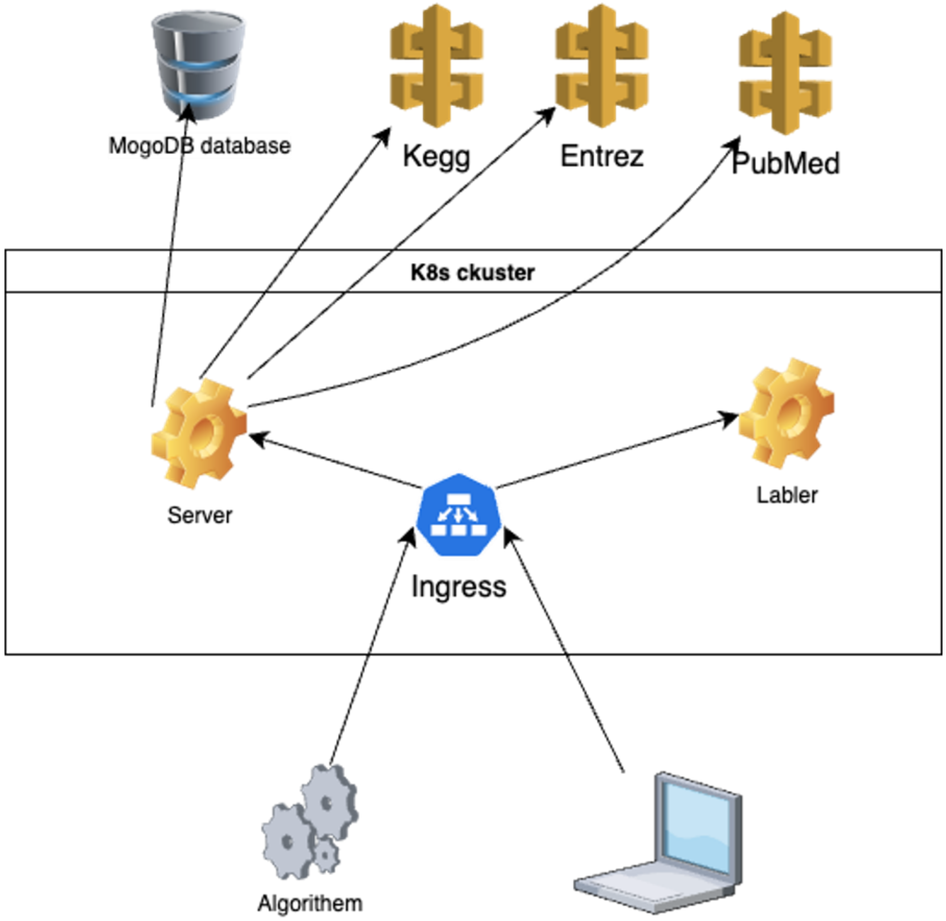


Figure 2 Risk Factor Annotation System

Components

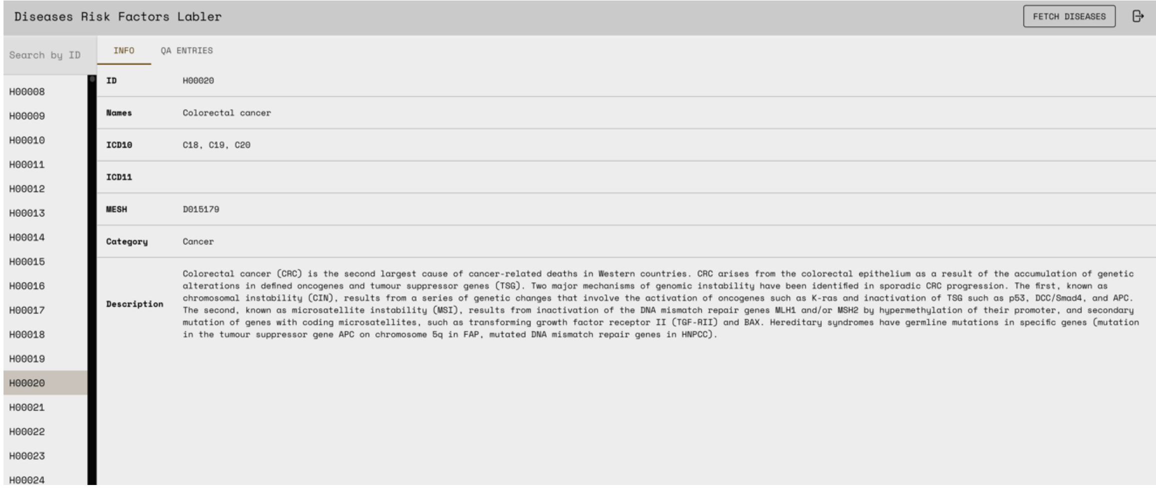
GraphQL Server

The backbone of the system is a GraphQL server, which serves as the central communication hub. Hosted on Kubernetes (k8s) for scalability and reliability, the server facilitates data exchange between the user interface and the database. It handles requests for data retrieval and submission, ensuring that the web application and the code can access and store data efficiently.

Web UI

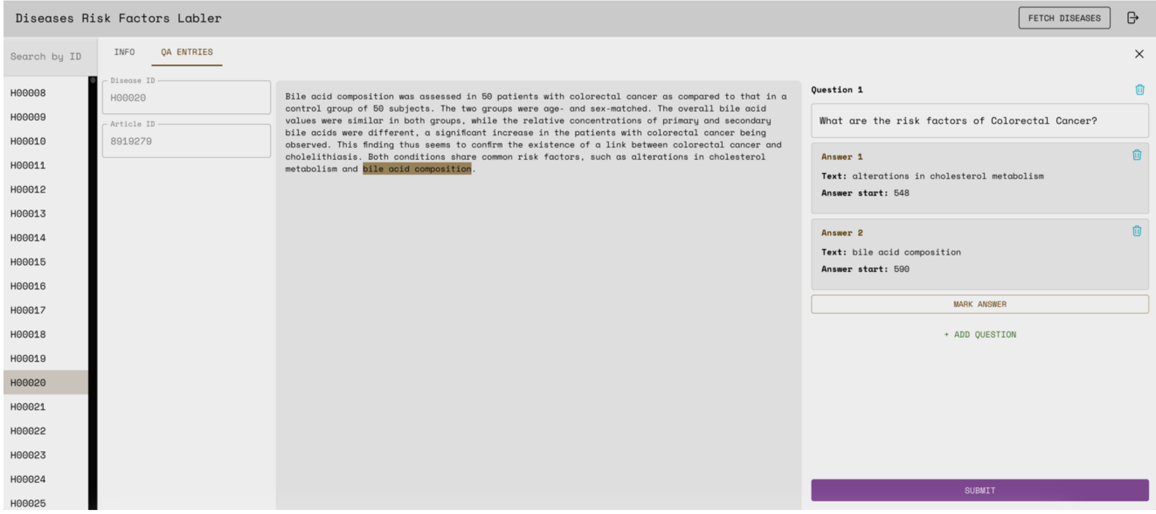
The front end of the system is a React-based web application, also deployed on Kubernetes for high availability. This intuitive user interface allows medical students and researchers to interact with the system, including retrieving medical articles, annotating risk factors within texts, and submitting these annotations back to the server. The design prioritizes ease of use to facilitate accurate and efficient annotation work.

The following two figures illustrate two screenshots of the application developed for manual annotation of risk factors; the system code will also be made available.



ID	H00020
Names	Colorectal cancer
ICD10	C18, C19, C20
ICD11	
MESH	D015179
Category	Cancer
Description	Colorectal cancer (CRC) is the second largest cause of cancer-related deaths in Western countries. CRC arises from the colorectal epithelium as a result of the accumulation of genetic alterations in defined oncogenes and tumour suppressor genes (TSG). Two major mechanisms of genomic instability have been identified in sporadic CRC progression. The first, known as chromosomal instability (CIN), results from a series of genetic changes that involve the activation of oncogenes such as K-ras and inactivation of TSG such as p53, DCC/Smad4, and APC. The second, known as microsatellite instability (MSI), results from inactivation of the DNA mismatch repair genes MLH1 and/or MSH2 by hypermethylation of their promoter, and secondary mutation of genes with coding microsatellites, such as transforming growth factor receptor II (TGF-RII) and BAX. Hereditary syndromes have germline mutations in specific genes (mutation in the tumour suppressor gene APC on chromosome 5q in FAP, mutated DNA mismatch repair genes in HNPCC).

Figure 3 Disease Risk Factor Annotation System: disease details as retrieved from KEGG and parsed



Disease ID: H00020

Article ID: 8919279

Bile acid composition was assessed in 50 patients with colorectal cancer as compared to that in a control group of 50 subjects. The two groups were age- and sex-matched. The overall bile acid values were similar in both groups, while the relative concentrations of primary and secondary bile acids were different, a significant increase in the patients with colorectal cancer being observed. This finding thus seems to confirm the existence of a link between colorectal cancer and cholelithiasis. Both conditions share common risk factors, such as alterations in cholesterol metabolism and bile acid composition.

Question 1

What are the risk factors of Colorectal Cancer?

Answer 1

Text: alterations in cholesterol metabolism

Answer start: 548

Answer 2

Text: bile acid composition

Answer start: 590

Figure 4 Disease Risk Factor Annotation System: manual annotation of spans containing risk factors; multiple risk factors for the same disease can be identified in the same abstract

Python Algorithm

Complementing the user interface is a Python-based algorithm that interacts with the GraphQL server. This component is responsible for processing medical articles, including sending requests to the server to fetch articles for annotation and submitting the results of automated risk factor identification processes. It plays a critical role in pre-processing and post-processing steps in the dataset creation pipeline.

Database

At the core of the system lies a MongoDB database hosted on Azure Cosmos DB. This NoSQL database was chosen for its scalability, flexibility, and robust support for storing unstructured data, such as medical article texts and annotations. It stores all data related to diseases, articles, and user annotations, providing a persistent and reliable data storage solution for the system.

System Use Cases

The Risk Factor Annotation System was developed to address the challenge of identifying and annotating risk factors in free-text medical articles.

Its design facilitates a comprehensive workflow:

- Medical students and researchers use the Web UI to access and annotate articles, marking text segments that represent risk factors for various diseases.
- The Python algorithm processes large volumes of articles, applying machine learning models to suggest potential risk factors.
- All interactions with the article data and annotations are managed through the GraphQL server, ensuring seamless data flow between the components.
- The MongoDB database serves as the central repository for storing and managing all data generated and used by the system.

Conclusion

The development of this system represents a significant step forward in the automation of risk factor annotation from medical literature. By leveraging modern web technologies, cloud computing resources, and advanced database management systems, the system offers a scalable and efficient tool for advancing medical research.

4 Findings (Results)

Human Evaluation

We next manually evaluated a random sample of 1,485 extracted risk factors spanning 29 various diseases (constituting roughly 1% of the full set of extracted factors), based on their validity and relevance to the disease in question.

Evaluation Scheme

We designed a specifically-tailored, four-tiered annotation scheme for the sake of reliable and accurate evaluation, as detailed below. Each risk factor was scored with one of three annotation marks, following the below annotation scheme:

(1) VALID RISK FACTOR FOR THE SPECIFIED DISEASE	Correctly identified risk factor extracted for the disease of interest, i.e., the disease in the question introduced to the QA system.
(2) VALID RISK FACTOR FOR A DIFFERENT DISEASE	Correctly identified risk factor for a different disease, i.e., not the disease in the question introduced to the QA system, indicating capabilities yet highlighting challenges in specificity.
(3) INVALID RISK FACTOR	Phrases and terms that are not considered medical risk factors.

Additional distinction was done within the first group (valid risk factor), annotating risk factors with strong statistical correlation, as evident from the abstract by inspecting statistical measurements as odd ratio (OR), and confidence intervals (CIs) -- metrics often used in medical literature for testing the significance of findings, such as the presence of a factor in one population but not the other. 41 out of the total of 1,485 were marked as *highly significant* risk factors; we release these annotations as well to facilitate further research in the community.

Evaluation Results

The following table presents error analysis of correctly- and incorrectly-identified risk factor examples (the first two rows), as well as an example for artifact that does not constitute a risk factor (the last row).

disease	abstract excerpt (identified risk factor highlighted)	marker
Chronic Myeloid Leukemia	[...] RESULTS: Previous diagnoses of dyspepsia, gastritis or peptic ulcers, as well as previous proton pump inhibitor (PPI) medication, were all associated with a significantly increased risk of CML (RRs, 1.5-2.0; P = 0.0005-0.05). Meanwhile, neither inflammatory bowel disease nor intake of NSAIDs were associated with CML, indicating that it is not gastrointestinal ulcer or inflammation per se that influences risk. [...]	**1**
Cystic Fibrosis	BACKGROUND: Cystic fibrosis, like other chronic diseases, is a risk factor for the development of elevated symptoms of depression and anxiety. [...] Patient anxiety (OR 2.33) and depression (OR 4.09) were significantly associated with forced expiratory volume in one second (FEV1) <40\% and forced vital capacity (FVC) <80\% (OR 1.60 and 1.61, respectively). CONCLUSIONS: Cystic fibrosis increases the risk of developing anxiety and depression in female	**2**

	patients and in mothers. [...]	
Renal Cell Carcinoma	RESULTS: A total of 888 incident RCCs and 356 RCC deaths were identified. In models including adjustment for body mass index and energy intake, there was no higher risk of incident RCC associated with consumption of juices (HR per 100 g/day increment = 1.03; 95\% CI, 0.97-1.09), total soft drinks (HR = 1.01; 95\% CI, 0.98-1.05), [...] CONCLUSIONS: Consumption of juices or soft drinks was not associated with RCC incidence or mortality after adjusting for obesity.	**3**

Table 5 Examples for automatic identification of risk factors in medical abstracts, marked by the annotator

- 1 (valid risk factor for the specified disease) -- stomach diseases are risk factors for CML.
- 2 (valid risk factor for a different disease) -- CF, the disease of interest, was found to be a risk factor for depression and anxiety.
- 3 (not a risk factor) -- juices were **not** identified as a risk factor for RCC.

We attribute most factors erroneously annotated with type 3 annotation --- not a risk factor --- to cases where the QA model was required to extract a risk factor from an abstracts that does not contain one. Since the model was trained (and fine-tuned) to *always* identify an answer span for a given context and question, it is expected to yield (admittedly) weak performance on a context lacking the factors at the first place. Notably, a relatively small amount of all manually evaluated examples (around 8.5%) fall into this category.

The next table further summarizes the evaluation results by disease family. The prevalence of type 1 and 2 annotations illustrates the model's effectiveness in identifying risk factors, yet also underscores the challenges in achieving precise disease-specific accuracy. The presence of type 3 annotations, although significantly lower, highlights the ongoing need for the classification model refinement to enhance both specificity and accuracy.

Family (sub-family)	(1) Valid risk factor for the specified disease	(2) Valid risk factor for a different disease	(3) Not a risk factor	Total in family
Carcinomas	317	285	60	662
Infection	45	51	6	102
Leukemias	208	192	46	446
Lymphomas	27	12	4	43
Metabolic disorders (GD)	4	60	8	72
Mucus malefunction (GD)	11	34	2	47
Cardiomyopathy	5	23	0	28
Sarcomas	15	5	1	21
Other hematological disorders	30	32	2	64
Total	662	694	129	1485

Table 6 Distribution of manual evaluation annotations by disease family

"GD" denotes "genetic disorder". Note the much high number of risk factors identified for common (and potentially fatal) diseases, due to the vast body of empirical literature. The numbers refer to the total number of (not necessarily unique) risk factors identified for a disease family. We hypothesize that abstracts concerning diseases with a significant, sometimes absolute, genetic component are less likely to address other contributing factors, between the dashed lines in the table.

Error Analysis

Additional observation can be made about error distribution between type 1 and 2 annotations within and across disease families. Evidently, while some disease families show a balanced ratio between type 1 and 2 annotations (e.g., Infection, Leukemias), others resulted in more mis-identified factors -- type 2 annotation (e.g., Metabolic disorders). We hypothesize that abstracts concerning diseases with a significant, sometimes absolute, genetic component are less likely to address other contributing factors. Consequently, research in this area predominantly focuses on stratifying potential risks for other diseases in individuals already affected by the genetic disorder.

5 Conclusion and Future Work

This study presented an approach to identifying and extracting disease risk factors from free-text medical articles using advanced natural language processing techniques, specifically leveraging the capabilities of the pre-trained BioBERT-based architecture. Our methodology involved a multi-step process, including the retrieval of relevant articles, binary classification to filter articles discussing risk factors, and a question-answering model to extract specific risk factor information.

We have demonstrated the potential of language technologies to significantly enhance the efficiency and effectiveness of risk factor identification in medical literature. Our contributions to this field are twofold: the presentation of an automated pipeline for risk factor extraction and the creation of valuable datasets for future research.

While our study marks an advancement in the automated extraction of risk factors from medical literature, several avenues remain for future research and development. Our future directions include introducing improvements to QA model's accuracy and specificity, integration of additional data sources, and evaluation of more advanced LLMs for the task of risk factors identification.

Furthermore, inspired by recent findings that automatic annotations generated by models like GPT-4 can achieve results comparable to human annotations, we plan to investigate the use of GPT-4 for the task of risk factors annotation, and compare its performance with human experts.

6 Ethical Considerations

We make use of publicly available data in the domain of healthcare, that have been broadly used in numerous studies. Manual annotations were conducted by one of the authors of the paper, with medical background. Due to the required expertise and the inherent difficulty of the task, the mean hourly rate for the annotator was much higher than the established minimum wage.

References

[1] Kronenberg, Florian and Mora, Samia and Stroes, Erik SG and Ference, Brian A and Arsenault, Benoit J and Berglund, Lars and Dweck, Marc R and Koschinsk, Marlys and Lambert, Gilles and Mach and Francois and others, "Lipoprotein (a) in atherosclerotic cardiovascular disease and aortic stenosis: a European Atherosclerosis Society consensus statement," *European heart journal,* vol. 43, pp. 3925-3946, 2022.

[2] Lee, Jinhyuk and Yoon, Wonjin and Kim, Sungdong and Kim, Donghyeon and Kim, Sunkyu and So, Chan Ho and Kang and Jaewoo, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, no. https://academic.oup.com/bioinformatics/article-abstract/36/4/1234/5566506, pp. 1234-1240, 2020.

[3] Chokwijitkul, Thanat and Nguyen, Anthony and Hassanzadeh, Hamed and Perez and Siegfried, "Identifying risk factors for heart disease in electronic medical records: A deep learning approach," *Proceedings of the BioNLP 2018 workshop,* no. https://aclanthology.org/W18-2303.pdf, pp. 18-27, 2018.

[4] Boytcheva, Svetla and Nikolova, Ivelina and Angelova, Galia and Angelov and Zhivko, "Identification of Risk Factors in Clinical Texts through Association Rules," *BiomedicalNLP@ RANLP,* no. https://www.acl-bg.org/proceedings/2017/RANLP_W4%202017/pdf/BioNLP009.pdf, pp. 64-72, 2017.

[5] Stubbs, Amber and Kotfila, Christopher and Xu, Hua and Uzuner and Ozlem, "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2," *Journal of biomedical informatics,* vol. 58, pp. 67-77, 2015.

[6] Sheikhalishahi, Seyedmostafa and Miotto, Riccardo and Dudley, Joel T and Lavelli, Alberto and Rinaldi, Fabio and Osmani and Venet and others, "Natural language processing of clinical notes on chronic diseases: systematic review," vol. 7, no. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6528438, 2019.

[7] Sabra, Susan and Mahmood, Khalid and Alobaidi and Mazen, "A semantic extraction and sentimental assessment of risk factors (sesarf): an NLP approach for precision medicine: a medical decision support tool for early diagnosis from clinical notes," no. https://ieeexplore.ieee.org/abstract/document/8029906, pp. 131-136, 2017.

[8] Abdelhamid, Abdelaziz A and Eid, Marwa M and Abotaleb, Mostafa and Towfek and SK and others, "Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques," *Journal of Artificial Intelligence and Metaheuristics,* vol. 4, pp. 45-53, 2023.

[9] Kavakiotis, Ioannis and Tsave, Olga and Salifoglou, Athanasios and Maglaveras, Nicos and Vlahavas, Ioannis and Chouvarda and Ioanna, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal,* vol. 15, pp. 104-116, 2017.

[10] Mehmood, Awais and Iqbal, Munwar and Mehmood, Zahid and Irtaza, Aun and Nawaz, Marriam and Nazir, Tahira and Masood and Momina, "Prediction of heart disease using deep convolutional neural networks," *Arabian Journal for Science and Engineering,* vol. 46, pp. 3409-3422, 2021.

[11] Naik, Aakanksha and Parasa, Sravanthi and Feldman, Sergey and Wang, Lucy Lu and Hope and Tom, "Literature-augmented clinical outcome prediction," *arXiv preprint,* 2021.

[12] Miyazawa, Yusuke and Katsuta, Narimasa and Nara, Tamaki and Nojiri, Shuko and Naito, Toshio and Hiki, Makoto and Ichikawa, Masako and Takeshita, Yoshihide and Kato, Tadafumi and Okumura and Manabu and others, "Identification of risk factors for the onset of delirium associated with COVID-19 by mining nursing records," *Plos one,* vol. 19, 2024.

[13] Roitero, Kevin and Portelli, Beatrice and Popescu, Mihai Horia and Della Mea and Vincenzo, "DiLBERT: Cheap embeddings for disease related medical NLP," *IEEE Access,* vol. 9, pp. 159714-159723, 2021.

[14] Yang, Xi and Chen, Aokun and PourNejatian, Nima and Shin, Hoo Chang and Smith, Kaleb E and Parisien, Christopher and Compas, Colin and Martin, Cheryl and Costa, Anthony B and Flores and Mona G and others, "A large language model for electronic health records," *NPJ digital medicine,* vol. 5, p. 194, 2022.

[15] Singhal, Karan and Tu, Tao and Gottweis, Juraj and Sayres, Rory and Wulczyn, Ellery and Hou, Le and Clark, Kevin and Pfohl, Stephen and Cole-Lewis, Heather and Neal and Darlene and others, "Towards expert-level medical question answering with large language models," *arXiv preprint,* 2023.

[16] H. Kilicoglu, G. Rosemblat, M. Fiszman and D. Shin, "Broad-coverage biomedical relation extraction with SemRep," *BMC Bioinformatics,* vol. 21, pp. 1-28, 2020.

[17] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat and T. C. Rindflesch, "SemMedDB: a PubMed-scale repository of biomedical semantic predications," *Bioinformatics,* vol. 28, pp. 3158-3160, 2012.

[18] G. Hong, Y. Kim, Y. Choi and M. Song, "BioPREP: deep learning-based predicate classification with SemMedDB," *Journal of Biomedical Informatics,* vol. 122, 2021.

[19] L. Luo, P.-T. Lai, Chih-Hsuan Wei, Cecilia N Arighi and Z. Lu, "BioRED: a rich biomedical relation extraction dataset," *Briefings in Bioinformatics,* vol. 23, 2022.

[20] Chen, Hao and Sharp and Burt M, "Content-rich biological network constructed by mining PubMed abstracts," *BMC bioinformatics,* vol. 5, pp. 1-13, 2004.

[21] David, Mary Rajathei and Samuel and Selvaraj, "Clustering of PubMed abstracts using nearer terms of the domain," *Bioinformation,* vol. 8, p. 20, 2012.

[22] Vinkers, Christiaan H and Tijdink, Joeri K and Otte and Willem M, "Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis," *Bmj,* vol. 351, 2015.

7 תקציר

מחקר זה מציג גישה לאוטומציה של זיהוי גורמי סיכון למחלות מתוך ספרות רפואית. תוך ניצול ההתקדמות בלמידת מכונה ועיבוד שפה טבעית, ובמיוחד ניצול מודלים מוכנים מראש בתחום הביו-רפואי תוך כוונון שלהם למשימה הספציפית - מחקר זה שואף לפתח מערכת מקיפה שתבצע באופן אוטומטי את החילוץ של גורמי סיכון ממגוון רחב של מאמרים רפואיים. המערכת מזהה מאמרים רלוונטיים, מסווגת אותם לפי נוכחות של דיונים על גורמי סיכון, ומחלצת מידע ספציפי על גורמי סיכון באמצעות מודל של QA שאלות-תשובות.

זיהוי גורמי סיכון למחלות הוא קריטי ברפואה מונעת, הוא מאפשר לאנשי מקצוע בתחום הבריאות לגבש אסטרטגיות מניעה יעילות ולשפר את תוצאות המטופלים. שיטות מסורתיות מסתמכות במידה רבה על סקירה ידנית של ספרות רפואית נרחבת - תהליך שהוא גם גוזל זמן רב וגם מצריך מאמץ רב. כלי אוטומטי שיכול לסרוק כמויות גדולות של ספרות מדעית ולזהות גורמי סיכון בולטים למחלות שונות אוטומטיים יכול לפשט משמעותית את התהליך הזה, לשפר את נגישות הידע ולהקל על השימוש היעיל במידע.

מחקר זה מתמקד בחילוץ גורמי סיכון במיוחד מתוך ספרות רפואית מדעית, בניגוד לניתוח של רשומות רפואיות אלקטרוניות. כאן, האתגר העיקרי הוא השונות והמבנה הלא מובנה של פרסומים רפואיים, בהם גורמי סיכון מתוארים בהקשרים ובפורמטים שונים. יתרה מזאת, הגילוי המתמשך של גורמי סיכון חדשים מצריך גישה דינמית שיכולה להתאים את עצמה לגוף הידע הרפואי המתפתח.

תחום המחקר של עבודת גמר זו מתמקד באופן ייחודי בחילוץ גורמי סיכון מתוך ספרות רפואית מדעית. תוך שימוש במודלים של שפה גדולים מוכנים מראש שהכנו במסגרת העבודה, המבוססים על BioBERT [2].

המחקר בונה מערכת רב-שלבית ייחודית שמזהה תחילה מאמרים רפואיים רלוונטיים, מסווגת אותם לפי נוכחות של גורמי סיכון באמצעות מודל Binary classification, ולאחר מכן מחלצת מידע ספציפי על גורמי סיכון באמצעות מודל של QA.

המכללה האקדמית תל-אביב

בית הספר למדעי החשב

מערכת אוטומטית לחילוץ גורמי סיכון ממאמרים רפואיים

חיבור זה הוגש כחלק מהדרישות לקבלת התואר “מוסמך " – M.Sc.

במכללה האקדמית תל-אביב

על ידי

מקסים רובצ׳ינסקי

העבודה הוכנה בהדרכתו של

ד"ר עדי שרייבמן

ד"ר דורית שוויקי

ד"ר אלה רבינוביץ'

יולי 2024