SimpleGermKG: Constructing a Gene-Disease Germline Knowledge Graph Using a Four-Stage Framework Based on BioBERT

By

Armando D. Diaz Gonzalez

Thesis

Submitted to the Faculty of the

Graduate School of Charleston Southern University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Computer Science

April 2023

North Charleston, South Carolina

Songhui Yue, Committee Chair

Kevin S. Hughes

Sean T. Hayes

Acknowledgments

This work would not have been possible without the invaluable support of Dr. Kevin S. Hughes and the Medical University of South Carolina. Dr. Kevin S. Hughes provided insights, feedback, and project data to help shape my research ideas and methodologies. His guidance and expertise in Cancer Genetics were instrumental in the success of my research project.

Dr. Songhui Yue, my research advisor, guided and supported me through the development of my research. His extensive experience in the field of Machine Learning and writing research helped me to successfully publish our work at the conference ICISDM 2023.

The contribution of Dr. Sean Hayes to my research project cannot be overstated. Through his feedback, critique, and guidance, I learned how to deliver writing of exceptional quality that met high academic standards.

Lastly, I must express my gratitude to my family who supported me throughout my academic journey. My wife, in particular, was a pillar of strength, always offering constant support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

Published biomedical information has and continues to rapidly increase. The recent advancements in Natural Language Processing (NLP), have generated considerable interest in automating the extraction, normalization, and representation of biomedical knowledge about entities such as genes and diseases. Our study analyzes germline abstracts in the construction of knowledge graphs of the immense work that has been done in this area for genes and diseases. This paper presents SimpleGermKG, an automatic knowledge graph construction approach that connects germline genes and diseases. For the extraction of genes and diseases, we employ BioBERT, a pre-trained BERT model on biomedical corpora. We propose an ontology-based and rule-based algorithm to standardize and disambiguate medical terms. For semantic relationships between articles, genes, and diseases, we implemented a part-whole relation approach to connect each entity with its data source and visualize them in a graph-based knowledge representation. Lastly, we discuss the knowledge graph applications, limitations, and challenges to inspire the future research of germline corpora. Our knowledge graph contains 82 genes, 181 diseases, and 309,283 triples. Graph-based visualizations are used to show the results.

Keywords: entity recognition, BioBERT, semantic relation, knowledge graph, germline mutations

CHAPTER I

Introduction

Certain genes that a person is born with protect us from developing cancer. Cancer susceptibility has mutations (i.e., have a DNA change that prevents their normal function), creating a higher risk of developing cancer. Looking for which mutated genes increase the risk of which specific cancers is of great interest and is known as the gene-disease association. Extracting germline genes and diseases from biomedical corpora for representing knowledge encoded in a Knowledge Graph (KG) requires complex, expensive, and time-consuming methods. Biomedical publications are increasing rapidly. For example, using the same search criteria, a PubMed search for BRCA1 and BRCA2 in 2010, fetched 478 papers compared to 830 papers in 2021, a 57% increase in annual new papers in 11 years. Because the total number of publications on these genes now exceeds 12,300 and there are an estimated 22,287 genes in the human genome (Salzberg, 2018), the magnitude of this task becomes overwhelming. As a result, manual extraction is essentially impossible. Many computational approaches have been proposed to extract gene-disease association information from the biomedical literature accurately and efficiently. For instance, in the fields of pharmacy (Kim et al., 2019), medicine (Choi & Lee, 2021), and biology (Singh et al., 2021), machine learning and deep learning models have enabled biomedical text-mining tasks such as summarizing, extracting, and analyzing large corpora with varying degrees of success (Al-Garadi et al., 2022).

Natural Language Processing (NLP), a field of artificial intelligence, is used to perform tasks such as *Named Entity Recognition* (NER), *Named Entity Normalization* (NEN), and Relation Extraction (RE) (Alshaikhdeeb & Ahmad, 2016; Cariello et al., 2021; Luo et al., 2022; Noh & Kavuluru, 2021; X. Wang et al., 2009). NLP systems can analyze immense amounts of text-based data and determine the correct meaning of a word in a specific context to extract key facts and relationships. To address the problem of gene-disease associations in an article, NER can be used for extracting genes and diseases (as entities) from the biomedical corpora (Wu et al., 2019).

The most recent approach is driven by transformer-based models that were recently developed by Google (Vaswani et al., 2017) and can be used for carrying out various NLP tasks (Bhatnagar et al., 2022). This approach can be pre-trained on biomedical literature and is known to outperform pre-trained models, such as ELMo and BERT (Lee et al., 2019).

Unlike relational databases, graph databases provide unique abilities to manage n-th degree relationships among complex types of biomedical data (Q. Zhu et al., 2020). Knowledge Graphs (KGs) have proven to be effective in representing large-scale heterogeneous data and visualizing the nature of underlying relationships. KGs provide a model of relevant facts and contextualized answers to specific questions, so that they can then be used to extract and discover deeper and more subtle patterns (Al-Moslmi et al., 2020). For example, KGs are suitable for representing hierarchical data, such as genes, diseases, and relationships that are interconnected.

Furthermore, many studies focus on a particular segment in the three-stage life cycle of the knowledge graph construction process that includes NER, NEN, and RE or Semantic Relation. In this paper, we present SimpleGermKG, a gene-disease knowledge graph based on germline corpora. The germline genes and diseases are extracted from abstracts using BioBERT (Lee et al., 2019). To our knowledge, no study has been conducted to analyze germline abstracts in the construction of knowledge graphs. Therefore, we examine the knowledge graph life cycle based on a hybrid procedure between deep learning, ontology-based, and rule-based approaches beginning with the data pre-processing, knowledge graph construction part, and ending with a discussion of graph applications and visualizations for further analysis.

Our contributions are summarized as follows:

- We developed SimpleGermKG, an automatic knowledge graph construction approach that connects germline genes and diseases. It uses BioBERT to extract genes and diseases from biomedical texts. Two algorithms were designed to explore, find new terms, and identify the exact master term for genes and diseases using an ontology-

rule-based approach with regular expressions. A part-whole relation approach to connect these gene-disease pairs with their references.

- We automated the construction process of SimpleGermKG, which visually organizes genes and diseases from germline abstracts. SimpleGermKG will expedite searches for gene-disease associations with references.
- Our study summarized the latest approaches to building a knowledge graph in the biomedical domain. We categorize these approaches into two main groups (1) ontology-based and (2) automatic knowledge graph construction, which is based on machine learning algorithms.
- In Chapter V, we proposed three relationship approaches for classifying relationships between germline genes and diseases. Two of them are based on a co-occurrence method, which indicates that there is a possible relationship between two entities when these appear in the text. The last approach could be used to find more granular relationships using a pre-trained language model such as BioBERT.
- The source code of our workflow is freely available at https://github.com/arm-diaz/SimpleGermKG.

The structure of this paper is as follows: Chapter II presents an overview of relevant approaches to the biomedical knowledge graph life cycle in previous studies. Chapter III explains a general description of the proposed methodological approach. Chapter IV describes details of the developed workflow, and the case study results. Chapter V discusses future work. Finally, the conclusions are highlighted in Chapter VI.

CHAPTER II

Background

II.I Germline Mutations

Genes are responsible for making proteins and act as instructions to determine features or characteristics that are passed from parent to child. Each cell in the human body contains about 22,000 genes. Proteins are building blocks and engines (Enzymes) that have specific functions in each part of the body. Hearts, arteries, lungs, skeletons, muscles, joints, hairs, teeth, and brains are all made up and managed by proteins (Truscott et al., 2016). Those proteins promote the growth and development of our body, help us repair cells and make new ones, and keep us healthy. Cancers start when one or more genes in a cell mutate. A gene mutation is a permanent change in the DNA that creates an abnormal protein or stops a protein's formation. Mutations can multiply uncontrollably and become dangerous. Some mutations can also cause genetic disorders or illnesses.

Human mutations can be classified into two basic types: germline and somatic. Germline mutations occur in sperm, eggs, and their progenitor cells and are therefore heritable (Meyerson et al., 2020). For example, germline mutations in the BRCA1 gene increase the risk of developing hereditary breast cancer. They also increase the risk of ovarian cancer in women and prostate cancer in men. Somatic mutations occur in other cell types and thus cannot be inherited by offspring (Meyerson et al., 2020). For instance, a breast cell may develop a mutation that causes it to divide out of control and form cancer. This is called sporadic cancer and the mutation is not passed to other family members. Gene changes that take place in breast cells may result from factors such as ultraviolet radiation, viruses, age, etc. (Thomson et al., 2014).

Research initiatives such as The Cancer Genome Atlas (TCGA) have generated 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data (Bell et al., 2011). Data has already led to improvements in the ability of physicians to diagnose, treat, reduce risk, target therapy, and prevent cancer. How- ever, researchers continue to study how genetic changes affect cancer development. Understanding the role that genes play in cancer is complicated because cancers involve multiple gene mutations. Advanced computational data analysis approaches, such as deep learning models, have revolutionized the biomedical domain and simplified the process of text mining.

II.II Knowledge Graph

The concept of a knowledge graph-based representation (i.e., a semantic network) was introduced by Richens (Richens, 1956). Decades later, increasing computing power made possible the introduction of Google's Knowledge Graph (Knowledge Vault) (Singhal, 2012). Furthermore, Resource Description Framework (RDF) (E. Miller, 1998) and Web Ontology Language (WOL) (Mcguinness & Harmelen, 2004) became the centerpiece of the Semantic Web (Berners-Lee & Hendler, 2001; S. Wang et al., 2022). The Semantic Web can represent the network model of system interactions that researchers refer to as pathways. Such pathways can be illustrated through a graph of connections between nodes of information representing the participating entities (Splendiani et al., 2011). In 2007, the semantic web and knowledge model concepts were investigated for data integration by computational experimentation in the biology domain (Post et al., 2007). Typical applications in the life sciences include drug discovery, drug development, indication expansion of existing drugs, pharmacovigilance, and medical imaging (Milošević & Thielemann, 2023; S. Wang et al., 2022).

II.III Deep Transformer

The introduction of the transformer model, an attention-based architecture developed by Google (Vaswani et al., 2017), marked a revolutionary innovation in the natural language processing domain. Prior to its introduction, NLP models were primarily based on recurrent neural networks (RNN), convolutional neural networks (CNN), and long short-term memory networks (LSTM). Transformers demonstrated a significant improvement because they do not require sequences of data to be processed in any fixed order, whereas RNNs, CNNs, and LSTMs do (Peters et al., 2018). Therefore, transformers can process data in any order and enable training on larger amounts of data than was possible before their existence. As a result, state-of-the-art (SOTA) results on common NLP tasks have been achieved through the success of transformer-based architectures such as Google AI's BERT (Devlin et al., 2018) and Open Ai's GPT (Radford & Narasimhan, 2018). The creation of pre-trained models facilitated the training on massive amounts of language data before BERT release.

II.III.I BERT

In 2018, Google introduced BERT, which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations. BERT learns information from both the left and the right side of a token's context during the training phase and is pre-trained on a large corpus of unlabeled text, including the entire Wikipedia (2,500M words) and Book Corpus (800M words). Its pre-training serves as a base layer of "knowledge" that carries encoded language and context. From there, BERT can be fine-tuned to solve specific tasks based on the ever-growing body of searchable content (Sanad, 2019). This process is known as transfer learning. Transfer learning created breakthroughs that were presented on the universal language model fine-tuning (ULMFiT) by fast.ai (Howard & Ruder, 2018).

Furthermore, BERT has been extended for domain-specific tasks in NLP such as drug discovery, clinical trials, and pharmacovigilance (Bhatnagar et al., 2022). A domain-specific

language is often challenging for NLP models because of the significant differences in vocabulary compared to standard language corpora such as Wikipedia. To solve this problem, BERT models are pre-trained for domain-specific tasks.

II.III.II BioBERT

BioBERT (Lee et al., 2019) trains a BERT model over a corpus of biomedical research articles sourced from PubMed and PubMed Central (PMC). PubMed is a database of biomedical citations and abstracts, whereas PMC is an electronic archive of full-text journal articles. Their contributions were a biomedical language representation model that could manage tasks such as named entity recognition, relation extraction, and question and answering. As a result of having a pre-trained model that embraces biomedical domain corpora, researchers and practitioners can encapsulate biomedical terms that would be challenging for a general language model to comprehend. Table II.1 summarizes recent transformer-based models for biomedical NER.

Table II.1: Transformer-based Models in the Biomedical Domain

Model	**Pretrained on**	**Performance**	**Year**
Bio Discharge Summary BERT (Alsentzer et al., 2019)	MIMIC III discharge summaries	Outperforms BERT and BioBERT on named entity recognition and natural language inference	June 2019
SciBERT (Beltagy et al., 2019)	Semantic Scholar	Outperforms SOTA for named entity recognition, relation extraction, patient enrollment task	September 2019
BioBERT (Lee et al., 2019)	PubMed and PMC	Outperforms SOTA for named entity recognition, relation extraction, question answering	October 2020

Model	Pretrained on	Performance	Year
BioMed-RoBERTa (Gururangan et al., 2020)	Semantic Scholar	Outperforms RoBERTa on text classification, relation extraction and named entity recognition	May 2020
BioALBERT (Naseem et al., 2020)	PubMed, PMC, MIMIC III	Outperforms SOTA for named entity recognition, relation extraction, question answering, sentence similarity, document classification	September 2020
ChemBERTa (Chithrananda et al., 2020)	PubChem	Outperforms baseline on one task of molecular property prediction	October 2020
ClinicalBERT (Huang et al., 2019)	MIMIC III	Outperforms deep language model for clinical prediction	November 2020
SciFive (Phan et al., 2021)	PubMed and PMC	Outperforms SOTA for named entity recognition, relation extraction, question answering, document classification and inference	May 2021
BioBART (Yuan et al., 2022)	PubMed	Outperforms BART on dialogue, summarization, named entity recognition and entity linking	April 2022

II.IV Ontology-based Named Entity Normalization

The majority of biomedical literature incorporates nonstandard naming conventions, abbreviations, and punctuations, which makes the required information difficult to use and understand (A. M. Cohen & Hersh, 2005). An ontology can provide a unique identifier for

describing information for each entity and link named entities in the biomedical text to make sense of the identified named entities (K. B. Cohen et al., 2002). Ontologies can help researchers integrate and link concept labels to their interpretations, concept definitions, and relations to other concepts (Karadeniz & Ozgur, 2019; Spasic et al., 2005).

Named Entity Normalization (NEN) is the task of mapping each named entity mentioned in a given text to its corresponding entity given in an ontology/dictionary (Simpson & Demner-Fushman, 2012). The term “ambiguity” occurs when the same term is used to refer to multiple concepts. NEN aims to support consistent and unambiguous knowledge integration (Bratus et al., 2011). This task is also called entity grounding, entity linking, or entity categorization. For instance, the diseases *Cancer of Pancreas*, *Carcinoma of pancreas*, and *Malignancy of Pancreas* can be normalized to the ontology concept term “Pancreatic Cancer”.

CHAPTER III

Related Work

Biomedical knowledge graphs may be constructed using various techniques which begin with large datasets that are extracted from pre-existing databases or texts. These pre-existing databases were created by domain experts using manually curated KGs and automatically extracted KGs (e.g., using machine learning methods). Manual curation is a time-consuming process, due to the required effort by the domain expert to review papers, annotate phrases and sentences, and define rules and constraints that help users make inferences. On the other hand, machine learning approaches in natural language processing tasks can be used to automate the process of building a knowledge graph. NLP can quickly detect sentences of interest and unveil complex relationships among the data. These methods require annotating only a subset of the data.

III.I Ontology-based Knowledge Graph Construction

In medicine, biomedical knowledge can be divided into many subdomains, such as genes, chemical compounds, diseases, organs, symptoms, and syndromes. The purpose of biomedical ontology goes beyond collecting names of entities, a dictionary of terms, and controlled vocabulary for a variety of entities. It defines biological classes of entities and the relations among them for building a knowledge base (Bodenreider et al., 2005). A well-defined ontology is essential for the creation of a biomedical knowledge graph because the ontology enables complex reasoning about biomedical knowledge.

BioPortal (Whetzel et al., 2011), an open repository of biomedical ontologies, has more than 1000 ontologies and 15 million classes of entities. These ontologies have been designed and developed by the community of research teams to summarize and organize information.

Maintaining an ontology life cycle is infeasible for a human expert, see Figure II.1, since it is expensive and time-consuming. In addition, the difficulty is compounded by the fact that scalable ontologies require reusing parts of other ontologies and applying automated quality control testing that guarantees best practices for software development (Matentzoglu et al., 2022). In the following subsections, we proceed with a detailed description of three recent studies on knowledge graphs for genes and diseases derived from biomedical ontologies.

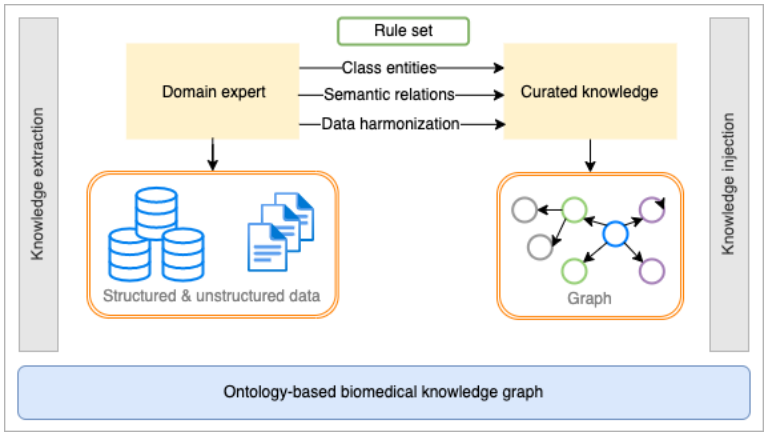


Figure II.1 Ontology-based Biomedical Knowledge. An overall methodology and architectural process for ontology-based biomedical knowledge.

III.II.I Genetic and Rare Diseases Information Center (GARD).

GARD (Hoskins, 2022) is currently managed by the Office of Rare Diseases Research (ORDR) within the National Center for Advancing Translational Sciences (NCATS). It was established to provide freely accessible consumer health information on over 6,500 genetic and rare diseases. The dataset includes curated disease information comprised of 15 different sections, such as summary, diagnosis, inheritance, and others. Through automated and manual mapping processes, a meta-ontology was pre-defined to capture and represent semantic relationships among different types of data. The knowledge graph includes a total number of 3,819,623 nodes and 84,223,681 relations from 34 different biomedical data resources.

III.II.II GenomicKB

GenomicKB (Feng et al., 2022) integrates over 30 well-established data sources, including GENCODE, the Eukaryotic Promoter Database (EPD), dbSuper, RNAcentral, Genotype-Tissue Expression (GTEx), and others. These ontologies provide different insights regarding human genome. The knowledge graph contains 347,378,103 nodes, 1,359,209,258 edges, and 3,902,460,300 node/edge properties. GenomicKB can be accessed through a web portal that supports customized queries of diverse entities, relations, and properties. With this portal, GenomicKB can answer human genomics-related questions and conduct multi-modal analysis with coding-free and interactive queries.

III.II.III The integrative Biomedical Knowledge Hub (iBKH)

iBKH (Su et al., 2022), a comprehensive biomedical knowledge graph, integrates data from 18 publicly available biomedical knowledge sources. The knowledge graph contains entity types that are commonly studied in biomedicine, such as genes, diseases, drugs, pathways, and a dietary supplement knowledge base (iDISK). Data sources include biomedical ontologies such as the Brenda Tissue Ontology, the Cell Ontology, the Disease Ontology, and the Uberon as well as manually curated biomedical knowledge bases such as the Bgee, the Comparative Toxicogenomics Database (CTD), the DrugBank, the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Pharmacogenetics Knowledge Base (PharmGKB), the Reactome, the Side effect resource (SIDER), and the TISSUE. iBKH contains a total of 2,384,501 entities of 11 types, including 23,003 anatomy entities, 19,236 disease entities, 37,997 drug entities, 88,376 gene entities, 2,065,015 molecule entities, 1,361 symptom entities, 2,988 pathway entities, 4,251 side-effect entities, 4,101 dietary supplement ingredient (DSI) entities, 137,568 dietary supplement product (DSP) entities, 605 dietary’s therapeutic class (TC) entities. In addition, there are 45 relation types within 18 kinds of entity pairs, which means multiple types of relations can exist between a pair of biomedical entities.

III.III Automatic Knowledge Graph Construction

Managing the increased rate of publications via manual curation is infeasible, requiring approaches that can automate part or all of the process. Natural Language Processing is commonly used to extract entities and their relations from biomedical text. Therefore, NLP can facilitate and automate knowledge graph construction. NER and NEN approaches have been developed to find relevant entities and connect these entities to meet the agreed data model (Milošević & Thielemann, 2023). Biomedical named entity recognition and named entity resolution techniques have been studied since the late 1990s (Fukuda et al., 1998; Hogan et al., 2021), and different approaches have been proposed and developed to solve NER systems. These approaches can be classified into (1) rule-based, which relies on linguistic experts designing accurate rules, (2) machine learning-based, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF), (3) deep learning-based such as RNN, LSTM, CNN, and pre-trained language models, and (4) hybrid approaches (J. Li et al., 2018; Pawar et al., 2017; J. Yang et al., 2021; Devlin et al., 2018).

We designed a taxonomy that illustrates key tasks and aspects of the process to automate the construction of biomedical knowledge graphs. Figure III.2 shows the taxonomy that was designed after examining recent KG creation approaches. This taxonomy aims to represent an overall architectural process of constructing a biomedical KG based on NLP-based algorithms discussed in this work (Figure III.3). The first step depicts a domain expert who is in charge of guiding and proving insights into the whole process, then knowledge extraction begins with an entity-level which describes the process of extracting key concepts from the text, normalization-level seeks to group terms with similar meanings into one unique term, relation-level aims to find associations between two entities, and after following the previous steps the results go through a data integration process to finally be stored as a knowledge graph.

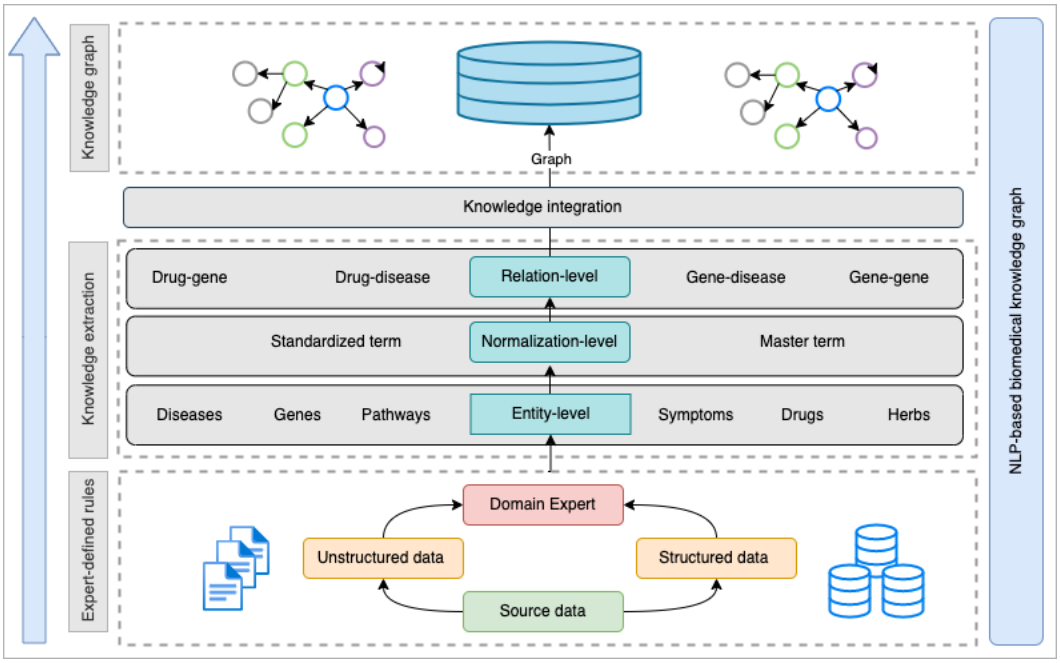


Figure III.2 Automatic Biomedical Knowledge Graph: A taxonomy of automatic biomedical.

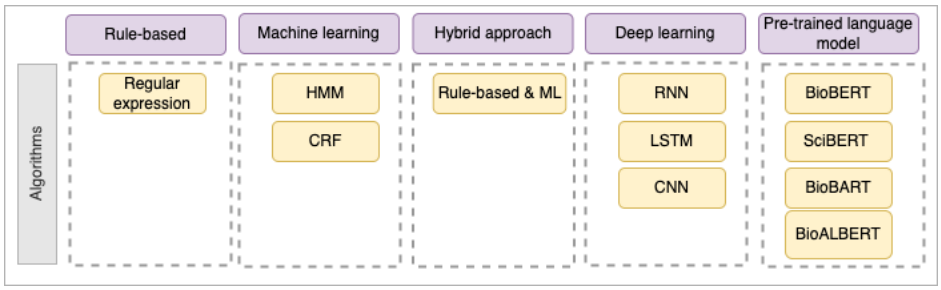


Figure III.3 NLP-based Algorithms to Automate Biomedical Knowledge Graph Construction.

Due to significant advances in deep learning, pre-training allows the model to incorporate domain-specific knowledge, which can further improve the ability of the pre-trained base model to achieve high performance on various tasks. The model can be fine-tuned on task-specific datasets to perform tasks such as named entity recognition and relation extraction, which are two critical tasks to construct domain-specific knowledge graphs. In contrast with a manual curation approach, pre-trained models (PTMs) can reduce computing costs, and save time and resources. In summary, PTMs can take advantage of large-scale corpora to learn a universal or domain-

specific language, better performance and speeds up convergence on the target task, as well as avoid overfitting on small datasets.

BERT-based models can be used to generate scalable knowledge graphs from new corpora that include undiscovered knowledge. Recent studies have attempted to apply this method to build knowledge that models genes, diseases, and their interactions from different perspectives. We summarize these studies that provide significant achievements for building a KG as follows:

III.III.I Herbal-molecular Medicine Knowledge Graph (HerbKG)

HerbKG (X. Zhu et al., 2022), a knowledge graph that bridges herbal and molecular medicine, includes bio-entities such as herbs, chemicals extracted from the herbs, genes that are affected by the chemicals, and diseases treated by herbs due to the functions of genes. The framework uses abstracts, chosen by a domain expert in herbal medicine, that go through the PubTator Central (PTC) NER model which provides automated annotations for text mining systems, followed by a custom BERT-based RE model to produce a list of identified relation triplets, which are used for the HerbKG construction. The constructed HerbKG supports multiple downstream applications, such as descriptive analysis, evidence-based graph query, similarity analysis, and drug repurposing. The proposed system analyzed a total of 516,393 PubMed abstracts and identified 4,130 herbs, 6,331 chemicals, 2,187 diseases, and 2,641 genes, with 53,754 distinct relations, including 19,872 HerbHasCompoundChemical (HHC), 13,627 HerbTreatsDisease (HTD), 9,984 ChemicalActsOnDisease (CAD), 3,353 ChemicalAssociatesGene (CAG), and 6,918 GeneInfluencesDisease (GID) relations.

III.III.II Biological Knowledge Graph Relating Genes, Diseases, and Drugs

A knowledge graph (Milošević & Thielemann, 2023), based on several rule-based and pre-trained language approaches, is proposed. The framework focuses on relationship normalization methods between drug, gene, and disease entities. The three pairs of relationships of interest are

Drug-Gene, Drug-Disease, and GeneDisease. Relationship types were initially modeled as a sentence-level classification task followed by traditional machine learning algorithms, such as Random Forests, and Naive Bayes. Then pre-trained language models such as DistilBERT, PubMedBERT, text-to-text transformer (T5), and SciFive were fine-tuned for sentence-level relationship type classification. Extracted relationships were loaded into a graph database to answer complex medical questions with evidence. For example, it can retrieve genes interacting with a given, drugs that have an effect on a certain disease, or genes that are important for a given disease. The model processed about 35 million abstracts and managed to extract a total of 4,784,985 relationships (with co-occurrences of 35,900,521). There were 631,573 named relationships found between Drug-Genes (6,885,810 including co-occurrences), 1,468,639 relationships between Drug-Diseases (8,378,599 including co-occurrences), and 2,684,742 relationships between Genes and Diseases (20,065,385 including co-occurrences).

III.III.III Stroke-related Knowledge Graph (StrokeKG)

StrokeKG (X. Yang et al., 2021), a stroke-related knowledge graph, combined information extracted from these scientific papers and existing knowledge bases. A pretrained BiLSTM-CRF model with the Plant-disease corpus was used to build a NER classifier to identify Herbs. DNorm method was applied to extract and normalize disease words, tmChem was used as a chemical named entity identifier, GNormPlus handled gene mentions and PWTEES for pathways detection. The knowledge graph contains nine named-entity types: diseases, drugs, genes, symptoms, pathways, Chinese Patent Medicines (CPMs), herbs, chemicals, and ingredients. BioBERT and a rule-based method were used as a sentence-level relationship type classification to indicate when there is a possible relationship between two entities. As a result, StrokeKG contains a total of 46,983 entities belonging to 9 entity types. The type-wise distribution of the entities. StrokeKG contains a total of 157,302 triplets belonging to 30 edge types with 659,838 properties.

CHAPTER IV

Method

We propose a four-stage pipeline, which constructs SimpleGermKG. First, we proceed with a detailed description of the dictionaries used for the NEN task. Then, we describe the workflow that consists of tokenizing and preparing the dataset for machine learning, extracting genes and diseases from germline corpora using BioBERT NER, standardizing entities through a named entity normalization process, and linking normalized entities using the semantic relation that associates entities with their PubMed ID as illustrated in Figure IV.4.

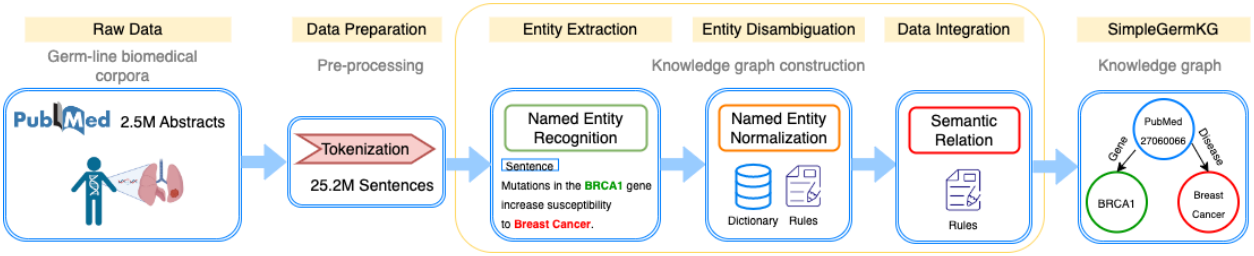


Figure IV.4 SimpleGermKG Architecture. This figure illustrates the overall workflow of SimpleGermKG.

IV.I Data Sources

Due to the complexity of properly defining and categorizing a large number of biomedical terms, we rely on two home-grown dictionaries (MUSC), one includes a list of diseases, and the other a list of genes. These dictionaries are relevant for mapping genes and diseases to a master term. The dictionary of diseases contains around 189 disease names and 1,567 synonyms, and the dictionary of genes contains around 84 gene names and 1,343 synonyms.

IV.II Pre-processing

Tokenization is the process of breaking down unstructured data and natural language text into smaller units of information (Webster & Kit, 1992). For instance, sentences, punctuation marks, words, and numbers can be considered tokens. Large input sizes for machine learning models

are not recommended, especially BERT-based models (Devlin et al., 2018), which have an input size restriction of 512 characters. Considering an abstract of a research paper is usually a paragraph of 300 words or less, an abstract may have over 512 characters with spaces included in the character count. To solve this problem, we used the PunktSentenceTokenizer (Marcus et al., 1993) method from the NLTK python library, which is trained on the Penn Treebank corpus and uses regular expressions to parse sentences and detect the sentence boundaries as shown in Figure IV.5.

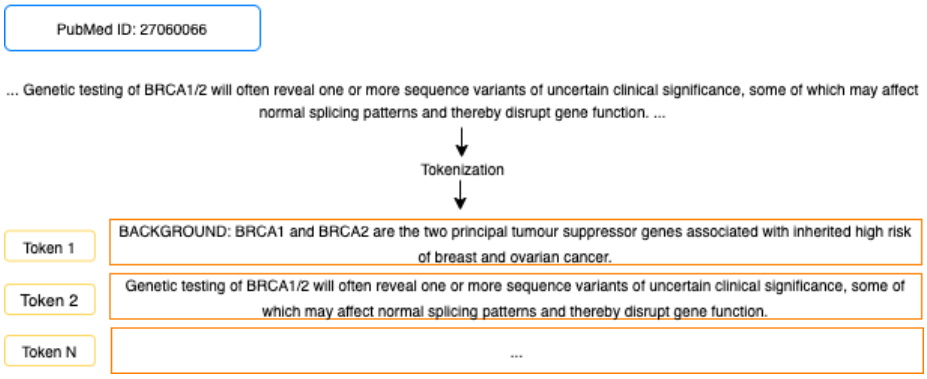


Figure IV.5 Tokenization Process. Break down text into sentences. Example with PubMed ID 27060066.

IV.III Named Entity Recognition.

We used a fine-tuned BioBERT model pre-trained on NCBI-disease corpus (Abreu Vicente, 2022b) to extract diseases from germline abstracts. The Natural Center for Biotechnology Information Disease (NCBI) (Doğan et al., 2014) disease corpus is a collection of 793 PubMed articles with 6,892 manually annotated disease mentions. For extracting genes from germline abstracts, we used a fine-tuned BioBERT model pre-trained on the BC2GM corpus (Abreu Vicente, 2022a). The BioCreative II Gene Mention (BC2GM) corpus (Smith et al., 2008) consists of sentences from PubMed abstracts with manually labeled gene and alternative gene entities.

IV.IV Named Entity Normalization

After extracting single and multi-word phrases from texts, Named Entity Normalization (Cho et al., 2017) is performed, which allocates suitable tags to the recognized entities. In biomedical articles, named entity normalization is a challenging task because biological terms, such as genes and diseases have multiple synonyms, and term variations, and are often referred to using abbreviations (Leaman et al., 2015). To resolve these ambiguities, machine-learning approaches have been investigated (Neves et al., 2010). However, many normalization tools rely on domain-specific ontologies, dictionaries, or rules. Domain-specific dictionaries can differentiate between synonyms, abbreviations, and punctuation marks. For instance, Hakenberg et al. (2011) and Wei & Kao (2011) implemented a dictionary-based approach to detect and normalize genes and proteins and normalize chemical names (Rocktäschel et al., 2012).

Our approach is divided into two steps. First, we used a dictionary-lookup approach using our two manually curated dictionaries and an approximate string-matching algorithm. Algorithm 1 maps *genes* and *diseases* mentioned from articles to a specific identifier/mater term. A successful match with a master term provides a numerical score that is used as the confidence level. Higher scores indicate higher certainty that the master terms matched to genes and diseases may be correct. Lower scores should only be considered for exploratory analysis purposes and collect those terms that are not available in the manually curated dictionaries to improve the confidence level.

To reduce the complexity of our Algorithm 1, we assumed that the BioBERT NER task classifies a single entity following the format BIO encoding scheme to represent the tokens and each entity that may be composed of several words must be unique. For example, the input sequence *“BRCA1 and BRCA2”* should be classified with the labels *B-GENE, I-GENE,* and *B-GENE*. This assumption allows us to map only one master term from the dictionary for each entity. However,

it is possible that an entity may be mapped to more than one master term because the BioBERT NER task may classify the previous example with the labels *B-GENE, I-GENE,* and *I-GENE*.

Algorithm 1: Exploratory Named Entity Normalization

Input: Ontology Set, O = {O_1, O_2, ..., O_n}

Named Entity Set, E = {E_1, E_2, ..., E_n}

Output: Best Match Entity Set, BM = {BM_1, BM_2, ..., BM_n}.

Best Match Score Set, S = {S_1, S_2, ..., S_n}.

$O\ (i, j)$ ←Ontology Set; /* Ontology of Genes and Diseases */

$E\ (i, j)$ ←Named Entity Set; /* Named Entity */

$\mathrm{BM}\ (i, j)$ ←Ø; /* Best Match Entity */

$\mathrm{S}\ (i, j)$ ←Ø; /* Score from 0 to 100*/

Condition: E, BM and S have the same size.

for all elements $(index, entity)$ in $E\ (i, j)$ do

if $entity$ in O then

$\mathrm{BM}\ (index, entity) \leftarrow O\ (entity)$;

$\mathrm{S}\ (index, entity) \leftarrow 100$;

else if StringMatch $(entity, O)$ then

$\mathrm{BM}\ (index, entity) \leftarrow StringMatch(entity, O)$ [Best Match];

$\mathrm{S}\ (index, entity) \leftarrow StringMatch(entity, O)$ [Score];

else

$\mathrm{BM}\ (index, entity) \leftarrow \emptyset$;

$\mathrm{S}\ (index, entity) \leftarrow \emptyset$;

```
        end
end
```

Line 1 of Algorithm 1 starts a loop that iterates over each element, which was found by the BioBERT NER task, in the Named Entity Set (*E)*. Then, line 2 checks whether the entity is present in the dictionaries. When the entity is in the dictionaries, the next line sets the corresponding element in the Best Match Entity Set (B*M*) that contains the dictionary master term that matches the entity and stores a score of 100, indicating a perfect match. If the entity is not in the dictionaries, line 5 calls a string match function that checks whether there is a string match between an entity and any master term. If there is a string match, line 6 sets the corresponding element in the Best Match Entity Set (B*M*) to the master term that best matches the entity (using a string-matching algorithm) and sets the confidence level score between the entity and its best matching master term. Finally, if the entity is not in the dictionary and there is no string match, we assign an empty set in the Best Match Entity Set (B*M)* and Best Match Score Set (*S)*.

As Algorithm 1 is used to explore, identify, and add new terms to our dictionaries, we implemented a second algorithm based on regular expressions generated from the dictionaries. Algorithm 2 aims to find the exact match between an entity and a master term. The algorithm loops through each element in the Named Entity Set (*E*) and for each element, it loops through each regular expression in the hash table *R*. If the regular expression matches the entity, it assigns the corresponding master term to the exact match entity in the Exact Match Entity Set (*EM*). If there is no match, the Exact Match Entity Set (*EM*) is set to an empty value. The *'break'* statement is used to exit the inner loop as soon as a match is found, because each entity can only have one match.

Algorithm 2: Entity Mapper

Input: Regex Set, R = {R_1, R_2, ..., R_n}

Named Entity Set, E = {E_1, E_2, ..., E_n}

Output: Exact Match Entity Set, EM = {EM_1, EM_2, ..., EM_n}.

R (master entity, regex) ←Regex Set; /* Regular Expression of Master Genes and Diseases */

E (index, entity) ←Named Entity Set; /* Named Entity */

EM (index, best match) ←Ø; /* Exact Match */

Condition: E and EM have the same size.

for all elements (eIndex, $entity$) in E do

for all elements (master entity, regex) in R do

if length (ApplyRegex (regex, $entity$)) > 0 then

exactMatch = master entity

EM (e$Index$, best match) ← exactMatch;

break

else

EM (e$Index$, best match) ← ∅;

end

end

end

IV.V Semantic Relation

Given a pair of entities, such as a gene and disease, a semantic relation consists of identifying the relation type between them. An important semantic relation for many applications is the part-

whole relation (Girju et al., 2006). Let us notate the part–whole relation as PART (<Tail Entity>, <Head Entity>), where <Tail Entity> is part of <Head Entity>. For example, the phrase “genes are found on tiny structures called chromosomes” contains the part-whole relation PART (genes, chromosomes). Winston, Chaffin, and Hermann (Simons, 1987) determined six types of part–whole relations: (1) component–integral object, (2) member–collection, (3) portion–mass, (4) stuff–object, (5) feature–activity, and (6) place–area. More recent studies, such as the SemEval 2018 Task 7 (Gábor et al., 2018), proposed a task on semantic relation extraction and classification in scientific paper abstracts that are practical for working on extracting specialized knowledge from domain corpora, such as biomedical information extraction.

Successful entity-relation linking requires detecting both the entity mentions in the abstracts, along with their respective entity types from the gene-disease dictionaries, and determining the type of relationship that exists between them. Based on psycholinguistic experiments and how the entities contribute to the structure of the part-whole relationship, we determined that the part-whole relationship from SemEval 2018 Task 7 can help us better identify and connect our entities to build the knowledge graph. SemEval 2018 Task 7 provides three comprehensive sets of classification rules, (1) composed of, (2) data source, and (3) phenomenon (Gábor et al., 2018).

Our main dataset contains germline abstracts and their PubMed ID. An abstract can include more than one gene and disease mentioned per sentence. Because germline mutations are passed on from parents to offspring, it is complicated to establish causal relationships in the germline association between genes and diseases, and thus they are not well-defined in the literature (Bonifaci et al., 2010). Therefore, we use a data source relationship in the form of PUBMED_ID-GENES_IN-GENE and PUBMED_ID-DISEASES_IN-DISEASE. Our approach matches all given genes and diseases to their given PubMed ID.

IV.VI Graph Database

The knowledge graph can be stored in either a relational database or a graph database. Relational databases may present problems such as the high overhead of complex queries and data redundancy. Graph databases, as opposed to traditional relational databases, can effectively manage, and execute complex queries. We used Neo4j, a graph database platform that can connect bodies of text and establish how they relate to each other, for the construction of the graph database and Cypher query language that uses graph pattern matching to read data and perform analysis of our results. Other studies have used Neo4j to create a graph representation from biomedical texts (Dörpinghaus et al., 2020; Gratzl et al., 2015; Himmelstein et al., 2017; Q. Zhu et al., 2020).

CHAPTER V

Results

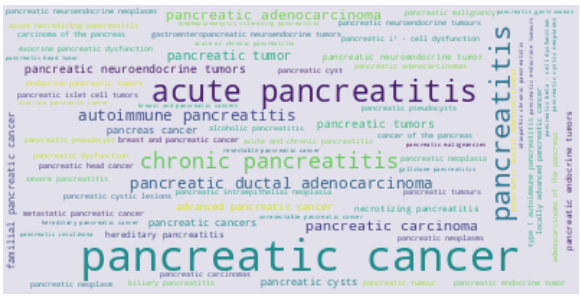
V.I Knowledge Graph Construction

Our experiments are conducted on germline corpora, which contain 2,556,715 abstracts from PubMed, and 25,225,925 sentences after tokenization. We used Algorithm 1 to explore genes and diseases within the abstracts. By conducting an exploratory data analysis, we were able to detect patterns among entities that exhibited a high matching score with our dictionaries. After that, it was observed that most of the gene and disease entities were synonyms and therefore belonged to the same entity. To eliminate the issue of data ambiguity, we applied an *ad hoc* mapping technique using regular expressions (Algorithm 2). Algorithm 2 standardized those entities that could be mapped to a dictionary of genes, diseases, and alternative names. To construct a germline-focused knowledge graph, we excluded entities that denoted objects not found in our dictionary of germline genes and diseases.

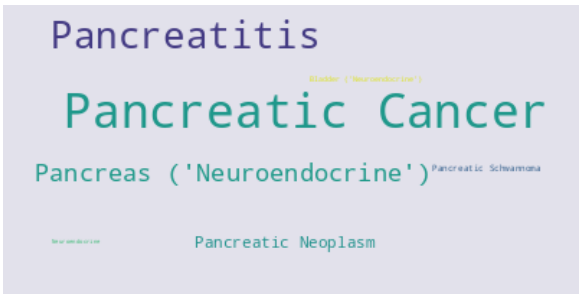
To summarize the results, BioBERT identified a total of 1,489,491 genes and 1,375,784 diseases. However, according to Algorithm 2 only 90 genes and 180 diseases were associated with master terms of germline mutations. Word-cloud visualizations were used to analyze both the entities extracted by BioBERT and the outcomes of the normalization procedure. Word clouds can help build familiarity with the content of a large collection of textual documents and identify their subject domains (Kalmukov, 2021). As depicted in Figure V.6, we limited the analysis to Pancreatic diseases which enabled us to showcase the effectiveness of our algorithms.

Then, we formally defined a semantic relation type to be a pair consisting of a domain class of type PART-DATASOURCE (Gene, PubMed ID) and PART-DATASOURCE (Disease, PubMed ID). We defined the semantic relation as “GENES_IN”, and “DISEASES_IN” to capture the connection between a PubMed ID and genes and/or diseases. Once we identified the

disambiguated entity types and relationships, we linked them together and built the knowledge graph.



(c) WordCloud NER Pancreatic diseases



(d) WordCloud NEN Pancreatic diseases

Figure V.6 Comparison Between Word Cloud NER and NEN Entities: Word cloud showing a comparison between the entities extracted through BioBERT NER task and normalized entities for Pancreatic diseases.

The knowledge graph was built with the Neo4j graph platform. As shown in Table V.2, we processed 2,556,715 abstracts, and SimpleGermKG contains 82,408 abstracts that mentioned at least one germline gene and disease. As a result, the generated output consists of 117,183 PubMed ID-gene relationships and 125,784 PubMed ID-disease relationships. By storing only the abstracts that mentioned at least one germline gene and disease, it will help us conduct additional analyses to infer associations between the genes and diseases discussed in the literature for future work.

Table V.2 SimpleGermKG summary: Gene and disease relationships with a PubMed ID.

Total # of abstracts	**# of abstracts with at least one gene and one disease**	**# of PubMed ID-gene relationships**	**# of PubMed ID-disease relationships**
2,556,715	82,408	117,183	125,784

V.II Evaluation

To validate the performance of SimpleGermKG, we compared the genes and diseases between annotated MUSC abstracts and SimpleGermKG. The evaluation of our framework was restricted to pancreatic diseases because reviewing a large number of abstracts manually would be unfeasible given the available time and resources. Our evaluation approach involved three steps: First, we identified abstracts that showed a perfect match between the annotated MUSC entities and SimpleGermKG. Then, we detected abstracts in which our framework captured genes and diseases that were not annotated by MUSC. Finally, we searched for abstracts in which our framework missed genes and diseases present in the MUSC annotations. As a result, SimpleGermKG achieved a 95% accuracy rate for genes and a 100% accuracy rate for Pancreatic diseases. Validation results are summarized in Table V.3.

Table V.3 Human validation between annotated MUSC abstracts and SimpleGermKG

Target	**Total # of abstracts**	**% of abstracts that matched the same genes and pancreatic diseases (similarity)**	**% of abstracts that detected new genes and pancreatic diseases (discovery)**	**% of abstracts missing genes and pancreatic diseases (error rate)**	**Accuracy (%) = similarity + discovery – error rate**
Genes	20,758	85%	10%	5%	95%
Pancreatic disease	20,758	49%	51%	0%	100%

V.III Knowledge Graph-based Visualization

So far, SimpleGermKG which covers germline abstracts from PubMed has been constructed with an integrated ontology of genes and diseases. We stored SimpleGermKG in the Neo4j graph

database, which allows researchers and clinicians to find relevant information and facilitates the navigation of biomedical data. To prove the visual management and ease-to-query of Neo4j, we show the search results of two queries through the Neo4j Cypher graph query language in Figure V.7. The blue node represents “PubMed ID” (abstract ID), the green one is “gene” (disambiguated gene names), and the red one is "disease" (disambiguated diseases) mentioned in the text. The “gene” and “disease nodes” can be also identified with the relationship (edge) “GENES_IN” and “DISEASES_IN” respectively. Figure V.7 shows articles that mentioned the disease "Pancreatic Cancer" and all gene entities mentioned in the abstracts.

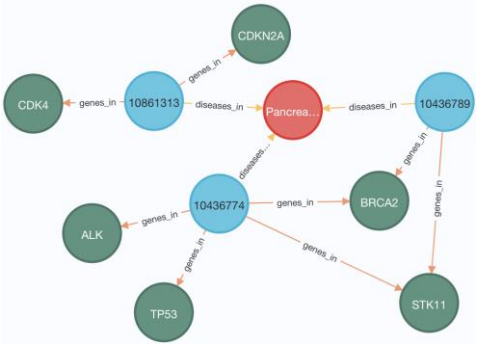


Figure V.7 Graph Representation of Gene-PubMed-disease Associations: PubMed IDs: “10436774, 10861313, 10436789” & Disease: “Pancreatic Cancer” & Genes.

V.IV Knowledge Graph Application

The proposed system re-organizes genes and diseases mentioned in the abstracts of germline papers and aims to serve as an educational resource and search engine to uncover implicit connections between heterogeneous data, which would facilitate further downstream applications, such as gene-disease associations, and question-answer systems.

CHAPTER VI

Limitations and Future Work

SimpleGermKG achieved a significant performance when compared with annotated MUSC genes and Pancreatic diseases. However, our study has the following limitations that will be addressed in future work. First, Algorithm 2 aims to map diseases mentioned in abstracts to a master term defined in the dictionaries provided by MUSC. This mapping only considers diseases and is excluding syndromes that have already been extracted by BioBERT. To solve this problem, we will add an in-house MUSC syndrome dictionary that will help us integrate new patterns and terms into our Algorithm 2. Lastly, we have limited the performance evaluation of SimpleGermKG on Pancreatic diseases. We will continue the validation process by expanding the analysis to include a wider range of genes and diseases.

Furthermore, the construction of knowledge graphs on germline corpora presents new opportunities as few studies focus on this domain. Some of the future activities on utilizing and improving our SimpleGermKG will involve:

- Developing an interface to enable a wider audience to use the SimpleGermKG. Neo4j platform can be queried via Cypher graph query language and driver APIs such as Python, Java, and JavaScript. These APIs are handy for developers but not medical practitioners who do not possess a large computational background. A graphical interface would allow non-engineers to evaluate hypotheses and analyze information stored in the graphs.
- Exploring possible applications and opportunities that could improve the lifestyle of individuals who present germline mutations. Germline mutations may affect people differently depending on genetic factors such as family background. These mutations may present a certain level of resistance to the effects of drugs. Therefore, it is important to explore opportunities in patients and identify possible risks, therapies, and clinical implications.

- Experimenting and exploring other approaches for the NER task. We aim to improve the precision of the gene-disease extraction by exploring pre-trained language models that have been fine-tuned on well-known gene and disease datasets in the literature. To measure the precision of the extracted entities, the first step is to locate tags for each term in the abstracts. It is necessary to use a tagging technique to label each token of the text and so recognize various tokens into named entities. The BIO encoding, which denotes Begin, Inside, and Outside, is the standard tagging and is more suited for the NER task. We recommend using web-based annotation tools such as ezTag (Kwon et al., 2018) and machine learning-based tagging approaches (Park et al., 2022) that can generate automatic tagging for a fine-tuning dataset in specific domains.
- Developing a transformer-based model pre-trained on germline corpora. Our work includes a dataset of abstracts from PubMed that have been categorized and verified by domain experts in germline corpora. We can take advantage of this dataset to create and train a new domain-specific model to solve tasks such as NER and RE. This process may require further annotations.
- Exploring a method of expanding our dictionaries for the NEN task. Larger gene-disease ontologies can be explored to enrich the vocabulary and improve the accuracy of the named entity normalization process. We can rely on other ontologies by combining concepts to generate a more complete vocabulary that includes more variations of the same terms from biomedical texts.
- Developing a technique for obtaining relationships from germline corpora. Due to the nature of germline mutations, conventional relation extraction techniques do not apply in the semantic relation of a germline corpus. Therefore, the relationship between genes and diseases needs to be defined for connecting the pathogenicity of germline variants in cancer. Training a model on germline corpora should consider the gene carrier probability

to select risk families for extracting relationships between cancer susceptibility genes. We propose three methods to identify the presence of associations between genes and diseases. The first two approaches are simple, and intuitive, and do not require annotations in the dataset. The BERT-based approach is more robust and promises more accurate relationships. Data must be pre-processed and annotated with the proposed relationships as opposed to the article-based and article-sentence-based approaches.

- Article level – Mentions of genes and diseases in the same article have a direct relationship with the PubMed ID. This approach cannot directly link a gene and disease. But we know that those entities have a contextual relationship within the text.
- Sentence level – In contrast to the article-based approach, we can assume a relationship between a gene and disease exists when those are mentioned in the same PubMed ID and sentence ID.
- BERT-based approach – A relation classification approach can extract sentences that contain the entity pair from the NER task which holds a semantic relation and then predict whether a certain relation exists between genes and diseases. We propose four cause-effect relationships where a gene may influence a certain degree of susceptibility to genetic disease.
    - Does not affect the disease - There is a clear indication that the gene does not represent any risk of causing disease.
    - Unlikely increases the risk of disease - There is no clear indication that gene does not affect the risk of causing disease.
    - May increase the risk of disease - There is no clear indication that gene increases the risk of causing disease.
    - Increases the risk of disease - There is a clear indication that gene is responsible for increasing the risk of causing disease.

CHAPTER VII

Conclusion

Digital biomedical information has been growing exponentially. To represent biomedical information effectively, we developed an automated knowledge graph construction framework, SimpleGermKG, to synthesize and store detailed information about germline genes and diseases associated with a PubMed ID. We employed BioBERT, a natural language processing model, to retrieve key information. A custom NEN algorithm based on regular expressions was designed to eliminate disambiguation. Our SimpleGermKG contains 90 genes, 180 diseases, and a total of 242, 967 relationships across 82,408 abstracts. The knowledge graph can store and represent medical knowledge from large biomedical corpora in such a way that researchers, students, and physicians can search, manage, share, and visualize.

SimpleGermKG has the potential to integrate information from electronic health records, genomic data, and other existing biomedical ontologies. The knowledge graph has already improved and speeded up the search capabilities for medical practitioners by helping them to retrieve relevant research papers for a particular disease or condition and genetic variation. Our results suggest that Algorithm 2 can be improved by adding more regular expressions to enable a broader of master terms within the biomedical text. As more information is added to SimpleGermKG, we expect to broaden its applications.

BIBLIOGRAPHY

Abreu Vicente, J. (2022a, March). *DrAbreu/bioBERT-NER-BC2GM_corpus*. https://huggingface.co/drAbreu/bioBERT-NER-BC2GM_corpus

Abreu Vicente, J. (2022b, March). *DrAbreu/bioBERT-NER-NCBI_disease*. https://huggingface.co/drAbreu/bioBERT-NER-NCBI_disease

Al-Garadi, M. A., Yang, Y.-C., & Sarker, A. (2022). The Role of Natural Language Processing during the COVID-19 Pandemic: Health Applications, Opportunities, and Challenges. *Healthcare*, *10*(11). https://doi.org/10.3390/healthcare10112270

Al-Moslmi, T., Gallofré Ocaña, M., L. Opdahl, A., & Veres, C. (2020). Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, *8*, 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. https://doi.org/10.48550/ARXIV.1904.03323

Alshaikhdeeb, B., & Ahmad, K. (2016). Biomedical Named Entity Recognition: A Review. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 889–895. https://doi.org/10.18517/ijaseit.6.6.1367

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., Dhir, R., Disaia, P., Gabra, H., Glenn, P., Godwin, A. K., Gross, J., Hartmann, L., Huang, M., Huntsman, D. G., Iacocca, M., Imielinski, M., Kalloger, S., Karlan, B. Y., … Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353), 609–615. https://doi.org/10.1038/nature10166

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. https://doi.org/10.48550/ARXIV.1903.10676

Berners-Lee, T., & Hendler, J. (2001). Publishing on the Semantic Web. *Nature*, 410, 1023–1024.

https://doi.org/10.1038/35074206

Bhatnagar, R., Sardar, S., Beheshti, M., & Podichetty, J. T. (2022). How can natural language processing help model informed drug development?: A review. *JAMIA Open*, *5*(2). https://doi.org/10.1093/jamiaopen/ooac043

Bodenreider, O., Mitchell, J., & McCray, A. (2005). Biomedical ontologies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *78*, 76–78. https://doi.org/10.1142/9789812704856_0016

Bonifaci, N., Górski, B., Masojć, B., Wokołorczyk, D., Jakubowska, A., Dębniak, T., Berenguer, A., Serra Musach, J., Brunet, J., Dopazo, J., Narod, S. A., Lubiński, J., Lázaro, C., Cybulski, C., & Pujana, M. A. (2010). Exploring the Link between Germline and Somatic Genetic Alterations in Breast Carcinogenesis. *PLOS ONE*, *5*(11), 1–8. https://doi.org/10.1371/journal.pone.0014078

Bratus, S., Rumshisky, A., Khrabrov, A., Magar, R., & Thompson, P. (2011). Domain-specific entity extraction from noisy, unstructured data using ontology-guided search. *International Journal on Document Analysis and Recognition*, 14, 201–211. https://doi.org/10.1007/s10032-011-0149-5

Cariello, M. C., Lenci, A., & Mitkov, R. (2021). *A Comparison between Named Entity Recognition Models in the Biomedical Domain*. 76–84. https://aclanthology.org/2021.triton-1.9

Chithrananda, S., Grand, G., & Ramsundar, B. (2020). *ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction*. https://doi.org/10.48550/ARXIV.2010.09885

Cho, H., Choi, W., & Lee, H. (2017). A method for named entity normalization in biomedical articles: Application to diseases and plants. *BMC Bioinformatics*, *18*. https://doi.org/10.1186/s12859-017-1857-8

Choi, W., & Lee, H. (2021). Identifying disease-gene associations using a convolutional neural network-based model by embedding a biological knowledge graph with entity descriptions. *PLOS ONE*, *16*(10), 1–27. https://doi.org/10.1371/journal.pone.0258626

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1), 57–71. https://doi.org/10.1093/bib/6.1.57

Cohen, K. B., Acquaah-Mensah, G. K., Dolbey, A. E., & Hunter, L. (2002). Contrast and Variability in Gene Names. *BioMed '02*, 14–20. https://doi.org/10.3115/1118149.1118152

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/ARXIV.1810.04805

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, *47*, 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

Dörpinghaus, J., Stefan, A., Schultz, B., & Jacobs, M. (2020). Towards context in large scale biomedical knowledge graphs. https://doi.org/10.48550/ARXIV.2001.08392

Feng, F., Tang, F., Gao, Y., Zhu, D., Li, T., Yang, S., Yao, Y., Huang, Y., & Liu, J. (2022). GenomicKB: a knowledge graph for the human genome. *Nucleic Acids Research*, *51*(D1), D950–D956. https://doi.org/10.1093/nar/gkac957

Fukuda, K., Tamura, A., Tsunoda, T., & Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 707–718.

Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H., & Charnois, T. (2018). *SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers*. 679–688. https://doi.org/10.18653/v1/S18-1111

Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic Discovery of Part-Whole Relations. *Computational Linguistics*, *32*(1), 83–135. https://doi.org/10.1162/coli.2006.32.1.83

Gratzl, S., Gehlenborg, N., Lex, A., Strobelt, H., Partl, C., & Streit, M. (2015). Caleydo Web: An Integrated Visual Analysis Platform for Biomedical Data.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. https://doi.org/10.48550/ARXIV.2004.10964

Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., Gonzalez, G., Nenadic, G., & Bergman, C. M. (2011). The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27(19), 2769–2771. https://doi.org/10.1093/bioinformatics/btr455

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, 6, e26726. https://doi.org/10.7554/eLife.26726

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Comput. Surv*., 54(4). https://doi.org/10.1145/3447772

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. https://doi.org/10.48550/ARXIV.1801.06146

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. https://doi.org/10.48550/ARXIV.1904.05342

Kalmukov, Y. (2021). Using word clouds for fast identification of papers' subject domain and reviewers' competences. https://doi.org/10.48550/ARXIV.2112.14861

Karadeniz, I., & Ozgur, A. (2019). Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC Bioinformatics*, 20. https://doi.org/10.1186/s12859-019-2678-8

Kim, J., Kim, J.-J., & Lee, H. (2019). DigChem: Identification of disease-gene-chemical relationships from Medline abstracts. *PLOS Computational Biology*, *15*(5), 1–16. https://doi.org/10.1371/journal.pcbi.1007022

Kwon, D., Kim, S., Wei, C.-H., Leaman, R., & Lu, Z. (2018). ezTag: Tagging biomedical concepts via interactive learning. *Nucleic Acids Research*, 46(W1), W523–W529. https://doi.org/10.1093/nar/gky428

Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, *57*, 28–37. https://doi.org/10.1016/j.jbi.2015.07.010

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Li, J., Sun, A., Han, J., & Li, C. (2018). *A Survey on Deep Learning for Named Entity Recognition*. https://doi.org/10.48550/ARXIV.1812.09449

Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N., & Lu, Z. (2022). BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, *23*(5). https://doi.org/10.1093/bib/bbac282

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Matentzoglu, N., Goutte-Gattat, D., Tan, S. Z. K., Balhoff, J. P., Carbon, S., Caron, A. R., Duncan, W. D., Flack, J. E., Haendel, M., Harris, N. L., Hogan, W. R., Hoyt, C. T., Jackson, R. C., Kim, H., Kir, H., Larralde, M., McMurry, J. A., Overton, J. A., Peters, B., … Osumi-Sutherland, D. (2022). Ontology Development Kit: A toolkit for building, maintaining and

standardizing biomedical ontologies. *Database*, *2022*. https://doi.org/10.1093/database/baac087

Mcguinness, D., & Harmelen, F. (2004). OWL Web ontology language: Overview. *W3C Recomm*, 10.

Meyerson, W., Leisman, J., Navarro, F., & Gerstein, M. (2020). Origins and characterization of variants shared between databases of somatic and germline human mutations. *BMC Bioinformatics*, 21. https://doi.org/10.1186/s12859-020-3508-8

Miller, E. (1998). An Introduction to the Resource Description Framework. *Journal of Library Administration*, 34, 245–255.

Milošević, N., & Thielemann, W. (2023). Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *Journal of Web Semantics*, *75*, 100756. https://doi.org/10.1016/j.websem.2022.100756

Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., & Kim, J. (2020). BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition. https://doi.org/10.48550/ARXIV.2009.09223

Neves, M., Carazo, J.-M., & Pascual-Montano, A. (2010). Moara: A Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics*, *11*, 157. https://doi.org/10.1186/1471-2105-11-157

Noh, J., & Kavuluru, R. (2021). Joint Learning for Biomedical NER and Entity Normalization: Encoding Schemes, Counterfactual Examples, and Zero-Shot Evaluation. *BCB '21*. Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, New York, NY, USA. https://doi.org/10.1145/3459930.3469533

Park, Y.-J., Lee, M.-A., Yang, G.-J., Park, S. J., & Sohn, C.-B. (2022). Biomedical Text NER Tagging Tool with Web Interface for Generating BERT-Based Fine-Tuning Dataset. *Applied Sciences*, 12(23). https://doi.org/10.3390/app122312012

Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). *Relation Extraction: A Survey*. https://doi.org/10.48550/ARXIV.1712.05191

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://doi.org/10.48550/ARXIV.1802.05365

Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., & Altan-Bonnet, G. (2021). SciFive: A text-to-text transformer model for biomedical literature. https://doi.org/10.48550/ARXIV.2106.03598

Post, L. J. G., Roos, M., Marshall, M. S., van Driel, R., & Breit, T. M. (2007). A semantic web approach applied to integrative bioinformatics experimentation: A biological use case with genomics data. *Bioinformatics*, 23(22), 3080–3087. https://doi.org/10.1093/bioinformatics/btm461

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

Richens, R. H. (1956). Preprogramming for mechanical translation. *Mech. Transl. Comput. Linguistics*, 3, 20–25.

Salzberg, S. L. (2018). Open questions: How many genes do we have? *BMC Biology*, *16*(1). https://doi.org/10.1186/s12915-018-0564-x

Sanad, M. (2019, September). Demystifying BERT: A comprehensive guide to the groundbreaking NLP framework. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

Simons, P. (1987). Parts: A Study in Ontology. *Revue de Métaphysique et de Morale*, 2, 277–279.

Simpson, M. S., & Demner-Fushman, D. (2012). Biomedical Text Mining: A Survey of Recent Progress. *Mining Text Data*.

Singh, G., Papoutsoglou, E. A., Keijts-Lalleman, F., Vencheva, B., Rice, M., Visser, R. G. F., Bachem, C. W. B., & Finkers, R. (2021). Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait. *BMC Plant Biology*, *21*(1). https://doi.org/10.1186/s12870-021-02943-5

Singhal, A. (2012, May). Introducing the Knowledge Graph: Things, not strings. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

Smith, L. L., Tanabe, L. K., Ando, R., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C. A., Povinelli, R. J., Vlachos, A., Baumgartner, W. A., Hunter, L. E., Carpenter, B., … Wilbur, W. J. (2008). Overview of BioCreative II gene mention recognition. *Genome Biology*, *9*, S2–S2.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3), 239–251. https://doi.org/10.1093/bib/6.3.239

Splendiani, A., Burger, A., Paschke, A., Romano, P., & Marshall, M. (2011). Biomedical semantics in the Semantic Web. *Journal of Biomedical Semantics*, 2 Suppl 1, S1. https://doi.org/10.1186/2041-1480-2-S1-S1

Su, C., Hou, Y., Rajendran, S., Maasch, J. R. M. A., Abedi, Z., Zhang, H., Bai, Z., Cuturrufo, A., Guo, W., Chaudhry, F. F., Ghahramani, G., Tang, J., Cheng, F., Li, Y., Zhang, R., Bian, J., & Wang, F. (2022). Biomedical Discovery through the integrative Biomedical Knowledge Hub (iBKH). *MedRxiv*. https://doi.org/10.1101/2021.03.12.21253461

Thomson, A., Heyworth, J., Girschik, J., Slevin, T., Saunders, C., & Fritschi, L. (2014). Beliefs and perceptions about the causes of breast cancer: A case-control study. *BMC Research Notes*, 7, 558. https://doi.org/10.1186/1756-0500-7-558

Truscott, R. J. W., Schey, K. L., & Friedrich, M. G. (2016). Old Proteins in Man: A Field in its Infancy. *Trends in Biochemical Sciences*, 41(8), 654–664.

https://doi.org/10.1016/j.tibs.2016.06.004

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

Verma, S., Bhatia, R., Harit, S., & Batish, S. (2023). Scholarly knowledge graphs through structuring scholarly communication: A review. *Complex & Intelligent Systems*, *9*(1), 1059—1095. https://doi.org/10.1007/s40747-022-00806-6

Wang, S., Lin, M., Ghosal, T., Ding, Y., & Peng, Y. (2022). Knowledge Graph Applications in Medical Imaging Analysis: A Scoping Review. *Health Data Science*, 2022. https://doi.org/10.34133/2022/9841548

Wang, X., Tsujii, J., & Ananiadou, S. (2009). *Classifying Relations for Biomedical Named Entity Disambiguation*. 1513–1522. https://aclanthology.org/D09-1157

Webster, J. J., & Kit, C. (1992). Tokenization as the Initial Phase in NLP. *COLING '92*, 1106–1110. https://doi.org/10.3115/992424.992434

Wei, C.-H., & Kao, H.-Y. (2011). Cross-species gene normalization by species inference. *BMC Bioinformatics*, 12 Suppl 8, S5. https://doi.org/10.1186/1471-2105-12-S8-S5

Whetzel, P., Noy, N., Shah, N., Alexander, P., Nyulas, C., Tudorache, T., & Musen, M. (2011). BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, *39*, W541-5. https://doi.org/10.1093/nar/gkr469

Wu, Y., Luo, R., Leung, H. C. M., Ting, H.-F., & Lam, T. W. (2019). *RENET: A Deep Learning Approach for Extracting Gene-Disease Associations from Literature*. Annual International Conference on Research in Computational Molecular Biology.

Yang, J., Han, S. C., & Poon, J. (2021). *A Survey on Extraction of Causal Relations from Natural Language Text*. https://doi.org/10.48550/ARXIV.2101.06426

Yang, X., Wu, C., Nenadic, G., Wang, W., & Lu, K. (2021). Mining a stroke knowledge graph from literature. *BMC Bioinformatics*, *22*(S10). https://doi.org/10.1186/s12859-021-04292-4

Yuan, H., Yuan, Z., Gan, R., Zhang, J., Xie, Y., & Yu, S. (2022). BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. https://doi.org/10.48550/ARXIV.2204.03905

Zhu, Q., Nguyen, D.-T., Grishagin, I., Southall, N., Sid, E., & Pariser, A. (2020). An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). *Journal of Biomedical Semantics*, *11*. https://doi.org/10.1186/s13326-020-00232-y

Zhu, X., Gu, Y., & Xiao, Z. (2022). HerbKG: Constructing a Herbal-Molecular Medicine Knowledge Graph Using a Two-Stage Framework Based on Deep Transfer Learning. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.799349