Ph.D. DISSERTATION

Clinical Information Extraction from Unstructured Free-Text for Pharmacovigilance

약물 감시를 위한 비정형 텍스트 내 임상 정보 추출 연구

BY

KIM SIUN

FEBRUARY 2023

DEPARTMENT OF APPLIED BIOENGINEERING
GRADUATE SCHOOL OF CONVERGENCE SCIENCE AND TECHNOLOGY
SEOUL NATIONAL UNIVERSITY

Clinical Information Extraction from Unstructured Free-Text for Pharmacovigilance

약물 감시를 위한 비정형 텍스트 내 임상 정보 추출 연구

지도교수 이 형 기
이 논문을 의학박사 학위논문으로 제출함

2023년 2월

서울대학교 대학원
융합과학기술대학원 응용바이오공학부
김 시 언

김시언의 의학박사 학위 논문을 인준함
2023년 2월

위 원 장 유 경 상 (인)
부위원장 최 진 욱 (인)
위 원 오 정 미 (인)
위 원 정 교 민 (인)
위 원 최 남 경 (인)

Abstract

Pharmacovigilance is a scientific activity to detect, evaluate and understand the occurrence of adverse drug events or other problems related to drug safety. However, concerns have been raised over the quality of drug safety information for pharmacovigilance, and there is also a need to secure a new data source to acquire drug safety information. On the other hand, the rise of pre-trained language models based on a transformer architecture has accelerated the application of natural language processing (NLP) techniques in diverse domains. In this context, I tried to define two problems in pharmacovigilance as an NLP task and provide baseline models for the defined tasks: 1) extracting comprehensive drug safety information from adverse drug events narratives reported through a spontaneous reporting system (SRS) and 2) extracting drug-food interaction information from abstracts of biomedical articles. I developed annotation guidelines and performed manual annotation, demonstrating that strong NLP models can be trained to extracted clinical information from unstructrued free-texts by fine-tuning transformer-based language models on a high-quality annotated corpus. Finally, I discuss issues to consider when when developing annotation guidelines for extracting clinical information related to pharmacovigilance. The annotated corpora and the NLP models in this dissertation can streamline pharmacovigilance activities by enhancing the data quality of reported drug safety information and expanding the data sources.

Keywords: Pharmacovigilance, Drug safety information, Natural language processing, Information extraction

Student number: 2018-20603

Contents

Abstract i

Contents ii

List of Figures vi

List of Tables viii

Chapter 1 1

1.1 Contributions of this dissertation 2

1.2 Overview of this dissertation 2

1.3 Other works 3

Chapter 2 4

2.1 Pharmacovigilance 4

2.2 Biomedical NLP for pharmacovigilance 6

2.2.1 Pre-trained language models 6

2.2.2 Corpora to extract clinical information for pharmacovigilance 9

Chapter 3 11

3.1 Motivation 12

3.2 Proposed Methods 14
3.2.1 Data source and text corpus 15
3.2.2 Annotation of ADE narratives 16
3.2.3 Quality control of annotation 17
3.2.4 Pretraining KAERS-BERT 18
3.2.6 Named entity recognition 20
3.2.7 Entity label classification and sentence extraction 21
3.2.8 Relation extraction 21
3.2.9 Model evaluation 22
3.2.10 Ablation experiment 23
3.3 Results 24
3.3.1 Annotated ICSRs 24
3.3.2 Corpus statistics 26
3.3.3 Performance of NLP models to extract drug safety information 28
3.3.4 Ablation experiment 31
3.4 Discussion 33
3.5 Conclusion 38
Chapter 4 39
4.1 Motivation 39

4.2 Proposed Methods 43
4.2.1 Data source 44
4.2.2 Annotation 45
4.2.3 Quality control of annotation 49
4.2.4 Baseline model development 49
4.3 Results 50
4.3.1 Corpus statistics 50
4.3.2 Annotation Quality 54
4.3.3 Performance of baseline models 55
4.3.4 Qualitative error analysis 56
4.4 Discussion 59
4.5 Conclusion 63
Chapter 5 64
5.1 Issues around defining a word entity 64
5.2 Issues around defining a relation between word entities 66
5.3 Issues around defining entity labels 68
5.4 Issues around selecting and preprocessing annotated documents 68
Chapter 6 71
6.1 Dissertation summary 71

6.2 Limitation and future works 72

6.2.1 Development of end-to-end information extraction models from free-texts to database based on existing structured information 72

6.2.2 Application of in-context learning framework in clinical information extraction 74

7.1 Annotation Guideline for “Extraction of Comprehensive Drug Safety Information from Adverse Event Narratives Reported through Spontaneous Reporting System” 76

7.2 Annotation Guideline for “Extraction of Drug-Food Interactions from the Abtracts of Biomedical Articles” 100

List of Figures

Figure 3.1: Overview of extracting comprehensive drug safety information from annotated ADE narratives reported through KAERS 14

Figure 3.2: Overview of proposed methods for developing annotated corpus and NLP models extracting comprehensive drug safety information 14

Figure 3.3: Examples of token tagging for NER 21

Figure 3.4: MedDRA system organ classes (SOCs) distribution of normalized ADE entities in annotated ADE narratives and ADEs normalized by reporters in KIDS-KD 27

Figure 3.5: Entity recognition for 12 key word entities by the KAERS-BERT model. ADE and RoA denotes adverse drug event and route of administration, respectively. Total sum of prediction proportions in a single row could be less than 100% because other 7 word entities are omitted in this table 30

Figure 3.6: NER performances of the KAERS-BERT model on total entities (a) and ADE entities (b) by the composition of training dataset. A *random only* dataset denotes a training dataset consisting of only (340 + M) randomly selected ADE narratives, while *ADE + random*, *indication + random*, *drug compound + random* and *drug product + random* datasets represent training datasets consisted of 340 ADE narratives reported with least reported ADE, indication, drug compound, drug product items plus M randomly selected ADE narratives, respectively. 32

Figure 4.1: Overview of proposed method for developing DFI corpus 44

Figure 4.2: Example of a manually annotated abstract for DFI extraction. Before annotating DFI, I selected abstracts that contained ≥1 drug word AND ≥1 food word simultaneously

in the same abstract. The entity types of annotated words are denoted by superscripts and highlighted in colors. In this example, two DFI key-sentences marked as light blue are annotated. Also, the relations between word entities or between a word and sentence entities were enumerated at “Relations” at the bottom, while the modalities of key-sentence entities at “Key-sentence Modality.” The evidence-level of the abstract, ‘in-vivo study’, was described at “Evidence-level.” 48

Figure 4.3: Distribution of (a) annotated word and (b) sentence entities in the DFI corpus. 51

Figure 4.4: Distribution of evidence levels of abstracts in the annotated corpus by the inclusion of DFI key-sentence. The size of the circles is proportional to the number of abstracts and the exact numbers and percentages are also reported. 53

Figure 4.5: Ternary plot showing the ratios of annotated entity types of given words in the DFI corpus. In this plot, ‘food’ denotes both ‘food’ and ‘food component’ entities, while ‘drug’ implies both ‘drug’ and ‘ambiguous’ entities. Also, ‘target’ includes ‘well-known target’, ‘drug metabolizer’, and ‘drug transporter’ entities. 54

Figure 7.1: Example of annotated drug entities in an abstract 102

Figure 7.2: Example of annotated well known target entities in an abstract 103

Figure 7.3: Example of annotated drug metabolizer entities in an abstract 104

Figure 7.4: Example of annotated food and food component entities in an abstract 105

Figure 7.5: Example of annotated DFI and DDI key-sentences in an abstract 107

Figure 7.6: Example of annotated supporting sentences in an abstract 108

List of Tables

Table 3.1: Summary characteristics of the total and annotated ICSRs 24

Table 3.2: Inter-annotator agreement on entity annotation 25

Table 3.3: Statistics of annotated entities in ADE narratives 26

Table 3.4: Entity labels and relations in annotated ADE narratives 28

Table 3.5: Performance metrics (%) of baseline models by task 29

Table 3.6: NER performance metrics of the KAERS-BERT model for entities labeled as '*occurred*' and '*not occurred*' 30

Table 4.1: Biomedical corpora developed for the extraction of drug interaction 41

Table 4.2: Frequency table for annotated abstracts 50

Table 4.3: Distribution of the annotated entity types in the DFI corpus 51

Table 4.4: Cohen's kappa by classification task between the annotators and the independent reviewer 55

Table 4.5: Performance of BERT models by DFI extraction task. The bolded and underlined performance scores indicate the best and second-best performances on a classification task, respectively. 55

Table 4.6: Examples of most likely true positive sentences in the validation dataset which were labeled as a key-sentence and also predicted as a key-sentence by the key-sentence classifier 56

Table 4.7: Examples of most likely true positive sentences in the validation dataset which

were labeled as a non key-sentence and also predicted as a non key-sentence by the key-sentence classifier 57

Table 4.8: Examples of most unlikely false negative sentences in the validation dataset which were labeled as a key-sentence but predicted as a non key-sentence by the sentence classifier were 58

Table 4.9: Examples of most unlikely false positive sentences in the validation dataset which were labeled as a non key-sentence but predicted as a key-sentence by the sentence classifier 58

Table 7.1: Entity labels for [seriousness] 92

Table 7.2: Definitions and examples for [seriousness] entity labels 93

Table 7.3: Entity labels for [action taken with drug] 97

Table 7.4: Definitions and examples for [WHO-UMC results of assessment] entity labels . 97

Table 7.5: Relation between entities representing synonyms and definition of synonym relation 109

Table 7.6: Relation between food and food component entities and definition of food component relation 109

Table 7.7: Relation between DFI key-sentence and word entities representing DFI and definition of relation 109

Table 7.8: Relation between DDI key-sentence and word entities representing DDI and definition of relation 110

Table 7.9: Relation between food-effect key-sentence and word entities representing food-effect and definition of relation 110

Chapter 1

Introduction

Pharmacovigilance is a scientific activity to detect, evaluate and understand the occurrence of adverse drug events (ADEs) or any other problems related to drug usage [1]. The current pharmacovigilance system is based on related regulations and standards that impose an obligation to monitor drug safety on countries and pharmaceutical companies through systemic efforts of regulatory agencies and WHO. One of the most important parts of routine pharmacovigilance is reporting ADEs occurred in clinical practice through a spontaneous reporting system (SRS) and analyzing safety reports to generate a safety signal. Moreover, post-marketing drug surveillance is essential to identify rare ADEs because all possible ADEs cannot be observed during clinical studies performed before drug approval. However, the data source for detecting safety signals is not limited to the post-marketing surveillance system and SRS, but also includes clinical trials and scientific literature.

In this context, natural language process (NLP) technology can modernize pharmacovigilance by enabling the extraction of clinical information from various types of unstructured free-texts. Indeed, many studies have tried to extract the ADE occurrence from social media data [2, 3] and to extract drug-drug interaction (DDI) information from biomedical literature [4-6] based on NLP techniques. Furthermore, the rise of transformer-based large language models (LLMs) [7-9] made NLP techniques easier to use, even for non-ML researchers,

because LLMs are an example of *a foundation model* adaptable to a wide range of downstream tasks through simple fine-tuning and relatively small datasets [10]. In this situation data-centric approach that emphasizes the fundamental importance of datasets rather than advanced model development has been proposed [11].

Therefore, in this dissertation, I defined the clinical information extraction for pharmacovigilance as an NLP task and showed that it is possible to develop strong NLP models extracting clinical information from unstructured free-texts through simple fine-tuning of LLMs using a quality annotated corpus. In addition, I identified issues around annotation elements for clinical information extraction based on our experience gained while performing task formulation and annotating free-text data.

1.1 Contributions of this dissertation

Contributions of this dissertation are as follows:

- I defined the extraction of clinical information from unstructured free-texts for pharmacovigilance as an NLP task and developed manually annotated corpora [12].
- I provided strong baseline models extracting drug safety information and drug-food interaction from free-texts through simple fine-tuning of transformer-based language models using quality annotated corpora [13].
- I identified issues around defining annotation elements for extracting clinical information related to pharmacovigilance.

1.2 Overview of this dissertation

I organized the remainder of this dissertation as follows. Chapter 2 describes the background and related work on the development of an NLP model to extract clinical information from free-texts

for pharmacovigilance. Chapter 3 describes the task configuration to extract drug safety information from unstructured adverse event (ADE) narratives as an NLP task. Defining the extraction of drug-food interaction information from biomedical articles is presented in Chapter 4. Chapter 5 provides experience in developing annotation guidelines for a biomedical or clinical corpus and details the challenges of defining annotation elements. Finally, this dissertation concludes in Chapter 6.

1.3 Other works

In addition to the research discussed in this dissertation, my Ph.D. coursework has also included works on biosimilar development [14] and pharmacokinetic-pharmacodynamic modeling [15].

Chapter 2

Background

2.1 Pharmacovigilance

Pharmacovigilance is a scientific activity essential to ensure the safe use of approved drugs. [1] Capturing rare and very rare ADEs is generally considered not feasible in randomized clinical trials where the study population consists of hundreds to thousands of relatively homogeneous patients. [16, 17] Moreover, safety issues resulting from manufacturing control failures are difficult to identify without continuous monitoring of ADE occurrence. [18] Therefore, I have tried to obtain clinical evidence for drug safety by monitoring the occurrence of ADEs in the post-marketing patient population.

After the thalidomide disaster in the 1950s and 1960s, an international effort was made to establish an ADE reporting and monitoring system. [19] WHO launched Program for International Drug Monitoring (PIDM) in 1968, and Uppsala Monitoring Centre (UMC), collaborating centre for PIDM, was established in 1978. In March 2022, 151 member countries participating in PIDM have a national ADE reporting system that collects individual case safety reports (ICSRs) and transmit the ICSRs to the WHO global database. [20] Additionally, regulatory agencies including the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have required pharmaceutical companies to conduct pharmacovigilance activities and pharmacoepidemiologic assessment during the post-approval period to investigate residual safety issues. [21, 22]

However, drug safety information collected in the post-marketing process inherently has data quality problems such as under-reporting [23] and completeness, i.e., missing data [24]. For example, only 10.6% of ICSRs reported to VigiBase, a global database operated by the WHO UMC, contained all necessary information like reaction onset and medicine treatment dates in 2000. [25] The quality issues of drug safety information databases are mainly because the drug safety information collected during the post-approval period is generated in a real-world setting rather than a preplanned clinical study. Although health authorities have tried to increase report quality by educating reporters or by developing guidelines providing good practices on drug monitoring, the report quality declined in many regions during late 2000s and early 2010s [24]. The cause of the decline in reporting quality is not clear, but it is suspected that the introduction of a comprehensive reporting format or the regulatory emphasis on timeliness, *e.g.*, within 15 days rules, may have had an impact. Moreover, because documenting ADEs is an extra task that physicians and nurses perform outside of their routine patient care, ADEs reporting can further increase the burden of documentation in hospitals. [26].

In this context, NLP techniques can help streamline ADE reporting and collection of drug safety information and improve the data quality in relevant databases. First, I can obtain other data sources of drug safety information like social media [27] and clinical notes [28], developing NLP models that automatically extract drug safety information from unstructured free-texts. In addition, NLP techniques can be used to systematically extract novel scientific knowledge helpful to assess drug safety, e.g., drug interaction [29, 30] and pharmacokinetics data [31], from biomedical articles. Second, NLP can improve the data quality of SRS databases and reduce the burden of reporting ADEs. NLP models that perform named entity recognition (NER) or entity linking can help reporters to structure drug safety information referenced in clinical free-texts and improve the reporting quality, alleviating the time and work burden required to report ADEs.

2.2 Biomedical NLP for pharmacovigilance

2.2.1 Pre-trained language models

The rise of transformer-based pre-trained language model (PLM) using self-supervised learning like BERT [9], GPT-2 [32] , RoBERTa [33], T5 [34] has significantly impacted the the field of NLP. As these models have achieved performances close to the state-of-the-art (SOTA) level in a wide range of downstream tasks, PLMs have become a foundation model [10]. PLMs as foundation models have brought a strong homogenization in the NLP fields, and almost all SOTA models in information extraction benchmarks also adopt transformer model architecture. Now, non-NLP researchers also can develop an NLP model of a near SOTA performance by fine-tuning PLMs on a small amount of labled data for downstream tasks of interest.

Berfore the emergence of transformer-based PLMs, static word embeddings such as Word2vec [35, 36], and recurrent neural networks (RNNs) [37] were commonly used as baseline models for various NLP tasks, including sentiment analysis, document classification and machine translation. Static word embeddings and RNNs have their own benefits.

Static word embeddings are a type of distributional representation that expresses the similarity and difference between words based on their distributional properties.These embeddings represent words as fixed-size vectors, instead of one-hot encodings, which can reduce the dimensoinality of word vectors. Addtionally, static word embeddings are trained using self-supervised learning frameworks, such as bag-of-words and skip-grams [36, 38], which allows for the creation of large amounts of training data by masking random words in text. Predicting the masked word based on surronding words can help a word embedding model learn the semantic and syntatic properties of words. Using static word embeddings as a starting point for training an NLP model can improve a model performance by leveraging the knolewdge captured in the

embeddings. In this sense, using word embeddings for NLP model development can also be seen as a form of transfer learning method, in which knowledge obtained from solving a specific problem is applied to other tasks.

Otherwise, RNNs are well-suited for dealing with key characteristics of text data as input for machine learning models: 1) text data is sequential and 2) the input length is not fixed. RNNs also introduce a strong inductive bias to NLP models [39], meaning that the current state of a system can be determined by incorporating information from previous time steps (*i.e.*, a hidden state) and the current input, and this process is recurrent. This plausible inductive bias allows RNNs to efficiently reduce the number of model parameters to estimate and make good predictions based on sequential patterns in the data, particularly when the data is limited [40]. However, RNNs has a problem of vanishing gradient, making it difficult to update the weights of the network as the gradients of weights are multiplied repeatedly during backpropagation through the RNNs [41]. A long-short term memory (LSTM) network, which introduces a cell state in addition to the hidden state of RNNs, was introduced as an alternative model structure to address the vanishing gradient problem [42].

Combining static word embeddings with RNNs has been successful in the filed of NLP [42-44], these frameworks transfer linguistic knowledge from a large amount of unlabeled texts to a target domain only through static word embeddings. Additionally, *context-sensitive* features are not properly represented by static word embddings. As a result, subsequent studies have focused on obtaining contextual word embeddings [45-47].

In this context, the transformer-based language model BERT [9] has emerged as a game changer. BERT is a transformer-based PLM that uses the attention mechanism [48] and is pretrained on masked language modeling (MLM) and next sentence prediction (NSP) tasks, which are similar to those used for training Word2vec. However, BERT differs from word2vec in

that it obtains contextual embedding through self-attention between input tokens. Since its introudction, various transformer-based PLMs such as GPT-2 [32], GPT-3 [49], XLNet [50], ELECTRA [8] have been developed by varying the text input method, model structure, model size, and self-supervised learning method,

Trasnformer-based PLMs have gained significant attention due to their state-of-the-art performance on various downstream tasks and the simplicity of the fine-tuning process [51]. They have also been successful in producing good results with relatively small amounts of training data, somtimes requring as little as hundreds of examples [52]. Furthermore, classification models can be trained by using the word embeddings provided by PLMs as input to even a simple linear model, resulting in satisfactory performance. The open-source organization HuggingFace also provides the model configuration of various transformer-based PLMs online, making it easier to utilize their capabilities for a wide range of NLP tasks [53].

In addition, PLMs specialized in the biomedical and clinical domains have been developed, as studies reported that domain-adaptive pretraining leads to performance gains [54-58]. For example, clinical BERT outperformed previous text mining models on NLP tasks in the clinical domain where drug, disease, and other medical jargon are frequently denoted in their abbreviated forms. Recently, researchers introduced a comprehensive benchmark dataset for biomedical NLP to facilitate evaluations of biomedical PLMs and accelerate progress in biomedical NLP. [55] Moreover, Korean medical BERT (KM-BERT) pre-trained on medical textbooks and health information news was developed and showed its applicability to biomedical NLP tasks in Korean clinical texts [59]. However, to the best of our knowledge, no pre-trained language model has ever been developed in both Korean and the clinical domain because large-scale Korean clinical narratives were scarce.

2.2.2 Corpora to extract clinical information for pharmacovigilance

Several human-annotated corpora of clinical narratives from diverse sources have been introduced to define NLP tasks related to pharmacovigilance such as detection of ADEs and extraction of medication information from free texts. The sources of those clinical narratives included clinical or physicians notes from electronic health records (EHRs) [28, 60], consumer reviews on medications [61], drug labels [62], social media [63-67], safety repots in the Vaccine Adverse Event Reportint System (VAERS) [68] and serious ADE reports collected during clinical trials [69]. These corpora principally focused on detecting ADEs and medication entities and normalizing dectected entities to medical ontology standards. For example, CADEC (CSIRO Adverse Drug Event Corpus), sourced from posts on *AskaPatient*, an online medical fourm dedicated to consumer reviews on medications, was created with two purposes: (1) entity identification for drugs, ADEs, symptoms, and diseases, and (2) entity normalization to the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), AMT (The Australian Medicines Terminology), and MedDRA (The Medical Dictionary for Regulatory Activities)[61] In 2018, the National NLP Clinical Changes (n2c2) shared a task and data on the extraction of ADEs and medication information from clincial narratives and tackled the NLP task in 3 steps: (1) concept extraction, (2) relation classification, and (3) construction of an end-to-end system integrating the two previous steps [60]. Also, medical text classifiers and ADE identifiers have been developed to extract ADEs from vaccine safety reports in VAERS [68, 70].

However, to the best of our knowledge, no human-annotated corpus of safety reports from SRS has ever been built except for those from VAERS. Furthermore, existing human annotated corpora for dectecting ADEs have rarely tried to extract other drug safety information hepful, and sometimes crucial, for assessing causality between a drug and an ADE. This information includes, but is not limited to, temproal relationship between drug administration and ADE occurrence and response to the withdrawal of the drug [71]. For example, the 2018 n2c2

shared task provided the annotated corpus for 9 areas including drug, reason (i.e., reason for medication or indication), and ADE in discharge summaries, but it did not contain the temporal relationship between drug administration and ADE occurrence [60]. To overcome those shortfalls, I annotated word entities and relations between entities according to the data elements defined in the ICH E2B(R3) guideline, which provides the formats and data requirements for electronic transmission of different types of ICSRs [72].

While numerous NLP models have been developed to extract ADEs from clincial narratives and some showed acceptable performances in detecting ADEs and related word entities [28, 60-62, 65, 66], the robustness of the extraction performances is still questionable due to the narrow clinical context of the annotated corpora. For example, while the best systems achieved F1-scores of 0.82-0.86 for the NER and relation extraction in MADE 1.0 challenges, the annotated corpus consisted of longitudinal EHR notes of only 21 randomly selected cancer patients at a single hospital [28]. Likewise, CADEC only included cosumer reviews on 12 drugs such as diclofenac or atorvastatin as their active ingredient [61], and human-annotated narratives in the VAERS only consisted of safety reports from patients with Guillain-Barre syndrome [68]. Therefore, extraction performance of an NLP system developed and evaluated only in limited clinical contexts could be lower in a wider clinical context. Furthermore, the risk of ADE occurrence and its description would be different according to a clinical setting for a patient who experiences an ADE (e.g., age, commorbidity and existence of an ADE reporting system in the hospital) [73, 74].

Chapter 3

Extraction of Comprehensive Drug Safety Information from Adverse Event (ADE) Narratives Reported through Spontaneous Reporting System

SRS is one of the most important sources of drug safety information for drug monitoring. However, ADE narratives reported through SRS has not been used to evaluate the drug safety, because they are reported as an unstructured form. Thus, I defined the extraction of comprehensive drug safety information from ADE narratives in SRS as an NLP task in this section. In addition, I developed manually annotated corpus based on the annotation guideline I developed and provide baseline models for NLP task related to the extraction of drug safety information. I expect that the annotated corpus and baseline models could be used to improve the data quality of SRS by further extracting drug safety information from ADE narratives and integrating them into structured database.

3.1 Motivation

Post-marketing surveillance is essential for monitoring and assessing ADEs, harms caused after appropriate or inappropriate use of a drug [75]. In many countries, ADE are voluntarily reported through a SRS as an ICSRs using a pre-structured format that investigates ADE(s) experienced by an individual patient [76, 77]. The regulatory agencies use the information on drug safety collected through an SRS to identify a potential safety concern and adjust strategies accordingly for efficient pharmacovigilance[77]. For example, the US FDA developed the FDA Adverse Event Reporting System (FAERS) to gather the drug safety information of marketed drugs and support their post-marketing surveillance program. Similarly, the Korea Adverse Event Reporting System (KAERS) was established in 2012 by the Korea Institute of Drug Safety and Risk Management (KIDS) to facilitate reporting ADEs and their management.

The number of ADE reports through SRS is substantial partly thanks to electronic submission of ICSRs. For example, more than two hundred thousand cases have been reported through KAERS every year since 2016 [78]. Truly, a large number of ADE reports is crucial to early detect a safety issue and to discover rare ADEs in the post-marketing phase [77]

However, concerns have been raised over the quality of drug safety information collected through SRS such as data incompleteness and under-reporting [23, 79]. Substandard data completeness has impeded the regulatory agency from appropriately assessing the relationship between an ADE and a drug based on ICSRs uploaded to an SRS. While KIDS has run an education program to improve the reporting quality of ICSRs, the completeness of several key data elements including drug indication and patient medical history was lower than 75% [80]. Moreover, data missingness has become more frequent as more comprehensive and lengthier forms were introduced [24].

Based on this understanding, I aimed to develop NLP models that automatically extract comprehensive drug safety information from ADE narratives, i.e., free texts detailing one or more

ADEs experienced by a patient and his/her clinical setting. Those NLP models are expected to greatly improve and complement the completeness of ICSRs collected through SRS. To this end, I constructed a manually annotated corpus and defined NLP tasks including NER, sentence extraction, relation extraction, label classification and entity normalization to formulate the extraction of comprehensive drug safety information from ADE narratives (Figure 3.1). In this study, I provided baseline models for NER, sentence extraction, relation extraction and label classification. Also, I pre-trained domain-specific BERT (Bidirectional Encoder Representations from Transformers) models specialized in clinical texts, where code-switching between English and Korean is frequent. Furthermore, I investigated how the performance of extracting drug safety information improves when a training dataset consists of more diverse ADE narratives as an ablation study.

Our contribution can be summarized as follows:

- From the KAERS system, I built datasets consisting of ADE narratives annotated with fine-grained drug safety information.
- I newly designed NLP tasks including named entity recognition, sentence extraction, relation extraction, text classification and entity normalization to extract drug safety information from unstructured ADE narratives.
- I proposed strong baseline models for our designed tasks using domain-specific BERT models, which incorporated code-switching language knowledge as well as medical knowledge.

3.2 Proposed Methods

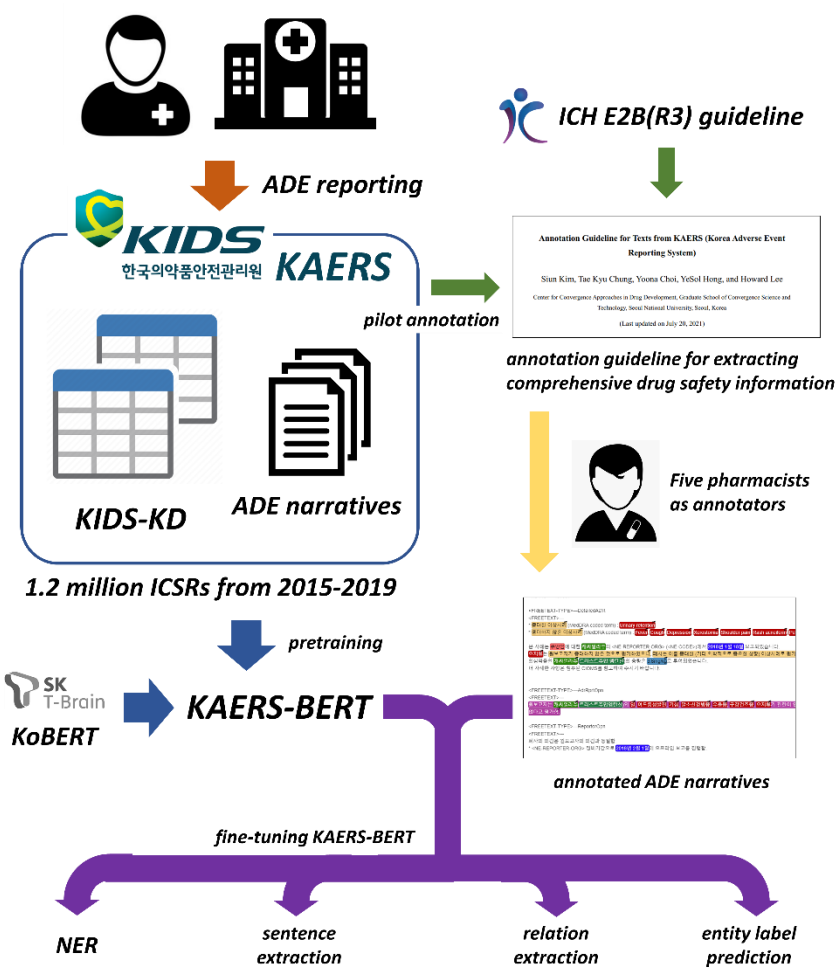

Figure 3.2: Overview of proposed methods for developing annotated corpus and NLP models extracting comprehensive drug safety information

Figure 3.1: Overview of extracting comprehensive drug safety information from annotated ADE narratives reported through KAERS

3.2.1 Data source and text corpus

ADE narratives and structured drug safety information were obtained from 1.2 million ICSRs reported through KAERS between January 1, 2015 and December 31, 2019. I created documents for the extraction of drug safety information by concatenating five types of ADE narrative in an individual ICSR: *disease history in detail*, *adverse event in detail*, *laboratory test in detail*, *reporter's opinion*, and *original reporter's opinion*. Then, I removed documents originated from ADE narratives either too short, i.e., <100 characters, or too long, i.e., >740 characters to control the reporting quality of ADE narratives and lighten the annotation burden. The 25^{th} and 95^{th} percentiles for the length of the documents were 100 and 740 characters, respectively. Additionally, I excluded a document where an ADE occurred during pregnancy because the data elements for ADEs experienced by pregnant women, infants and children were not defined in our annotation system. Furthermore, I anonymized all the ADE narratives by replacing patient identification information including patient names, address, and hospital name with special tokens such as <NE-PERSON-NAME> or <NE-HOSP-NAME> using a rule-based de-identification algorithm I developed.

I selected ADE narratives for annotation in two ways. First, I randomly selected ADE narratives for annotation. Second, to diversify the type of the documents, I selected additional ADE narratives from ICSRs that contained the least reported items related to drug safety information, i.e., adverse events, indications, drug compounds and drug products. The least reported items were determined based on the KIDS-KAERS Database (KIDS-KD), a structured drug safety information database by KIDS. To compare the clinical context between total and annotated ICSRs, I performed an exploratory data analysis on the report types and structured drug safety information of ICSRs using KIDS-KD.

3.2.2 Annotation of ADE narratives

I defined data elements for the extraction of drug safety information from ADE narratives based on the ICH E2B(R3) guideline [72]. Furthermore, the data elements that were rarely described in ADE narratives (e.g., reporter's name, whether autopsy was done) and related to pregnancy (e.g., parent information, gestation period at the time of exposure) were excluded from the annotation. Then, I developed an annotation guideline that defined word entities, relations between entities, entity labels to formulate the extraction of drug safety information as an NLP task. The annotation guideline was reviewed by three pharmacoepidemiology experts to ensure that entities and relations were correctly defined. The annotation guideline was converted into a web-based annotation system using *tagtog* service[1].

In this study, I defined 21 types of word entities to capture comprehensive drug safety information in the ICH E2B(R3) guideline. These word entities were divided into six categories: *clinical finding*, *drug*, *dosing information*, *date*, *patient information*, and *others*. 'ADE' and 'Disease' (i.e., patient's prior or present disease) entities belong to *clinical finding* along with 'ADE Seriousness' and 'ADE at last observation'. A clear distinction between 'ADE' and 'Disease' is the key component for extracting drug safety information, because mistaking 'Disease' for 'ADE' could undermine the safety of medical products. Thus, I recognized signs, symptoms and diseases diagnosed after the administration of the concerned drug as 'ADE', while those diagnosed before the administration of the concerned drug as 'Disease'. In the *drug* and *dosing information* categories, I defined word entities for capturing drug names (i.e., 'Drug compound', 'Drug product' and 'Drug group') and word describing dosing information (e.g., 'Dose' and 'Dosing Interval'). 'Date' and 'Date Period' entities, collectively classified as the *date* category,

[1] https://www.tagtog.net/

were defined to capture the temporal information of disease diagnosis, ADE occurrence, drug administration and more. Word entities that help assess the causality between drug and ADE including 'Test name', 'Test result', 'Non-drug treatment' and 'Action taken with drug' were put into the *others* category. 'WHO-UMC assessment' entity is the only sentence entity in our annotation guideline.

Additionally, I defined 59 types of relations between word entities. For example, a relation between 'Disease' and 'Drug Compound' indicated that the drug compound was prescribed for the disease. In the annotation guideline, I provided clear instructions how to annotate a relation between two entities. Furthermore, 6 entity labels were created to give detailed information on annotated entities. For example, I put an 'occurred' label to 'ADE', 'Disease' and three drug entities, i.e., 'Drug Compound', 'Drug product' and 'Drug group, to denote whether a mentioned entity actually occurred in a patient or was administered to. Likewise, a 'concerned' label was added to three drug entities to indicate whether a mentioned drug entity was a suspected drug. Then, I performed an entity normalization for 'ADE', 'Disease' and three drug entities using MedDRA 24.0 (English and Korean) and the national drug code directory provided by the Ministry of Food and Drug Safety.

Detailed explanations of word entities, relations and entity labels in the annotation guideline can be found in Chapter 7 Appendix, section 7.1.

3.2.3 Quality control of annotation

Five pharmacists who had experience in monitoring and reporting of ADEs in a pharmaceutical company or a pharmacy were recruited as annotators. They underwent a one-week education program, through which they became familiar with the annotation guideline and performed

preliminary annotations to understand how the annotation system works. Furthermore, confusing annotation examples were presented and explained to the annotators to help them understand annotation principles in the guideline. Then, 80 to 120 documents a week were annotated by each annotator for 7 weeks. Ten percent of the documents were assigned to two different annotators at the same time in order to calculate the inter-annotator agreement. ~4,000 documents were annotated by five annotators for the entire period given dual annotations. In contrast, annotators performed an entity normalization for ADE and drug entities only in 20 documents a week to lift their annotation burden.

To examine annotation quality, an independent reviewer (Siun Kim) separately reviewed 15% of documents annotated by the five annotators. When annotation agreement between the independent reviewer and an individual annotator was <80% of documents, the annotator was asked to re-do all of the document assigned to him or her for that week. In addition, I investigated all the annotated documents in the form of JSON file to check whether the annotators accurately followed the annotation guideline. When an annotator was found to obviously violate the annotation guideline for a document (e.g., missing entity labels), I manually re-annotated the document. Annotation quality for NER was assessed using Cohen's kappa [81] defined as follows:

$$\kappa = \frac{Po - Pc}{1 - Pc} \tag{3.1}$$

where *Po* represents the proportion of annotations that two annotators agreed with each other, *Pc* represents the proportions of annotations agreed between two annotators by chance.

3.2.4 Pretraining KAERS-BERT

Since ADE narratives collected through KAERS were written in Korean and contained large medical jargons and abbreviations, I newly developed a domain-specific Korean BERT (KAERS-

BERT) model[2] to incorporate the semantic knowledge from ADE narratives in KAERS. I trained KAERS-BERT by pretraining KoBERT using masked language modeling on 1.2 million ADE narratives reported through KAERS. I only used ADE narratives in the *'disease history in detail'* and *'adverse event in detail'* to pretrain KAERS-BERT as narratives in the *'laboratory test in detail'*, *'reporter's opinion'*, and *'original reporter's opinion'* sections tended to contain similar information redundantly such as causality assessment of ADEs. I tokenized ADE narratives using the KoBERT WordPiece tokenizer, which was developed based on the Korean Wikipedia, randomly masked 15% of tokens with a '[MASK]' token. I used a maximum sequence length of 200, a learning rate scheduling was 5×10^{-5}, and Adam as the optimizer. I used a warm-up scheduler with a warmup ratio of 0.1 and trained the model for up to 5 epochs. I set the weight decay to 0.1, with the exception of the weights of normalization layers and all the bias parameters in the model.

3.2.5 Training baseline models

I developed and trained four deep-learning baseline models for four defined extraction tasks such as NER, sentence extraction, relation extraction, and label classification (i.e., 'occurred' and 'concerned'): 1) long short-term memory (LSTM), 2) bi-LSTM (bidirectional LSTM), 3) KoBERT, and 4) KAERS-BERT. In all of the settings, I used the KoBERT Word-Piece tokenizer. The KoBERT model configurations were from Huggingface[3]. The LSTM and bi-LSTM models consisted of 300-hidden layers. I used the Adam optimizer [82] with a learning rate of 5×10^{-5} to train all the baseline models, while a drop-out probability and a batch size set to 0.4 and 8,

[2] https://github.com/SKTBrain/KoBERT (There have been no academic papers published on KoBERT pretraining yet)

[3] https://huggingface.co/monologg/kobert

respectively. I applied gradient clipping [83] to ensure stable learning and predicted the token label by inputting the logits of each token into conditional random field (CRF). The loss function used for training was the negative log-likelihood calculated from the CRF. Additionally, since the proportion of 'none' type tokens was overwhelming, I excluded them from the calculation of performance metrics when they were correctly predicted as 'none' type tokens. The warmup schedule was set to linear, and the warmup step was set to 100.

3.2.6 Named entity recognition

I formalized an NER task as a token-level sequence classification based on the BIO scheme [84]. Annotated tokens were tagged using the BIO scheme, where a token at the beginning of, inside, and outside the entity was labeled as "B", "I", and "O", respectively. I gave an NER tag based on word-level tokenization instead of the WordPiece tokenization. Thus, I tagged and used the first WordPiece token in each word in training NER models (Figure 3.3). Thus, WordPiece tokens tagged as 'X' were excluded in calculating a loss function and performance metrics. I combined "Date end" and "Date start" entities into a "Date" entity. Likewise, "Event admission" and "Event discharge" entities were combined into an "Event hospitalization" entity, resulting in a total of 19 entity types for the NER task.

I developed baseline models for the NER task by training BERT and RNN with CRF (conditional random fields) to capture the semantic dependency within a sentence [85]. I used a negative log-likelihood as a loss function of CRF for training. I predicted the tags of word entities to inference entity types of tokens by generating the tag sequence that maximizes a log-likelihood via the Viterbi algorithm [86].

Word-level phrase	삼중음성*triple-negative*	유방암*breast cancer*	진단*diagnosis*	이후에*after*
Entity type	ADE	ADE	None	None
WordPiece tokens	[_삼] [중] [음] [성]	[_유] [방] [암]	[_진단]	[_이후] [에]
Tagging result	<B-ADE> <X> <X> <X>	<I-ADE> <X> <X>	<O>	<O> <X>

Figure 3.3: Examples of token tagging for NER

3.2.7 Entity label classification and sentence extraction

Label classification for ‘occurred’ and ‘concerned’ labels was formalized as a token-level sequence classification with three label types: 1) ‘positive’, 2) ‘negative’ and 3) ‘unrelated’. I gave a token a ‘positive’ label when the token was annotated as ‘occurred’ or ‘concerned’, while I considered a token ‘negative’ when the token was annotated as ‘not occurred’ or ‘not concerned’. Word entities, for which neither ‘occurred’ nor ‘concerned’ label classification was applicable to the entity, were labeled as ‘unrelated’. Sentence extraction for the ‘WHO-UMC assessment’ entity was formalized as a token-level sequence classification with a binary IO scheme, where a token at the inside and outside the entity was labeled as ‘I’ and ‘O’, respectively. Thus, sentence extraction models were trained to predict whether a token is positioned inside a ‘WHO-UMC assessment’ sentence. I used CRF to train baseline models for both label classification and sentence extraction.

3.2.8 Relation extraction

I formalized relation extraction as a pair-wise binary classification for token pairs, for which entity types were defined as possibly related to each other in the annotation guideline. I originally defined a total of 59 types of relations to capture comprehensive drug safety information in the ADE narratives. However, I used only 15 types of relations to improve the quality of an annotation dataset and balance the ratio between the annotated and negative relations in a dataset. Negative

relation was an entity pair that was not annotated as related. Also, I limited the maximum number of negative relations in a single ADE narrative up to 40 to balance positive (i.e., annotated) and negative relations in the training dataset.

I obtained average pooling of tokens in related entities and concatenated token embeddings of [CLS] and two mention pools [87] to develop relation extraction models as follows:

$$\textbf{Token embeddings}: \begin{bmatrix} [\text{CLS}], t_0, \dots, [E_{1,start}], t_i, \dots, t_j, [E_{1,end}], \dots, \\ [E_{2,start}], t_k, \dots, t_l, [E_{2,end}], \dots t_n \end{bmatrix} \quad (3.2)$$

$$\textbf{Mention pools}: E_{1,pool} = AveragePool([t_i, \dots, t_j]), E_{2,pool} = AveragePool([t_k, \dots, t_l]) \quad (3.3)$$

$$\textbf{Total Representation}: [[\text{CLS}], E_{1,pool}, E_{2,pool}] \quad (3.4)$$

where [CLS] is an embedding of special classification token representing a whole sentence; t_i denotes an individual token embedding; $E_{i,start}$ and $E_{i,end}$ are the start and end tokens embeddings of two-word entities, respectively, related to each other; *AveragePool* is a pooling operation based on the element-wise average on tokens of same size. I used a multi-layer perceptron to train baseline models on the relation extraction.

3.2.9 Model evaluation

I split the annotated ADE narratives into training and test datasets with a proportion of 9:1 (Table 3). I calculated precision, recall and F1-score to evaluate the information extraction performances of baseline models as follows:

$$\textbf{Precision} = \frac{Number\ of\ correct\ predictions\ about\ all\ unit\ samples}{Number\ of\ all\ unit\ samples\ in\ test\ documents} \quad (3.5)$$

$$\textbf{Recall} = \frac{Number\ of\ correct\ predictions\ about\ positive\ unit\ samples}{Number\ of\ positive\ unit\ samples\ in\ test\ documents} \quad (3.6)$$

$$\textbf{F1 score} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3.7)$$

where a unit sample means a valid input-output pair for a given NLP task. A unit sample for the NER task was a token that was assigned a NER tag except 'X' and the token's NER tag, while a unit sample for the label classification was a token, for which the label annotation and the token's label (i.e., 'concerned' or 'occurred' label) was applicable to the entity. In the sentence extraction, all of the tokens in the dataset were used as an input token and it was predicted whether tokens positioned inside a 'WHO-UMC assessment' sentence. In the relation extraction, I used positive and negative relations and their binary modality as unit samples.

The proportions of predicted entity types were reported to investigate how well the KAERS-BERT model performed in classifying word entities. I observed that entities labeled as 'not occurred' were more likely to appear as a list of the most common ADEs of a concerned drug in ADE narratives, e.g.,*"Tramadol can cause nausea, vomiting, constipation, or drowsiness."*, while entities labeled as 'occurred' were not. For this reason, I were concerned that the difference in the way entities were written in an ADE narrative between those labeled as 'occurred' and 'not occurred' might have an effect on the NER performances. Thus, I assessed the NER performances of the KAERS-BERT model separately for 'occurred' and 'not occurred' entities.

3.2.10 Ablation experiment

I performed an ablation experiment to investigate whether a baseline NER model was improved when a training dataset contained more diverse ADE narratives. To this end, I created five training datasets consisting of 1) 340 ADE narratives randomly selected from the total ICSRs and 340 ADE narratives from ICSRs that contained the least reported items, i.e., 2) ADE, 3) indication, 4)

drug compound and 5) drug product. First, I trained KAERS-BERT using the five training datasets of 340 ADE narratives (M = 0). Then, I added more randomly selected ADE narratives to each of the five training datasets at a number (M) of 340, 1020, and 1700. Thus, training dataset 1 consisted of only randomly selected 340 + M ADE narratives of (*random only)*, while the other four datasets consisted of 340 ADE narratives randomly selected and M ADE narratives with the least reported items (*ADE* + *random*, *indication* + *random*, *drug compound* + *random*, and *drug product* + *random*, respectively). The performance of baseline NER models was calculated as M was increased.

3.3 Results

3.3.1 Annotated ICSRs

I annotated 3,723 ADE narratives out of the 1,199,498 total ICSRs reported through KAERS between January 1, 2015 and December 31, 2019 (

Table 3.1). The overall characteristics of ADE narratives were similar between the total and annotated ICSRs. A total of 235 (6.3%) ADE narratives were doubly annotated by two different annotators, and 580 (15.6%) by the independent reviewer and an annotator. The agreement was high not only between the annotators and the main reviewer, but between any two annotators (Cohen's kappa 96.5% and 85.9%, respectively, Table 3.2).

Table 3.1: Summary characteristics of the total and annotated ICSRs

ICSRs categories†	Total ICSRs (N = 1,199,498)	Annotated ICSRs (N = 3,723)
Patient Age, years	54.1 ± 19.0	52.3 ± 20.3
Female patient, N (%)	720,882 (60.1)	2,127 (57.1)
No. of reported drugs	2.3 ± 4.4	3.4 ± 6.0
No. of reported ADEs	1.5 ± 1.3	1.7 ± 1.3
No. of reported medical histories	1.3 ± 1.2	1.5 ± 1.7
Reports with serious ADE, N (%)	120,258 (10.0)	555 (14.9)
Report type, N (%)		
Spontaneous reports	984,332 (82.1)	2,998 (80.5)
Reports from survey research	180,224 (15.0)	586 (15.7)
Reports from literature	13,225 (1.1)	71 (1.9)
Other reports	21,669 (1.8)	68 (1.8)
Initial or follow-up, N (%)		
Initial report	1,132,176 (94.4)	3,500 (94.0)
Follow-up	67,322 (5.6)	223 (6.0)
Reporter type, N (%)		
Regional pharmacovigilance center	844,773 (70.4)	2,518 (67.6)
Manufacturer	306,678 (25.6)	1,098 (29.5)
Medical institution	36,594 (3.1)	69 (1.9)
Consumer	5,945 (0.5)	26 (0.7)
Other	5,508 (0.5)	12 (0.3)

Abbreviations: ICSR, individual case safety report; N, number; KAERS, Korean Adverse Event Reporting System; ADE, adverse drug event; ATC, Anatomical Therapeutic Chemical classification system; ICD-10, the 10th revision of the International Statistical Classification of Diseases and Related Health Problems; WHO-ART, World Health Organization-Adverse Reaction Terminology; SOC, system-organ classes

† Except where indicated otherwise, values are the mean ± standard deviation.

Table 3.2: Inter-annotator agreement on entity annotation

Annotators	IAA	Number of documents annotated by two annotator
IAA between the main reviewer and an annotator		
Annotator 1	97.55%	134
Annotator 2	94.98%	108
Annotator 3	94.91%	102
Annotator 4	97.68%	148
Annotator 5*	95.87%	16
Weighted average, without annotator 5	96.48%	492

Weighted average	96.46%	508

Abbreviation: IAA, inter-annotator agreement.
* Annotator 5 discontinued annotation program after the first week

3.3.2 Corpus statistics

The annotated corpus contained a total of 86,750 entities extracted from 2,378 randomly selected ADE narratives and 336, 336, 337, 336 ADE narratives with least reported ADEs, disease, drug compounds, and drug products, respectively (Table 3.3). All the entities defined in this study were annotated more than 300 times. The most frequent entity was 'ADE' (39.8%), while drug entities including 'Drug compound', 'Drug product' and 'Drug group' comprised 19.8% of the total annotated entities. The overall distributions of system organ class (SOC) were similar between the normalized ADE entities in annotated ADR narratives and the ADEs normalized by reporters in KIDS-KD (Figure 3.4).

Table 3.3: Statistics of annotated entities in ADE narratives

Entity types, no. per narrative (%)	**Total Narratives (N = 3,723)**	**Selection methods of ADE narratives for annotation**	
		randomly selected (N = 2,378)	**with least reported items (N = 1,345)**
Pathological finding			
ADE	9.27 (39.8)	9.55 (41.6)	11,787 (36.7%)
Disease	0.97 (4.2)	0.87 (3.8)	1,529 (4.8%)
ADE seriousness	0.15 (0.7)	0.11 (0.5)	293 (0.9%)
ADE at the last observation	0.53 (2.3)	0.51 (2.2)	756 (2.4%)
Drug			
Drug compound	1.45 (6.2)	1.37 (6.0)	2,120 (6.6%)
Drug product	2.61 (11.2)	2.56 (11.2)	3,635 (11.3%)
Drug group	0.56 (2.4)	0.63 (2.8)	584 (1.8%)
Dosing information			
Dose	0.93 (4.0)	0.97 (4.2)	1,136 (3.5%)
Dosing interval	0.2 (0.9)	0.18 (0.8)	317 (1%)
RoA or formulation	0.6 (2.6)	0.64 (2.8)	736 (2.3%)

Date			
Date	2.21 (9.5)	2.06 (9.0)	3,324 (10.3%)
Period	0.25 (1.1)	0.22 (1.0)	415 (1.3%)
Patient information			
Patient sex	0.13 (0.5)	0.11 (0.5)	210 (0.7%)
Patient age	0.08 (0.4)	0.07 (0.3)	139 (0.4%)
Others drug safety information			
Hospitalization event	0.26 (1.1)	0.22 (1.0)	426 (1.3%)
Test name	0.58 (2.5)	0.53 (2.3)	904 (2.8%)
Test result	0.51 (2.2)	0.45 (2.0)	822 (2.6%)
Non-drug treatment	0.27 (1.1)	0.25 (1.1)	408 (1.3%)
Action taken with drug	0.58 (2.5)	0.56 (2.4)	823 (2.6%)
WHO-UMC assessment result	1.18 (5.1)	1.10 (4.8)	1,775 (5.5%)
Average numbers in a narrative	23.3 (100.0)	22.97 (100.0)	32,139 (100%)
Total numbers	86,750	54,611	7,468

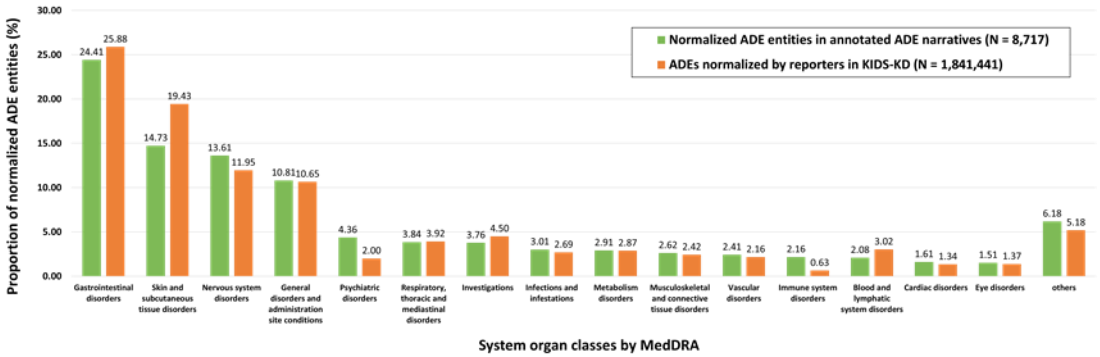

Figure 3.4: MedDRA system organ classes (SOCs) distribution of normalized ADE entities in annotated ADE narratives and ADEs normalized by reporters in KIDS-KD

Furthermore, the annotated corpus contained a total 81,828 entity labels and 45,107 relations related to the extraction of drug safety information (Table 3.4). In total, 40.0% of 'ADE', 'Disease' and drug entities were labeled as 'not occurred', while 17.1% of drug entities were labeled as 'not concerned'. Among 59 relation types, 24 relations were used more than 500 times

for annotation.

Table 3.4: Entity labels and relations in annotated ADE narratives

Entity label and relation types	Annotated ADE narratives		
	Total (N = 3,723)	**Train (N = 3,408)**	**Test (N = 315)**
Entity labels ("occurred" and "concerned")			
"occurred" label			
‘occurred’	33,291 (60.0)	30,386 (60.1)	2,905 (58.5)
‘not occurred’	22,229 (40.0)	20,168 (39.9)	2,061 (41.5)
` **Total**	55,520 (100.0)	50,554 (100.0)	4,966 (100.0)
"concerned" label			
‘concerned’	14,309 (82.9)	13,088 (83.5)	1,221 (77)
Not Concerned	2,945 (17.1)	2,580 (16.5)	365 (23)
Total	17,254 (100.0)	15,668 (100.0)	1,586 (100.0)
Relations (most frequent 10)			
ADE and *WHO-UMC assessment*	7,711 (17.1)	7,030 (17.1)	681 (16.6)
Drug product and *WHO-UMC assessment*	3,947 (8.8)	3,604 (8.8)	343 (8.4)
ADE and *Date*	3,134 (7.0)	2,809 (6.9)	325 (7.9)
ADE and *ADE*	3,043 (6.7)	2,746 (6.7)	297 (7.2)
ADE and *ADE at the last observation*	2,330 (5.2)	2,150 (5.2)	180 (4.4)
Drug product and *Dose*	2,366 (5.2)	2,135 (5.2)	231 (5.6)
Drug product and *Date*	2,040 (4.5)	1,838 (4.5)	202 (4.9)
Test name and *Test result*	1,877 (4.2)	1,712 (4.2)	165 (4)
Drug product and *RoA or Formulation*	1,544 (3.4)	1,432 (3.5)	112 (2.7)
Drug product and *Action taken with drug*	1,473 (3.3)	1,361 (3.3)	112 (2.7)
Total	45,107 (100.0)	4,1001 (100.0)	4,106 (100.0)

The total number (column %) is displayed.
Abbreviations: ADE, adverse drug event; N, number; WHO-UMC, WHO-UMC, World Health Organization-Uppsala Monitoring Centre

3.3.3 Performance of NLP models to extract drug safety information

The KAERS-BERT model outperformed all the other models including the KoBERT model in four NLP tasks to extract comprehensive drug safety information (Table 3.5). The F1-scores of the KAERS-BERT model were >80% for the tasks of NER, sentence extraction, and label classification for the 'occurred' entity label, which were 2.35-4.85 percentage points higher than the second-best performing KoBERT model.

Table 3.5: Performance metrics (%) of baseline models by task

Task*	**Baseline models**			
	KAERS-BERT + CRF	**KoBERT + CRF**	**BILSTM**	**LSTM**
NER				
Precision	**83.90**	80.48	77.47	72.38
Recall	**83.72**	78.95	68.63	57.28
F1-score	**83.81**	79.71	72.78	63.95
Sentence extraction				
Precision	76.62	65.73	**85.51**	74.80
Recall	**73.00**	53.47	72.08	69.28
F1-score	**80.63**	78.28	78.23	71.94
Relation extraction				
Precision	**69.48**	58.09	59.15	68.32
Recall	**59.95**	59.61	46.87	48.46
F1-score	**64.37**	58.84	52.30	56.71
Label classification				
'occurred' label				
Precision	**79.85**	79.13	75.11	72.01
Recall	**82.87**	74.01	72.14	61.48
F1-score	**81.33**	76.48	73.60	66.33
'concerned' label				
Precision	**78.84**	74.89	74.75	67.41
Recall	**76.43**	65.91	62.74	38.58
F1-score	**77.62**	70.12	68.22	49.08

Abbreviations: KAERS, the Korea Adverse Event Reporting System; BERT, bidirectional encoder representations from transformers; CRF, conditional random fields; LSTM, long short-term memory; BILSTM, bidirectional LSTM; NER, named entity recognition

* The best and second-best performance scores are highlighted in bold and underline, respectively.

Even in cases that the KAERS-BERT model failed to correctly recognize entities, most

of them were reasonably predictable (Figure 3.5). For examples, about half of misclassified ‘Drug compound’ entities (20.21%) were recognized as ‘Drug group’ (10.24%), which was still related to ‘Drug compound’. (Table 5).

True entity types	Predicted entity type (%)												
	none	ADE	Disease	Drug compound	Drug product	Drug group	Dose	Dosing interval	RoA or formulation	Date	Period	Test name	Test result
ADE	5,85	91,99	0,89	0,07	0,1	0,24	0	0,05	0,1	0	0	0,29	0,1
Disease	4,4	10,26	80,77	0,18	0,18	1,1	0	0	0	0	0,18	0,37	0,55
Drug compound	4,57	0,55	1,28	78,79	1,83	10,24	0,55	0	0,55	0	0	0,55	0
Drug product	7,34	0,7	0,35	0,35	84,27	0,7	0	0	2,8	0	0	0,7	0
Drug group	4,07	0,85	0,09	4,64	0,28	86,74	0,19	0	2,84	0	0	0,19	0
Dose	3,23	0	0	0	0	0,6	92,14	2,82	0	0,6	0,2	0,2	0
Dosing interval	9,09	0	0	0	0	0	11,57	76,03	1,65	1,65	0	0	0
RoA or formulation	8,33	1	0	0	2	2,67	0,33	0,67	84,33	0	0	0	0,33
Date	13,27	0	0	0	0	0	6,64	1,42	0	65,88	11,85	0	0,47
Period	1,69	0,17	0	0,08	0	0,17	0,34	0,08	0	1,01	95,53	0,08	0,59
Test name	9,59	2,4	2,05	1,37	0,68	0,34	0	0	0,68	0	0,34	76,37	4,45
Test result	20	2,17	1,74	0	0	0,43	1,74	0	0	0,87	1,3	3,48	63,04

Figure 3.5: Entity recognition for 12 key word entities by the KAERS-BERT model. ADE and RoA denotes adverse drug event and route of administration, respectively. Total sum of prediction proportions in a single row could be less than 100% because other 7 word entities are omitted in this table

Furthermore, the NER performances on entities labeled as ‘occurred’ were generally not lower than those on entities labeled as ‘not occurred’ except ‘Drug compound’ entity (Table 3.6). In ‘Drug compound’, an F1-score on ‘occurred’ entities was 5.08% lower than that on ‘not occurred’ entities.

Table 3.6: NER performance metrics of the KAERS-BERT model for entities labeled as ‘*occurred*’ and ‘*not occurred*’

Performance metrics	**ADE**	**Disease**	**Drug compound**	**Drug product**	**Drug group**

Precision					
‘occurred’	97.38	82.02	90.46	91.75	92.41
‘not occurred’	93.69	86.31	87.75	92.48	90.40
Difference†	3.69	-4.29	2.70	-0.73	2.02
Recall					
‘occurred’	90.69	79.46	69.90	89.48	88.06
‘not occurred’	88.31	75.90	80.18	86.00	80.73
Difference†	2.38	3.56	-10.28	3.48	7.33
F1-score					
‘occurred’	93.80	80.43	78.60	90.19	89.98
‘not occurred’	90.92	80.77	83.67	88.83	85.27
Difference†	2.88	-0.34	-5.08	1.36	4.71

Abbreviations: NER, named entity recognition; KAERS-BERT, Korea Adverse Event Reporting System-bidirectional encoder representations from transformers; ADE, adverse event
†The difference is performance score on entities labeled as ‘occurred’ minus that on entities labeled as ‘not occurred’.

3.3.4 Ablation experiment

NER performances were better when using both randomly selected ADE narratives and those with least reported items than when using only randomly selected ADE narratives, i.e., *random only*, particularly when the sample size of the training dataset was sufficiently large (Figure 3.6). When the KAERS-BERT model was trained using only 340 ADE narratives (M = 0), the NER performance on total entities was better when using the *random only* dataset (F1-score of 68.08%). However, when M became greater than or equal to 340, the NER performances on total entities were consistently better when using *drug product* + *random* datasets than when using the *random only* dataset (Figure 3.6a). Also, the NER performance on ADE entities was better when using datasets contained ADE narratives with least reported items than when using the *random only* dataset (F1-score of 65.70%) at M = 1700. The NER performance on ADE entities at M = 1700 was

the highest when using the *drug product* + *random* dataset (F1-score of 68.13%).

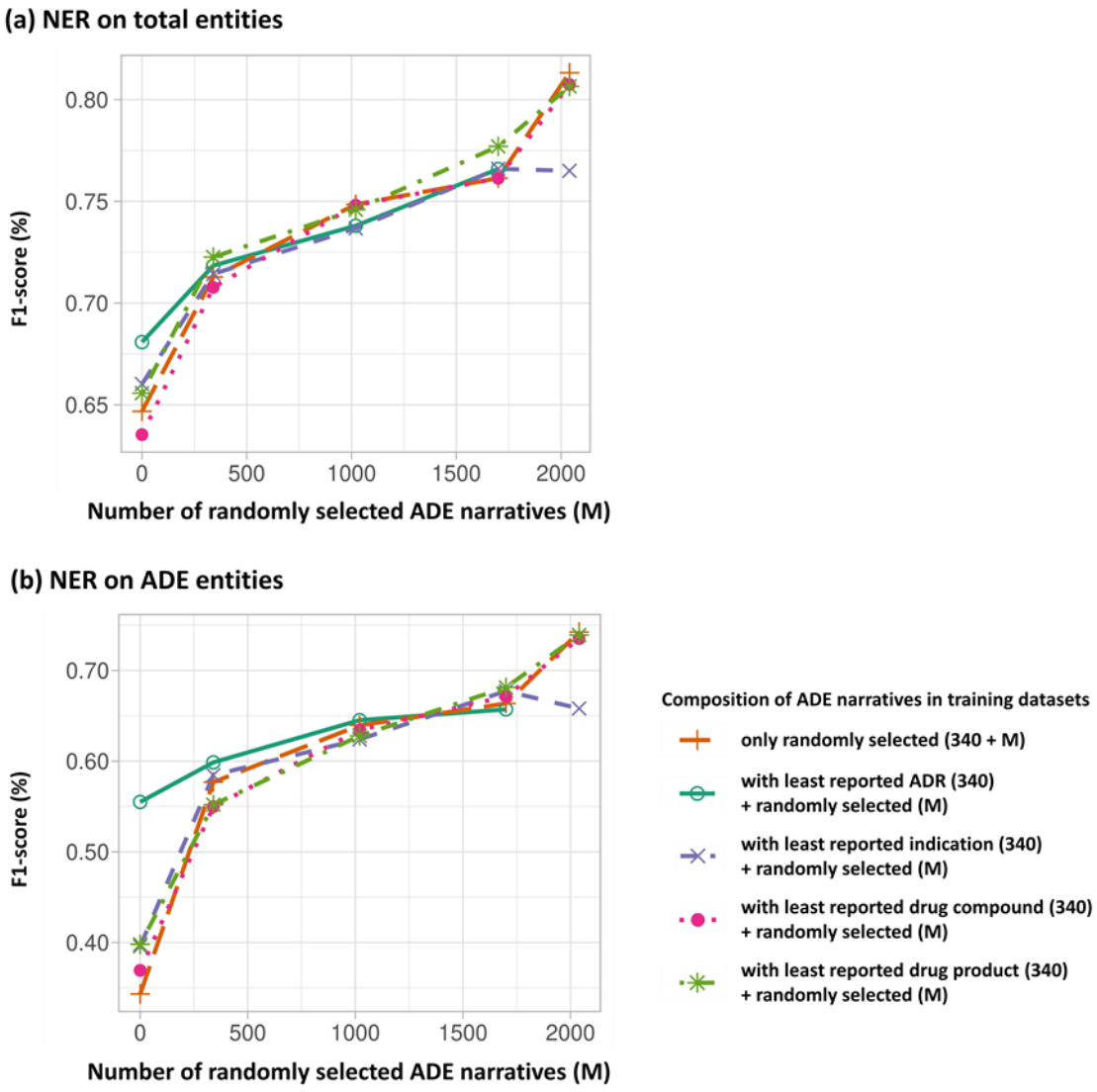

Figure 3.6: NER performances of the KAERS-BERT model on total entities (a) and ADE entities (b) by the composition of training dataset. A *random only* dataset denotes a training dataset consisting of only (340 + M) randomly selected ADE narratives, while *ADE* + *random*, *indication* + *random*, *drug compound* + *random* and *drug product* + *random* datasets represent training datasets consisted of 340 ADE narratives reported with least reported ADE, indication, drug compound, drug product items plus M randomly selected ADE narratives, respectively.

3.4 Discussion

I successfully created an annotated corpus to extract comprehensive drug safety information from ADE narratives reported through SRS, and proposed well-performing baseline models for various NLP tasks. Furthermore, I pre-trained the KAERS-BERT model specialized in clinical texts with frequent code-switching between English and Korea using 1.2 million ADE narratives. The KAERS-BERT model outperformed the KoBERT model on all NLP tasks (Table 3.5). All of those were possible because a thorough and consistent annotation guideline was adopted to extract drug safety information according to the standardized definitions of the data elements used in actual drug safety monitoring as seen in the ICH E2B(R3) standard. As a result, annotation quality was appropriately maintained (Table 3.2).

While previous NLP corpora to extract drug safety information have rarely attempted to capture information other than ADE occurrence and drug dosing information [28, 60, 61, 63, 64], our annotated corpus additionally covered other drug safety information helpful for pharmacovigilance, including the WHO-UMC assessment results and temporal information. In addition, the NER performance of the KAERS-BERT model is comparable to or even higher than those of previous NLP models even though the numbers of used entity types were much larger in this study than in previous studies [28, 60, 62, 68]. The model extracting drug safety information from ADE reports in the VAERS [68] and the best performing model of the MADE 2018 challenge [28] detecting adverse drug events from EHRs achieved F1-scores of 67.35% and 82.9% on the NER, respectively. On the other hand, the KAERS-BERT model I developed resulted in an F1-score of 83.91%. However, the number of used entity types was 19 in this study, which more than doubled and tripled 9 and 6, respectively, in the MADE 2018 challenge and the VAERS study. Given that an NER task becomes difficult as the number of entity types increases, the KAERS-BERT model showed a considerable performance improvement in recognizing word entities related to drug safety information.

The high performance of the KAERS-BERT model on the NER task might be due to three reasons. First, the annotation guideline satisfactorily distinguished each word entity used to express drug safety information from others (see Chapter 7 Appendix, section 7.1). Second, pre-training BERT model using the large volume of ADE narratives might help the model to be tailored to the target domain of drug safety information extraction [88]. Last, ADE narratives mentioned drug safety information more explicitly than other clinical texts of ADEs because the main purpose of an ADE narrative was to describe an adverse event and the clinical settings of a patient experiencing the adverse event.

For example, drug safety information helpful to determine the causality between drug usage and ADE occurrence, such as whether a sign or symptom occurs before or after drug administration, is more explicitly described in ADE narratives than in other clinical texts. Indeed, the proportion of 'ADE' entities correctly predicted was much greater in this study (91.99%) than in the 2018 n2c2 shared task (58.7%) where ADE and medication information was extracted from clinical notes [60]. This larger difference in ADE recognition performance was partly because ADE narratives point out ADEs more explicitly than other clinical texts.

While the KAERS-BERT model achieved decent performances on the NER task, some confusion arose particularly in identifying individual entities (Figure 3.5). Even in this case, if confusion cases can be fixed through rule-based post-processing, they are not fatal to end-to-end systems that can extract drug safety information and save it into the SRS database. In addition, I initially had a concern that the NER performance could be lower on entities labeled as 'occurred' than on entities labeled as 'not occurred' because 'not occurred' entities tended to be written as a list of the most common ADEs of a concerned drug. Unlike our concern, however, the NER performances on 'occurred' entities were comparable to those on 'not occurred' entities (Table 3.6).

Furthermore, I showed that the NER performance was improved when adding ADE narratives with the least reported items to training datasets (Figure 3.6). I postulate that more diverse clinical settings and ADEs in the training dataset with the least reported items added could improve the model performance. The model performances were improved as the sample size of the training dataset increased, and the performance improvement was larger when using training datasets containing ADE narratives with the least reported items than when using the *random only* dataset. As I hypothesized, this finding indicates that diverse clinical settings in ADE narratives with the least reported items could improve the model performance when the model became familiar with the dominant clinical settings represented in randomly selected ADE narratives. The NER performance was better when using the smallest training datasets than when using the *random only* dataset (Figure 3.6). This finding was not totally unexpected in that the validation dataset was also composed of randomly selected ADE narratives.

I expect that the baseline models I developed can improve the data quality of SRS by capturing the drug safety information left out when transmitting ICSRs to the SRS database. For example, I observed that the proportion of entities labeled as ‘Caused hospitalization’ among ‘ADE at the last observation’ was greater in the total ICSRs than in the annotated ADE narratives (Table 3.4). Also, ‘ADE’ entities normalized to psychiatric and immune system disorders were more frequent in the annotated ADE narratives than in KIDS-KD (Figure 3.4). These differences may indicate that certain ADEs and clinical settings, e.g., psychiatric or immunologic ADEs and hospitalization caused by ADEs, tend to be left out when transmitting ICSRs to the SRS database. Because the annotated corpus appropriately captured this information, the NLP models I developed will be capable of extracting the drug safety information left out untapped.

In this study, I defined and annotated the 'ADE Seriousness' entity , but not define a word entity for the severity of ADEs. This decision was mainly due to the fact that the severity of ADEs is not a mandatory reporting information field in the ICH E2B(R3) guideline. In general, the

severity of an ADE refers to the intensity of the ADE or its symptoms. Therefore, severe ADEs, such as severe headaches, may be unrelated to the medical significance of an ADE. On the other hands, the seriousness of an ADE is determined by whether it threatens the patient's life or body, such as causing death, being life-threatening, requiring hospitalization, or resulting in disability [89]. The severity of an ADE is typically assessed using terminologies such as the CTCAE criteria [90]. Since the goal of this study was to increase the reporting fidelity of essential items, I decided to exclude severity in order to increase annotation efficiency and reduce confusion with seriousness. However, in many cases, modifier expressions representing the severity of an ADE are captured as ADE entities, so if the entity normalization step is improved, it is expected that a significant portion of severity information will also be recognized.

I also tested the feasibility of the baseline models developed in this study for extracting drug safety information from free-texts of social networks and web portals, such as Twitter and Naver, using a demo model. The performance of these models was found to be satisfactory. Extracting safety information from new sources, such as social media networks, has the potential to significantly improve the issue of under-reporting [91]. However, it is important to establish appropriate search terms to identify relevant posts containing drug safety information on social media and pre-process irregular spacing and misspellings in the posts. Additionally, caution should be exercised to avoid mistaking previous indications or symptoms as ADEs when extracting drug safety information from social media, as the distinction between ADEs caused by drugs and preexisting symptoms may not be clearly described in the these sources compared to ADE narratives reported through traditional reporting systems.

Our study had several limitations. First, I did not annotate parent-child drug safety information, which is critical to evaluate drug safety during pregnancy. Second, I did not provide baseline models for the entity normalization task and the classification tasks for entity labels except 'occurred' and 'concerned' although I finished annotation for the tasks. Third, the high

performances of the KAERS-BERT model do not guarantee that the model improve the data quality of drug safety information collected through SRS. The usefulness of the NLP model extracting drug safety information depends on the post-processing module that inputs drug safety information into SRS database based on the inference results of the NLP model. To address those limitations, an end-to-end system that could extract drug safety information from ADE narratives and incorporate it into the SRS database may demonstrate the utility of the NLP models in the real-world pharmacovigilance setting [24].Lastly, one potential issue with using NLP models to extract drug safety information is the risk of over-reporting, where existing indications or symptoms are mistaken for ADEs. This can lead to false positive safety signals, which can cause fatigue in pharmacovigilance efforts. In this study, ADEs were defined as any signs or symptoms that occurred after drug administration. This means that drug safety information extracted by NLP models may include a wider range of symptoms than those reported by humans, who may separate ADEs from preexisting indications. To reduce false positive safety signals, I can develop an algorithm that excludes ADEs that are clinically similar to preexisting indications by normalizing medical entities and measuring clinical similarity. Alternatively, I can use a classification model to select only those ADEs that reporters would typically report, based on structured drug safety information from KIDS-KD.

I anticipated that the annotated corpus and NLP models developed in this study would improve pharmacovigilance by reducing under-reporting through the extraction of drug safety information from free-texts. Under-reporting is a significant limitation of SRS and one of the main causes of poor-quality reporting. In the US, only 6% of adverse events are reported to FEARS [23]. However, under-reporting is not only problematic because it leads to a lack of reporting, but also because it can create a reporting bias. For instance, it is known that the reported rate of serious ADEs or ADEs that are expectable for a given drug is higher than that of ADEs that are not [92, 93]. While under-reporting is somewhat inevitable given the nature of SRS, I

believe that by extracting additional drug safety information from free-texts, I can improve under-reporting and related reporting bias. In this study, I also expect to see an effect on improving the reporting bias on the therapeutic area of the ADE, as the distribution of ADEs in terms of system organ class differs between the KIDS-KD and free-texts (Figure 3.4). In addition, while under-reporting has been investigated using expected reports counts estimated using literature review [94, 95], obtaining drug safety information contained in free-texts can help estimate the reporting rate through the final information entry stage of the SRS.

3.5 Conclusion

In summary, I defined the extraction of comprehensive drug safety information from ADE narratives reported through SRS as a series of NLP tasks and successfully developed well-performing baseline NLP models for the tasks. Specifically, I developed the KAERS-BERT model suited for clinical texts written in Korean and English using 1.2 million ADE narratives collected through KAERS. The KAERS-BERT model also outperformed other baseline models including the KoBERT on all the NLP tasks. The annotated corpus and the KAERS-BERT model can streamline pharmacovigilance activities and eventually increase their efficiency by improving the data quality of SRS database.

Chapter 4

Extraction of Drug-Food Interactions from the Abstracts of Biomedical Articles

Although drug-food interaction (DFI) is a drug interaction that has posed a threat to the safe usage of medicine next to drug-drug interaction (DDI), there is no database that systematically collects DFI information. Thus, in this study, I tried to define the extraction of DFI information from abstracts of biomedical articles for building DFI database as an NLP task. In addition, I developed manually annotated corpus for the DFI extraction, i.e., 'the DFI corpus' and provided baseline models for the defined tasks through simple fine-tuning of diverse BERT models.

4.1 Motivation

Drug interaction occurs when the exposure to a drug (or *victim*) or its efficacy and/or safety is affected by another substance (or *perpetrator*) consumed together with the drug. Perpetrators include, but are not limited to, another drug, food, beverage, or a chemical. Drug interaction may increase or decrease the activity of the victim drug. For example, grapefruit can increase the blood pressure-lowering effect of some anti-hypertensive when ingested together [96]. In contrast, atorvastatin can reduce the efficacy of co-administered clopidogrel, which is used to

manage unstable angina [97].

Of various types of drug interactions, those with another drug (drug-drug interaction, DDI) or food (drug-food interaction, DFI) are clinically important because DDIs and DFIs can be avoided if they were recognized beforehand. Drug interactions can be identified at various stages of drug development, post-marketing pharmacovigilance, and routine clinical use through prospective systematic investigations or by accident. No matter how drug interactions are identified, a drug interaction database can assist clinicians to choose a set of medications, which can be safely co-administered to avoid harmful drug interactions.

An NLP system that automatically extracts drug interaction information from biomedical texts can facilitate the construction of a drug interaction database. For example, several DDI extraction NLP models developed on the DDIExtraction 2013 corpus have shown good performance in extracting and classifying DDIs [4, 98, 99]. A similar NLP model to extract DFIs from scientific literature was trained on the DDIExtraction 2013 corpus on the assumption that the structure of DDI-describing sentences is close to that of DFI-describing ones [100]. This assumption, however, may not be substantiated given the differences between DFIs and DDIs in the objectives, designs, and evidence levels of studies investigating drug interactions [101]. To support this notion, many DFIs are identified incidentally and communicated typically via case reports, whereas most DDIs are affirmed or nullified in well-designed, prospective *in-vitro* or *in-vivo* studies. Therefore, 'extraction task' should be separately approached between DFIs and DDIs. In other words, the sources of corpora for extraction task should vary by the type of drug interactions.

At the time of writing this article, the POMELO corpus was the only publicly available corpus for DFI extraction [102] (Table 1). The POMELO corpus consists of abstracts of biomedical articles indexed by the Medical Subject Heading (MeSH) term 'Drug-Food

Interaction' (MeSH Unique ID: D018565). However, many DFIs have been also reported in a study on the biological effects of a food or its components, which were not necessarily indexed by the MeSH term 'Drug-Food Interaction'. Thus, it is not a good idea to limit the DFI corpus to texts explicitly referring to a DFI, such as texts from the 'Drug interaction' section in DrugBank or biomedical articles with the 'Drug-Food Interaction' MeSH Term.

Table 4.1: Biomedical corpora developed for the extraction of drug interaction

Drug interaction type	**Corpus**	**Data source**	**Numbers of annotated sentences and documents**	**Annotated information**
DDI	DDI corpus [99]	DrugBank MEDLINE	4701/792 327/233	types of entities, relations between entities, types of DDI statements
	PK-DDI [103]	Drug package inserts	592/68	modality of DDI, types and roles of drug entities
	PK [104]	MEDLINE	1333/428	types of entities
DFI	POMELO [102]	MEDLINE	1084/639	types of entities, relations between entities
	DFI corpus (this study) [13]	PubMed	2498/2270	types of entities, sentences containing DDI or DFI, modality of DFI statements, relations between entities, evidence level of abstracts

Based on that understanding, I newly created an annotated corpus for DFI extraction from the abstracts of biomedical articles, i.e., *the DFI corpus*. To this end, I have broadened the definition of DFI such that it is not only the change in the safety and efficacy profile of a drug,

but indirect interactions between a food and a drug via biological components or pathways. Additionally, I have expanded the concept of DFI to include the interaction between a food and molecules closely related to the pharmacological action and pharmacokinetic properties of the drug. This expansion has enabled us for identifying more DFIs, many of which were not prospectively studied, and, therefore, should be suspected based on a small number of case reports and our understanding on the molecular mechanisms.

The annotated DFI corpus can be useful to construct a DFI database for healthcare professionals such as clinicians and pharmacists.[4] This is why the DFI corpus contains not only a DFI key-sentence and an extractive summarization of DFI, but also the word entities related to the extracted DFI, the relations between the annotated entities, and the evidence levels of the extracted DFI. Eventually, the DFI corpus is also suited for the multi-task learning approach because the classification tasks defined in the DFI corpus are related [105].

To assess the appropriateness of the DFI corpus, I established a baseline performance for DFI extraction using a BERT (Bidirectional Encoder Representations from Transformers) model. BERT is one of the most successful and widely-used language models in language understanding tasks [9]. However, BERT models pre-trained on general domains are not adequate for text mining tasks in the biomedical area because the word distributions and semantics in biomedical texts are quite different from those in general texts. For this reason, various BERT models, including BioBERT [56], ClinicalBERT [57], and PubMedBERT [106], have been introduced to facilitate language understanding tasks in the biomedical domain. In this study, I used those BERT models pre-trained on biomedical texts to report the performance benchmarks.

In brief, our main contributions are as follows:

1. I constructed ‘the DFI corpus’, a manually annotated corpus for extracting DFI

[4] Available here: https://github.com/ccadd-snu/corpus-for-DFI-extraction

information from the abstracts of biomedical articles. I identified a 'DFI key-sentence' as a target entity for DFI extraction.

2. The DFI corpus I created is the largest in its kind, i.e., manually annotated corpora for extracting drug interactions.
3. Simple fine tuning of the pre-trained BERT models using the DFI corpus performed remarkably well in the classification tasks for DFI extraction.

4.2 Proposed Methods

In this section, I explained how I selected biomedical articles annotated in the DFI corpus and described an outline of annotation guideline I developed. In addition, I elaborated the methods for the quality control of annotation and baseline model development.

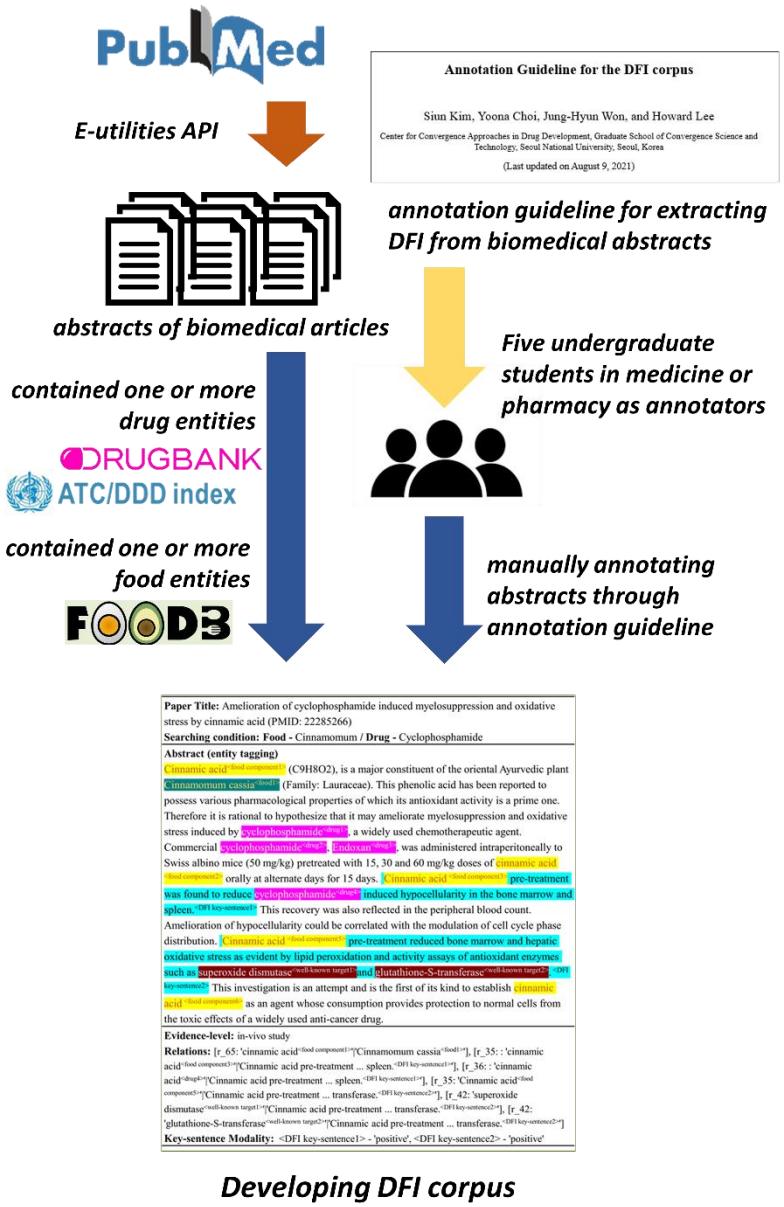

Figure 4.1: Overview of proposed method for developing DFI corpus

4.2.1 Data source

I collected the abstracts of biomedical articles from the journals that have published one or more papers between January 1, 1970 and October 2, 2019, indexed by the MeSH term of 'Drug-Food

Interaction' (MeSH Unique ID: D018565) or ‘Herb-Drug Interaction’ (MeSH Unique ID: D041743) (Supplementary data 1). MeSH terms, a hierarchical subject vocabulary for biomedical documents, are annually updated by the National Library of Medicine in the US [107]. I used the Entrez Programming Utilities (E-utilities, last updated on June 24, 2015; the National Center for Biotechnology Information, Bethesda, MD, USA) to collect the abstract texts from biomedical articles and their metadata [108]. All the annotated abstracts were accessible through PubMed.

To increase the number of the abstracts containing DFI information in the DFI corpus, I selected abstracts that simultaneously included both a food word and a drug/drug-related molecule word (i.e., well-known targets, drug metabolizers, and drug transporters). A word was regarded as a food, drug, or drug-related molecule word when the word is listed in the manually-curated vocabularies of foods, drugs, and drug-related molecules, each of which was based on the FooDB [109], the 2019 ATC index and the DrugBank, respectively (Supplementary data 2). Some words were used to denote both food and drug/drug-related molecule depending on the context, e.g., ascorbic acid (vitamin C) and Zn (zinc). This manual annotation enabled for deciding the category for a word in a certain context.

4.2.2 Annotation

Annotation for the DFI corpus was conducted in the following five steps: (1) labeling the type of a word entity, (2) labeling the type of a sentence entity, (3) tagging a relation between word entities or between a word and sentence entities, (4) tagging a sentence modality for key-sentence entities, and (5) assigning an evidence-level to an abstract document. Each step is described below in more detail.

(1) I classified words into one of the following eight entities: ‘drug’, ‘food’, ‘food component’, ‘ambiguous’, ‘well-known target’, ‘drug metabolizer’, ‘drug transporter’,

and ‘none’. Word lists are constructed for each entity category using the DrugBank, the 2019 ATC index, and FooDB. A word could be labeled as >1 entity depending on the context. For example, insulin injected into the body was labeled as a ‘drug’, which is a medicine or substance used to treat or prevent a disease or alleviate its symptoms. At the same time, insulin was labeled as a ‘well-known target’ when its level was determined to evaluate the physiological effect of a chemical compound. Detailed guidance and examples for labeling word entities can be found in the annotation guideline (See Chapter 7 Appendix, section 7.2).

(2) I labeled a sentence as one of the following five sentence entities: ‘DFI key-sentence’, ‘DDI key-sentence’, ‘food-effect key-sentence’, ‘supporting sentence’, and ‘none sentence’. A ‘DFI key-sentence’ refers to a sentence that contains DFI information, while a ‘DDI key-sentence’ is a sentence that represents DDI information. A ‘food-effect key-sentence’ denotes a sentence that provides information as to how food intake affects the bioavailability of a drug [110]. Lastly, a ‘supporting sentence’ does not provide information about the occurrence of DFI or DDI by itself, but must be read in advance to understand the DFI or DDI key-sentence that follows.

(3) I tagged a relation between the word entities. For example, apple and citric acid are a food and its food component, respectively. Synonyms or abbreviations are identified such as ‘green tea extract’ and ‘GTE’. Furthermore, I tagged a relation between a word and sentence entities that were associated with a labeled DFI key-sentence. Thus, a ‘DFI key-sentence’ should have at least two tagged relations, one with a drug-related entity and the other with a food-related entity.

(4) I determined a DFI key-sentence was ‘positive’ if the study described in the abstract showed an interaction between a food and drug entities, and ‘negative’ if not.

(5) I assigned one of the following seven evidence-levels of DFI information to each document: ‘clinical trial’, ‘observational study’, ‘case study’, ‘in-vivo study’, ‘in-vitro study’, ‘bioanalysis’, or ‘others.’

More detailed instructions and examples can be found in the annotation guideline for the DFI corpus (see Appendix, section 7.2). The annotation was performed using *tagtog* (tagtog.net), an web-based annotation tool [111].

Paper Title: Amelioration of cyclophosphamide induced myelosuppression and oxidative stress by cinnamic acid (PMID: 22285266)

Searching condition: Food - Cinnamomum / **Drug -** Cyclophosphamide

Abstract (entity tagging)

Cinnamic acid<food component1> (C9H8O2), is a major constituent of the oriental Ayurvedic plant Cinnamomum cassia<food1> (Family: Lauraceae). This phenolic acid has been reported to possess various pharmacological properties of which its antioxidant activity is a prime one. Therefore it is rational to hypothesize that it may ameliorate myelosuppression and oxidative stress induced by cyclophosphamide<drug1>, a widely used chemotherapeutic agent. Commercial cyclophosphamide<drug2>, Endoxan<drug3>, was administered intraperitoneally to Swiss albino mice (50 mg/kg) pretreated with 15, 30 and 60 mg/kg doses of cinnamic acid <food component2> orally at alternate days for 15 days. Cinnamic acid <food component3> pre-treatment was found to reduce cyclophosphamide<drug4> induced hypocellularity in the bone marrow and spleen.<DFI key-sentence1> This recovery was also reflected in the peripheral blood count. Amelioration of hypocellularity could be correlated with the modulation of cell cycle phase distribution. Cinnamic acid <food component5> pre-treatment reduced bone marrow and hepatic oxidative stress as evident by lipid peroxidation and activity assays of antioxidant enzymes such as superoxide dismutase<well-known target1> and glutathione-S-transferase<well-known target2>. <DFI key-sentence2> This investigation is an attempt and is the first of its kind to establish cinnamic acid <food component6> as an agent whose consumption provides protection to normal cells from the toxic effects of a widely used anti-cancer drug.

Evidence-level: in-vivo study

Relations: [r_65: 'cinnamic acid<food component1>'|'Cinnamomum cassia<food1>'], [r_35: : 'cinnamic acid<food component3>'|'Cinnamic acid pre-treatment ... spleen.<DFI key-sentence1>'], [r_36: : 'cinnamic acid<drug4>'|'Cinnamic acid pre-treatment ... spleen.<DFI key-sentence1>'], [r_35: 'Cinnamic acid<food component5>'|'Cinnamic acid pre-treatment ... transferase.<DFI key-sentence2>'], [r_42: 'superoxide dismutase<well-known target1>'|'Cinnamic acid pre-treatment ... transferase.<DFI key-sentence2>'], [r_42: 'glutathione-S-transferase<well-known target2>'|'Cinnamic acid pre-treatment ... transferase.<DFI key-sentence2>']

Key-sentence Modality: <DFI key-sentence1> - 'positive', <DFI key-sentence2> - 'positive'

Figure 4.2: Example of a manually annotated abstract for DFI extraction. Before annotating DFI, I selected abstracts that contained ≥1 drug word AND ≥1 food word simultaneously in the same abstract. The entity types of annotated words are denoted by superscripts and highlighted in colors. In this example, two DFI key-sentences marked as light blue are annotated. Also, the relations between word entities or between a word and sentence entities were enumerated at "Relations" at the bottom, while the modalities of key-sentence entities at "Key-sentence Modality." The evidence-level of the abstract, 'in-vivo study', was described at "Evidence-level."

4.2.3 Quality control of annotation

The DFI corpus was manually annotated by five annotators, and their annotations were reviewed by an independent reviewer, who had expertise in both clinical pharmacology and NLP. The annotators had at least ≥2 years of study in medicine or pharmacy, and had a good command of English. I trained the annotators for two weeks to help them become familiar with the annotation guideline, to teach them how to use tagtog, and to fix any mistake in preliminary annotation exercises.

After two weeks of training, each annotator labeled a hundred abstracts per week. To control and improve the annotation quality, the independent reviewer gave real-time feedbacks to the annotators one-on-one. The independent reviewer also thoroughly examined and scored the annotation results by each annotator on a weekly basis. If the proportion of correct answers was <80% in randomly selected 20% sentences/abstracts, the annotator was asked to re-do the entire annotation task for the week.

The inter-annotator agreement (IAA) was measured by comparing the annotation results between each of the five annotators and the independent reviewer through the Cohen's kappa coefficient (κ) defined by [81]. IAA was assessed for three classification tasks defined in the DFI corpus: (1) named entity recognition, (2) key-sentence classification and (3) evidence-level classification.

4.2.4 Baseline model development

I built baseline models for DFI extraction using the BERT-Base model and publicly available BERT models pre-trained on biomedical texts such as BioBERT, ClinicalBERT, PubMedBERT [9, 106, 112, 113]. All of the tokenizers and model configurations of those BERT models were

directly downloaded from the Huggingface's repository.[5] The fine-tuning of the BERT models was performed using simple fully connected networks of two layers. I performed hyperparameter tuning on the learning rate, drop-out probability, batch size and epoch number. I evaluated the baseline models by measuring F1 scores and performed the qualitative error analysis.

4.3 Results

4.3.1 Corpus statistics

A total of 2,270 abstracts were collected, from which 2,498 sentences were extracted to create the DFI corpus (Table 4.1). Most (94.6%) of the abstracts containing DFI key sentences were not indexed by a DFI MeSH term (Table 4.2). The DFI corpus contained 33,386 (6.0%) word entities out of 552,371 words, and the most common word entity was food (13,353, 40.0%) followed by well-known target (8,467, 25.3%) and drug (8,088, 24.2%). (Figure 4.2a and Table 4.3). Of 20,550 sentences in the DFI corpus, 2,488 sentences (12.1%) included information about DDI or DFI, of which 2,145 (85.9 %) and 211 (8.4%) sentences were DFI and DDI key-sentences, respectively. (Figure 4.3b)

Table 4.2: Frequency table for annotated abstracts

Included DFI key-sentence	**Indexed by DFI MeSH term**	
	Yes	No
Yes (N=1001)	54 (5.4%)	947 (94.6%)
No (N=1269)	8 (0.6%)	1261 (99.4%)

[5] https://huggingface.co/models

Total (N=2270)	62 (2.7%)	2208 (97.3%)

* MeSH, medical subject headings; DFI, drug-food interaction.

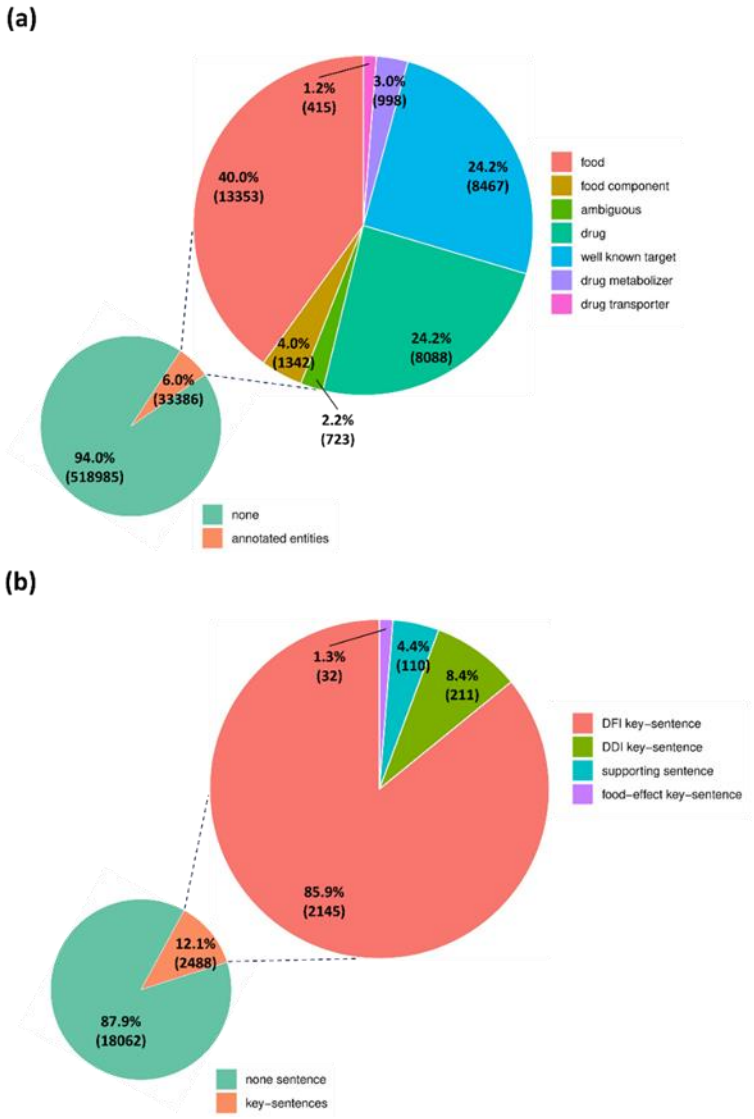

Figure 4.3: Distribution of (a) annotated word and (b) sentence entities in the DFI corpus.

Table 4.3: Distribution of the annotated entity types in the DFI corpus

Entity type	**Percent of abstracts, n (%)**

	Training	Development	Test
'drug'	5,632 (1.46)	1,669 (1.48)	787 (1.43)
'food'	9,384 (2.44)	2,621 (2.33)	1,348 (2.45)
'food component'	902 (0.23)	377 (0.34)	63 (0.11)
'ambiguous'	452 (0.12)	118 (0.10)	153 (0.28)
'well-known target'	6,065 (1.58)	1,723 (1.53)	679 (1.24)
'drug metabolizer'	697 (0.18)	176 (0.16)	125 (0.23)
'drug transporter'	288 (0.07)	113 (0.10)	14 (0.03)
'none'	361,545 (93.92)	105,688 (93.96)	51,752 (94.23)
total	**384,965 (100.0)**	**112,485 (100.0)**	**54,921 (100.0)**

In terms of evidence levels, the majority of the abstracts were from in-vitro or in-vivo studies regardless of the inclusion of a DFI key-sentence (75.6% and 56.2%, respectively, Figure 4.4). On the contrary, only 7.9% of abstracts including a DFI key-sentence were associated with clinical studies (i.e., case report, clinical trial and observational study) with only 3.1% being from clinical trials.

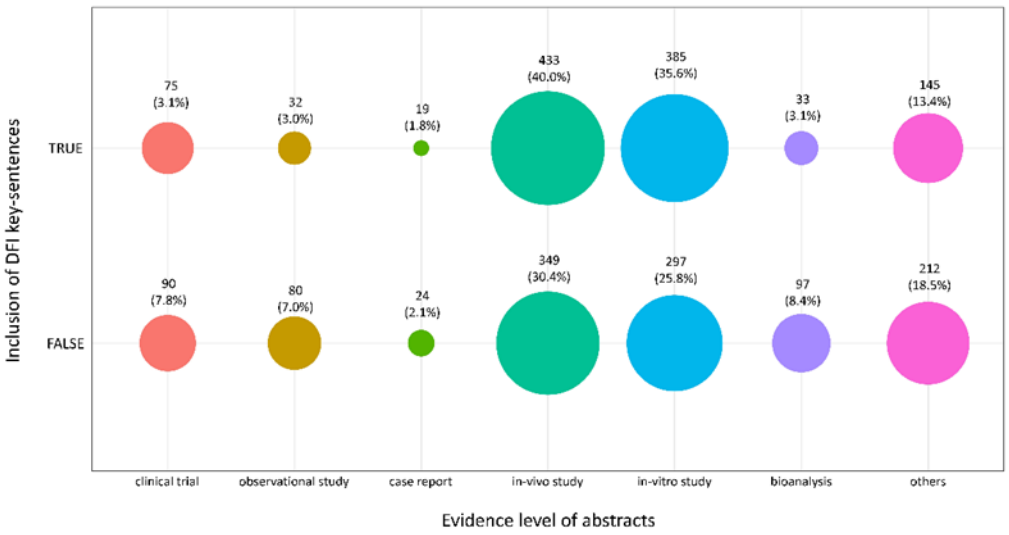

Figure 4.4: Distribution of evidence levels of abstracts in the annotated corpus by the inclusion of DFI key-sentence. The size of the circles is proportional to the number of abstracts and the exact numbers and percentages are also reported.

Many words were labelled as >1 entity, particularly for food, drug, and well-known target (Figure 4.5). For examples, while coffee and cisplatin were consistently labeled as a food and a drug entity, respectively, cannabidiol and iron were annotated as a food or a drug depending on the context.

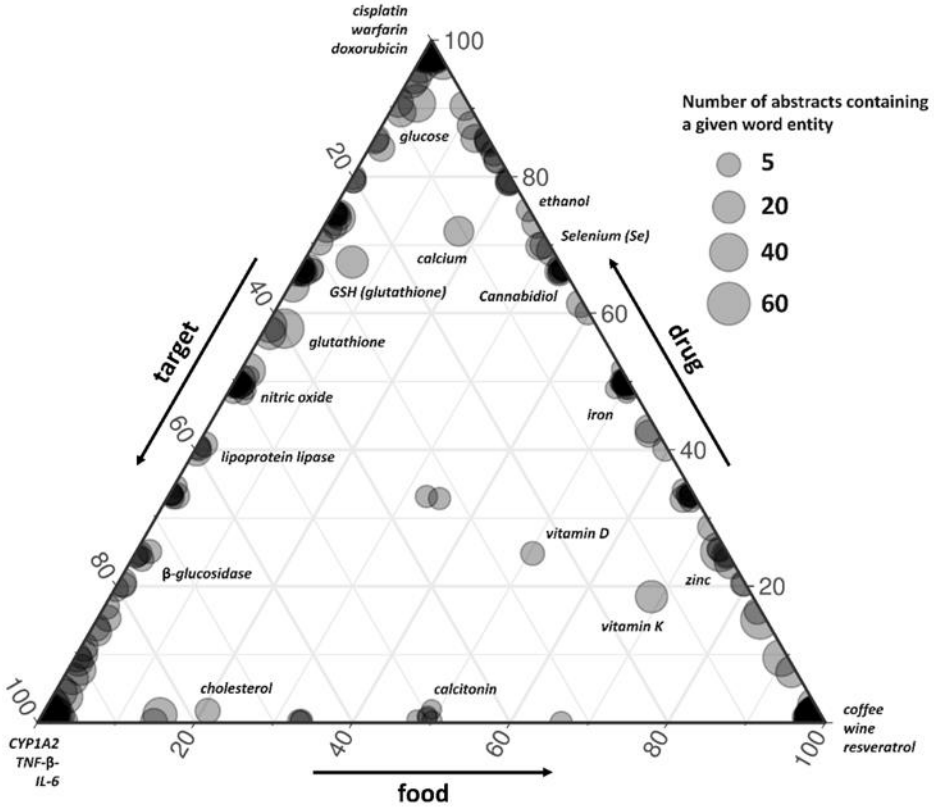

Figure 4.5: Ternary plot showing the ratios of annotated entity types of given words in the DFI corpus. In this plot, 'food' denotes both 'food' and 'food component' entities, while 'drug' implies both 'drug' and 'ambiguous' entities. Also, 'target' includes 'well-known target', 'drug metabolizer', and 'drug transporter' entities.

4.3.2 Annotation Quality

The total IAA was >0.845 for key-sentence classification and evidence-level classification tasks between the annotators with the independent reviewer, while the average of total IAAs was 0.797 (Table 4.4). For name entity annotation, only three annotators finished annotation schedule and the total IAA of the three annotators was 0.638.

Table 4.4: Cohen's kappa by classification task between the annotators and the independent reviewer

Classification task	Annotator					total
	1	2	3	4	5	
Named entity annotation	0.797	NA	NA	0.545	0.585	0.638
Key-sentence classification	0.910	0.878	0.591	0.887	0.920	0.865
Evidence level classification	0.913	0.869	0.669	0.935	0.935	0.888
Average	0.873	0.798	0.619	0.789	0.813	0.797

* NA, not available;
* #2 and #3 annotators did not finish the pre-defined annotation schedule for named entity recognition.

4.3.3 Performance of baseline models

All of the BERT models fine-tuned on the DFI corpus performed well in the classification tasks, particularly for named entity recognition and key-sentence classification (F1 scores > 83%, Table 4.5). In all three tasks, the PubMedBERT and BioBERT models showed the best and second-best performances [56, 106].

Table 4.5: Performance of BERT models by DFI extraction task. The bolded and underlined performance scores indicate the best and second-best performances on a classification task, respectively.

Classification task (%)	Base-BERT	BioBERT	PubMedBERT	ClinicalBERT
Named entity recognition				
weighted F1 score	83.1	<u>85.2</u>	**86.1**	82.3
macro F1 score	80.0	<u>83.1</u>	**83.8**	79.3

Key-sentence classification	82.0	82.6	**85.1**	81.4
Evidence level classification				
weighted F1 score	70.6	**72.8**	70.4	67.3
macro F1 score	61.9	**65.6**	63.1	53.6

4.3.4 Qualitative error analysis

As a result of qualitative error analysis, the most likely true positive sentences in the validation dataset which were classified as a key-sentence by both annotators and the key-sentence classifier are syntactically typical to express DFI (Table 4.6). The typical syntax of sentence describing DFI is "<food, food compound or related abbreviations)> significantly <change, increase, decrease etc.> <the exposure of drug, drug-induced toxicity or the expression of gene in the mode of action of interested drug>." The most unlikely true false sentences seem to be irrelevant to a study investigating DFI (Table 4.7).

Table 4.6: Examples of most likely true positive sentences in the validation dataset which were labeled as a key-sentence and also predicted as a key-sentence by the key-sentence classifier

Examples of most likely true positive sentences
- Pretreatment with 23-HTA (10 mg/kg/d, per os (p.o.)) significantly reduced cisplatin-induced elevations in blood urea nitrogen (BUN) and serum creatinine level, whereas NIF_1 (10 mg/kg, p.o.) slightly reduced these levels. [114]
- Co-administration of berberine increased the initial plasma concentration and AUC of metformin and decreased systemic clearance and volume of distribution of metformin in rats, suggesting that berberine inhibited disposition of metformin, which is governed by OCT1 and OCT2. [115]
- Moreover, green tea significantly inhibited OATP1A2-mediated nadolol uptake (half-maximal inhibitory concentration, IC50 = 1.36%). [116]

- Furthermore, BA ameliorated mRNA and protein expression of NF-κB, iNOS, TNF-α, Nrf2, HO-1 and NQO1 in the kidney. [117]
- Ginger significantly decreased the area under the concentration-time curve of isoniazid, whereas Vz and Cl were increased. [118]

Table 4.7: Examples of most likely true positive sentences in the validation dataset which were labeled as a non key-sentence and also predicted as a non key-sentence by the key-sentence classifier

Examples of most likely true negative sentences
- The morphologies of Laminaria were studied by scanning electron microscopy and transmission electron microscopy. [119]
- Whole flour samples were analyzed by ICP-AES/MS, HPLC and Elemental CHNS Analyzer. [120]
- Technological advances in the past 30 years have enabled the production of pure, stable proteins in vast amounts. [121]
- Here, a randomized plot field experiment was performed to study the GHG emissions for various farming systems during the rice growing season. [122]
- Allogeneic stem cell transplantation (alloSCT) is a curative procedure for myelofibrosis. [123]

On the other hand, the most unlikely false negative sentences which were labeled as a key-sentence but predicted as a non key-sentence by the sentence classifier were frequently wrongly parsed by the sentence tokenizer. Thus, the most unlikely false negative sentence actually contained more than one sentence or was an incomplete sentence. In the most unlikely false positive sentences, syntactic features were similar to those of DFI key-sentences. However, these false positive sentences were related to only drug entities or it is unlikely to describe DFI in an given experiment setting of a source document, *i.e.*, abstract of biomedical article. Also, abbreviated drug or gene names seemed to be often mistaken for a food entity.

Table 4.8: Examples of most unlikely false negative sentences in the validation dataset which were labeled as a key-sentence but predicted as a non key-sentence by the sentence classifier were

Examples of most unlikely false negative sentences
- TRB was also effective against oral toxicity of BoNT/A, B and E. Thus, TRB may be of potential benefit in protecting the paralytic actions of botulinum neurotoxins (BoNTs), but its use is limited by mixing with the toxin. [124]
- Such an inhibitory effect could be due to reduced levels of S.
- The leaf extract had moderate anti-elastase activity (54%) but was inactive agains collagenase. [125]
- Limited data suggest that thiamine supplementation is capable of increasing left ventricular ejection fraction and improving functional capacity in patients with heart failure and a reduced left ventricular ejection fraction who were treated with diuretics (predominantly furosemide). [126]
- CYP3A4, raising the question which effect prevails in vivo.

Table 4.9: Examples of most unlikely false positive sentences in the validation dataset which were labeled as a non key-sentence but predicted as a key-sentence by the sentence classifier

Examples of most unlikely false positive sentences
- However, a combination of ICI182780 and MK886 significantly inhibited resveratrol-induced eNOS mRNA expression. [127]
- In addition, it significantly up-regulated the level of t-PA and down-regulated the level of PAI-1 ($p<0.05$). [128]
- HG enhanced expression of fibrosis biomarkers such as collagen IV and connective tissue growth factor (CTGF), which was markedly attenuated by Oryeongsan. [129]
- In addition, ROS levels in bladder tissues and serum lipid peroxidation (TBARS assay) were markedly higher in obese compared with lean mice, all of which were reduced by resveratrol treatment. [130]

- Gelatin caused a decrease in thrombin-antithrombin complexes (-45% vs. -4%, $p<0.05$) and F1+2 (-40% vs. +1%, $p<0.05$). [131]

4.4 Discussion

I successfully created a new, manually-annotated, domain-specific corpus to extract DFI information from biomedical texts, which I named 'the DFI corpus'. To the best of our knowledge, the DFI corpus is largest and most comprehensive in its kind. For example, the numbers of sentences and documents are two and 3.5 times, respectively, as many as those in POMELO (Table 4.1). Furthermore, the number of documents annotated in the DFI corpus is much bigger than that in any other corpora on DDI (2,270 vs. 792 in the DDI corpus, Table 4.1).

Besides, the DFI corpus is well-suited for various NLP tasks in DFI extraction. This versatility has been made possible by including not only word and sentence entities, but the evidence-level of the extracted DFIs and the relations between annotated entities, such as synonym relation between a food and its abbreviation form, and inclusion relation between a food and its component. Furthermore, what is learned from each classification task defined in the DFI corpus is helpful to learn other types of NLP tasks. For example, the ability to recognize food and drug word entities acquired from the named entity recognition task in the DFI corpus could improve the performance of identifying DFI key-sentences. Thus, I expect that the multi-task learning framework incorporated into the DFI corpus will improve the performance of any classification model trained or fine-tuned on the DFI corpus [105]. This expectation appears to be realized if I look at the remarkable performance in the classification tasks achieved by the BERT models, particularly the PubMedBERT and BioBERT models (Table 4.5).

The inclusion of both DDI and DFI key-sentences in the DFI corpus can improve the performance of any models trained on the DFI corpus. In the DFI corpus, I annotated not only DFI but DDI key-sentences because they will help differentiate the syntax of DFI key-sentences

from that of sentences not containing any drug interaction information (i.e., none key-sentences). As the definitions of 'interaction' were identical when describing DFI and DDI (Supplementary data 3), the syntactic features of DFI and DDI key-sentences are very similar. Therefore, if DDI key-sentences are separately annotated and labeled, a model could have more chance to learn how to distinguish DFI key-sentences from none key-sentences by utilizing the syntactic information of DDI key-sentences as well. This flexibility was realized by including both DFI and DDI key-sentences in the DFI corpus, which eventually allows for identifying the structural difference between sentences describing DFI and DDI.

Unlike the previous drug interaction corpora that exclusively relied on MeSH terms in selecting abstracts, the DFI corpus was constructed by selecting biomedical abstracts that included both a food word and a drug/drug-related molecule word. This resulted in a huge increase in the number of the abstracts containing DFI information in the DFI corpus. For example, 44.1% of annotated abstracts (1,001 out of 2,270 abstracts, Table 4.2) contain one or more DFI key-sentences. In fact, I found out that the vast majority of abstracts containing DFI key-sentences could not be tagged when searched only by the DFI MeSH term, i.e., only 5.4% or 54 out of 1,001 abstracts (Table 4.2). This is why DFI extraction models trained on the POMELO corpus, which consists of abstracts of biomedical articles tagged only by the DFI MeSH term, might not be practical to capture DFI from abstracts of biomedical articles. Indeed, the DFI corpus has many more 'negative' sentences (18,062 sentences, Figure 4.3b), which appeared around DFI key-sentences but did not contain evidence for DFI, while the total number of 'positive' sentences describing drug interactions are comparable to or even larger than those of previous corpora [99, 102].

In addition, the concept of 'ambiguous' word entity was introduced in this corpus to address the difficulty in determining whether a word refers to a food or drug entity. Distinguishing between these two entities is also important for differentiating DFI from DDI. However, despite

drug and food entities making up 40.0% and 24.2% of annotated word entities, respectively, ambiguous entity made up only 2.2% of all word entities (Figure 4.3). This clear distinction between food and drug entities may be due to the fact the only documents containing one or more drug and food words on manually curated, mutually exclusive lists were used as annotation targets in this corpus. For instance, I selected one document for annotation [114] with the title "*The ameliorative effect of 23-hydroxytormentic acid isolated from Rubus coreanus on cisplatin-induced nephrotoxicity in rats"* because it contained a drug word 'cisplatin' and a food word 'rubus'. These pre-recognized words were served as cues for the annotators to distinguish drug and food entities, making it easy for them to identify clear examples of drug and food entities.

Another strength of the DFI corpus is that it was annotated based on the scientific understanding of the design of the study investigating DFI and the context where a food or entity was used. Those annotation rules are particularly useful for distinguishing a drug entity from a food entity while not confusing DDIs with DFIs. For example, *'cholesterol'*, which usually represents a target entity, was labeled as a food entity if used to denote a food component. To support this notion, the overall distribution of labeled entity types was plausible (Figure 4.5).

The DFI key-sentences in the DFI corpus are not just statements as a sentence level, but complex entities that provide a context each document denotes. I labeled a DFI key-sentence only when the *factual* evidence of the DFI information is presented in a given abstract (see Appendix, section 7.2). For example, a sentence like *"According to previous studies, green tea may interact with cardiovascular drugs such as warfarin, simvastatin and nadolol."* was not tagged as a DFI key-sentence because an annotated abstract did not provide any factual evidence to support the acclaimed DFIs. This is why the sentence entities in the DFI corpus can provide more information than other corpora as I also labeled the evidence level of the documents that contained each key-sentence.

The whole steps of annotation to create the DFI corpus was well planned and performed adequately. The detailed annotation guideline was prepared in full advance of actual annotation (see Appendix, section 7.2). Additionally, I carefully chose qualified annotations, trained them, gave feedbacks, and fixed any error on a real-time basis, which dramatically improved the quality of the DFI corpus. Evidence is that the average of total IAAs on the three classification tasks was 79.6%, which is higher than or comparable to IAAs reported previously in other well-annotated medical corpora [61, 132]. The total IAAs for classifying key-sentences and evidence-levels (86.5% and 88.8%, respectively, Supplementary table 1) were even 15% higher than the highest IAAs for annotating relational information from MedLine corpus in the DDI corpus (72.5%) [99]. This study had a major limitation. The total IAA for named entity recognition in the DFI corpus was lower than that in the DDI corpus (63.4% and 79.6%, respectively, Table 4.4) [99]. This might be because the named entities in the DFI corpus were more diverse and heterogeneous than those in the DDI corpus [99]. Unlike the DDI corpus, where only a 'drug' and its subtype entities (e.g., brand name and group of drugs) were included as named entities, more diverse and heterogeneous entities were contained in the DFI corpus (i.e., 'drug', 'food', 'food component', 'ambiguous', 'well-known target', 'drug metabolizer', 'drug transporter', and 'none'). In spite of this limitation, however, the overall quality of annotation in our study cannot be overestimated because the average IAA for all of the classification tasks using the DFI corpus was sufficiently high (79.6%, Supplementary table 4), and all the BERT models fine-tuned on the DFI corpus performed well in the classification task for named entity recognition (F1 scores >83%, Table 3).

In addition, I found that the sentence tokenizer for baseBERT did not successfully parse the sentences in the abstracts of biomedical article and wrongly parsed sentences had a bad impact on the key-sentence classification. Moreover, drug and gene names arbitrarily abbreviated in a given abstract seemed to be often mistaken for a food entity. I expect that these problem could be solved by using sentence tokenizers specialized in the biomedical domain and replacing

abbreviated words by their definitions in preprocessing.

In this study, I discuss the development of the DFI corpus which focuses on systemically collecting scientific evidence on DFIs. Unlike other annotated corpora for drug interaction extraction that focused on detecting sentences describing drug interactions, the DFI corpus defines and selects DFI key-sentences that experimentally support the existence (or non-existence) of DFIs. The dataset and model developed in this study have the potential to enable the systemic collection of DFI information as well as its reference papers. The relatively low performance on extracting key-sentences may be due to the fact that DFI key-sentences are determined based on the experimental design in the abstract rather than the content of the sentence itself. This suggests that NLP models may need to have some understanding or reasoning ability about biomedical experiment design in order to properly extract DFI key-sentences. Therefore, future work could explore the use of the broader contextual information from the experimental settings of the abstracts to improve the performances of DFI key-sentence extraction beyond the baseline model in this study, which only used individual sentence representation as input.

4.5 Conclusion

I constructed the DFI corpus, the largest and the most comprehensive corpus in its kind to extract drug interaction, based on the proper pharmacological understanding about DFI and domain-specific knowledge as to how scientific evidence of DFI is documented. The DFI corpus was well annotated, and baseline BERT models based on the DFI corpus achieved remarkable performances on the classification tasks related to DFI extraction. Because the annotation guideline for the DFI corpus resolved diverse problems to extract DFI from abstracts of biomedical articles, any NLP model developed, revised, and fine-tuned on the DFI corpus would be useful to build an up-to-date DFI database.

Chapter 5

Experiences in Developing Annotation Guideline for Extracting Clinical Information from Unstructured Free-texts

In this section, I elaborate on our experiences in developing annotation guidelines to extract clinical information for pharmacovigilance and identify issues around defining annotation elements like a word entity and a relation between word entities. Although the issues I identified are only based on our experience gained from two studies explained in Chapters 3 and 4, I decided to write this section in the dissertation because the issues to be considered when defining annotation elements for clinical information extraction have not been systemically reviewed before.

5.1 Issues around defining a word entity

The consistency of annotations between different annotators is one of the major indicators of annotation quality. An annotation guideline is a document that provides the purpose of the annotated corpus, clear definitions of entity types, and annotation instructions. Instructions

contained in an annotation guideline should be enough clear enough for a different annotator to perform annotation consistently.

In clinical information extraction, annotation guidelines should provide instruction about a mention span of clinical word entities to ensure consistency in setting mention spans. A mention span means a range of strings annotated as word entities for a given entity type. Two different annotators may recognize the same word entity in a given but set mention spans differently for the recognized word entity. For example, one annotator labels the string ‘orally tablet’ as the *DrugFormulation* word entity, while a different annotator labels only the string ‘tablet’ except ‘orally’ as the *DrugFormulation* word entity in the given text.

Especially, compositional concepts in clinical texts which do not map to a single medical concept are problematic to setting a mention span. [133] For example, one annotator labels ‘chest and back pain’ as an *AdverseEvent* entity while a different annotator labels ‘chest’, ‘back’, and ‘pain’ separately. The rationale of different mention spans might be based on which medical concepts the annotators think about when finding clinical word entities to annotate. If an annotator thinks ‘chest and back pain’ is an individual symptom, then the annotator will label the three words as one word entity. However, when an annotator thinks that ‘pain’ is the main concept of a symptom and considers ‘chest’ and ‘pain’ as medical modifiers, the annotator can label the three words separately. Therefore, I recommend specifying a medical vocabulary corresponding to a word entity regardless of whether the word entity is included in the entity linking tasks or not. Indeed, I provided lists of drug products and drug compounds used by the MFDS to annotators when annotating ADE narratives and the drug lists helped the annotators to consistently set mention spans of drug entities.

Also, the overlapping of mention spans of different word entities is more likely to happen when defining multiple word entities in the annotation guideline. Mention span overlapping also

raises a problem when training NER models because I usually defined a NER task as a token-level classification. I should give a word entity type (or label) to a token based on annotation information to construct a NER dataset. However, when a token is annotated as more than one word entity type, I need to decide which of the annotated word entity types the token will have as a label. Thus, I suggested two ways to solve the problem caused by the overlapping of mention spans: 1) develop separate NER models for word entities that frequently overlap with each other, 2) set priorities between word entity types so that a word entity label of higher ranking is assigned to each token.

Training separate NER models for frequently overlapped word entities has the advantage of using all the annotation information for model training but has the disadvantage of connecting multiple NER models properly to develop a final end-to-end pipeline that extracts clinical information of interest from free-texts. On the other hand, when setting priorities between word entity types to give a single label to a token, I lose some annotation information during assigning token labels, but I can keep the NER model as one.

5.2 Issues around defining a relation between word entities

Extracting relations between word entities is a crucial task to extract semantic relations between entities from unstructured free-texts. Especially, in clinical information extraction, relations between word entities are useful to annotate temporal information, e.g., administration time or occurrence time of adverse events, and scientific reasoning, e.g., causality assessment results between an ADE and a drug, in clinical information extraction. However, there are several factors that make it difficult to obtain high-quality annotation data of relation between word entities.

Firstly, it is frequent in clinical texts that word entities referring to the same object appear repeatedly. Taking the case of extracting DFI from a biomedical abstract as an example, a food of

interest appears multiple times in the biomedical abstract, whether expressed as the same word, a pronoun or other abbreviation. However, the annotation burden would greatly increase when I annotate all relations between all food words mentioned in the abstract and a DFI key-sentence. In addition, a relation unrelated to the semantic relation could be annotated when labeling a relation between the DFI key-sentence and food entities at a distance only because the food words denote the same object.

Alternatively, I can instruct to annotate a relation only with the closest word among words indicating the same object. However, if I annotate a relation only between the closest words, relations between word entities at a distance were considered as negatively related in a training dataset for relation extraction. Thus, ideally, I should identify word entities indicating the same object in a text and then predict a relation only between the closest word entities among words indicating the same object. This is also an implausible story.

Thus, I recommend to determined how to obtain a negative relation sample, a pair of word entities labeled as negatively related in training datasets for the relation extraction, before annotating relations between word entities. In general, relations not manually annotated by an annotator among all the possible combinations of word entities are used as negative samples in the relation extraction. However, a dataset for relation extraction becomes highly imbalanced if the number of actual annotated relations is much less than the number of all the possible combinations of word entities. To address severe label imbalance in a training dataset for the relations extraction, I limited the maximum number of negative relations in a single document to balance positive, i.e., annotated, and negative relations (see Chapter 3, section 3.2.8).

5.3 Issues around defining entity labels

The entity label is used to provide additional information about an entity, such as the identifier of a drug or whether a patient has been given the drug. When the number of entity types becomes large and difficult to manage, the use of entity labels can help reduce the burden of selecting the correct entity type during annotation. As discussed in section 5.1, clinical texts often have a high number of entity types, which can make the annotation process time-consuming. In these cases, introducing an entity label instead of increasing the number of entity types can improve the efficiency of the annotation process.

For instance, when annotating a 'Date' entity, if you want to indicate whether the date represents the start or end of a medical action, you can add a 'start_or_end' label instead of creating two separate entity types for 'Date start' and 'Date end'. Additionally, if the majority of dates in the text indicate the start of an action, setting the default value of the 'start_or_end' label to 'start' can further streamline the annotation process, only requiring the use of the 'end' value for dates that indicate the end of an action. Thus, it is important to understand the distribution of information through pilot annotation in order to set an effective default value for an entity label.

Furthermore, it is recommended to extract the annotation data after the annotation process is complete and check if all required entity labels have been specified. This is because in most annotation tools, the entity label information is not immediately visible on the screen, making it difficult for researchers to quickly notice if some annotators are not consistently labeling entities or are using labels incorrectly. To avoid this problem, researchers should double-check the annotation data to ensure that all required labels have been applied correctly.

5.4 Issues around selecting and preprocessing annotated documents

Selecting the appropriate target documents for annotation is crucial step in designing the annotation guideline and managing the annotation process efficiently. Annotations are a labor-intensive and costly task, so the chosen document should be of high quality and contain a sufficient amount of information to be annotated. If selected documents have little information to extract, the cost of obtaining the same amount of annotated information will increase, as the price of annotation is typically set per document. Furthermore, documents with little information to extract may decrease the concentration of the annotators and overall annotation quality. On the other hand, if target documents are too long or contain too much information, annotators may become overwhelmed, leading to a decrease in annotation quality, such as potentially resorting to cheating.

In the studies discussed in Chapter 3 and 4, the length of the target document was used as a proxy for description quality, and documents that were too long or short were excluded from annotation. However, using this method for selecting target documents may result in a difference in the information extraction performance of the NLP model when measured on the annotated corpus compared to the whole document pool. To minimized this difference and accurately measure performance, it is important to ensure that the overall descriptive characteristics and clinical situations of the selected documents are similar to those of the entire document pool. Additionally, the clinical context in which a clinical text is written also could be a great proxy for the document's meta-information (e.g., who wrote the document and the writing purpose).

Also, I showed that using structured clinical information to select diverse target documents can improve the performance of drug safety information extraction (see section 3.4.4). I hypothesized that the expression of drug safety information may differ depending on the clinical situation of the patient. Although I did not directly test this hypothesis, I believe that incorporating ADE narrative reported with diverse data items into training data may diversify the patterns of description of ADEs, leading to improved information extraction performance. Furthermore,

because structured clinical information and clinical texts documented appear together, such as EHRs and clinical notes, I believe this approach can be applied to a range of clinical information extraction scenarios.

In addition, I suggest expanding any medical abbreviation in the target document before performing annotation. While a clinical language model may have some knowledge of medical abbreviations, providing the full phrase will improve the input for NLP models. Furthermore, it can be difficult to change text string after annotation, so it is important to expand abbreviations beforehand. For example, chemotherapy regimen names often consist of the initial letters of multiple drugs. In these cases, it would be beneficial to replace the abbreviated chemotherapy name with a detailed drug list before annotation.

Chapter 6

CONCLUSION

6.1 Dissertation summary

In this dissertation, I investigated the extraction of clinical information from unstructured free-texts for pharmacovigilance purpose.

In Chapter 3, I defined the extraction of comprehensive drug safety information from ADE narratives reported through SRS as an NLP task and developed a manually annotated corpus for this purpose. I also developed the KAERS-BERT model, which is pretrained on 1.2 million ADE narratives reported in KAERS and used it to provide strong baseline performance on the defined NLP tasks. Furthermore, I showed that the NER performance could be improved by adding ADE narratives with the least reported items to training datasets. I also discussed the differences between structured drug safety information in KIDS-KS and the annotated information in our corpus.

In Chapter 4, I introduced the DFI corpus, a manually annotated corpus for extracting DFI information from the abstracts of biomedical articles. The DFI corpus aims to systematically extract scientific evidence of DFIs from biomedical abstracts, defining DFI key-sentence as the prediction target. In contrast to a previous dataset for extracting DFI [134], which exclusively relied on MeSH terms to select abstracts, the DFI corpus was constructed by selecting biomedical

abstracts that included both a food word and a drug/drug-related molecule word. In fact, I found that the vast majority of abstracts containing DFI key-sentences could not be tagged when searched only by the DFI MeSH term

In Chapter 5, I investigated the issues surrounding the definition of annotation elements for extracting clinical information related to pharmacovigilance. I discussed the problem caused by overlapping mention span and suggested two possible solutions. I also emphasized the importance of obtaining negative relations before annotating relations between word entities.

6.2 Limitation and future works

6.2.1 Development of end-to-end information extraction models from free-texts to database based on existing structured information

In this dissertation, I demonstrated that I can develop effective NLP models for extracting clinical information of interest from unstructured free-texts by fine-tuning BERT models and creating a clear annotation guideline for obtaining a high-quality annotated corpus. However, achieving strong performance on individual NLP tasks does not guarantee the success of an end-to-end NLP pipeline for extracting information of interest from the target documents and storing it in a structured form in an existing database. Furthermore, in order to develop a complete end-to-end pipeline for extracting clinical information, it is necessary to identify and improve the performance characteristics of NLP models trained on individual tasks, such as NER, relation extraction, and sentence classification, or supplement them with appropriate human supervision or post-processing modules.

However, developing a post-processing module that combines the results of individual NLP models to extract clinical information can be a time-consuming and resource-intensive task

that requires expert guidance and access to relevant data. Additionally, using these models sequentially can lead to compounding errors and decreased performance in the final information extraction. To improve the accuracy of the extracted information, it is necessary to identify and address the specific error patterns of individual models, such as by using rule-based post-processing to correct errors in NER. Moreover, evaluating the performance of the end-to-end information extraction model requires access to a gold standard dataset. Researchers have explored methods such as simultaneous NER and relation extraction [116-119] and multi-task learning [120] to reduce error propagation through NLP models and improve final information extraction performance. In recent research, task formulation has been performed in the form of text-to-table [135] to directly derive the database structure from natural language data, or NER and relation extraction tasks have been solved simultaneously by filling an enhanced table [136].

In this situation, developing an NLP model for clinical information extraction using weak supervision with structured clinical data in a database can reduce hugely annotation burden and enable immediate evaluation of the model performance by comparing extracted information from free-texts to the structured data. However, there is a risk of discrepancy between information in the free-text and structured data when using structured clinical information for supervision. This means that the free-text may contain information not present in the database, or vice versa. Additionally, relying on structured data as weak supervision may lead to an NLP model that simply reaffirms already reported clinical information, rather than extracting missing information from the free-text.

On the other hand, using structured data as a gold standard for developing an information extraction allows for easier evaluation of the model 's usefulness in streamlining the actual reporting processes. Furthermore, manually annotating differences between structured data can help identify which clinical information is frequently omitted in the reporting process, allowing for the development of more targeted NLP models

6.2.2 Application of in-context learning framework in clinical information extraction

In this dissertation, I developed NLP models for extracting clinical information related to pharmacovigilance. I chose to create human-annotated corpora to train these models because previous information extraction models suffered from issues with existing annotated corpora, and the target domain of the document significantly different from those used in existing models. Additionally, I was able to provide a concrete guide for extracting clinical information of interest for both annotators and NLP models. However, as is typical, a large amount of time and resources were required to develop the annotated corpora for two research projects in this dissertation. I recruited five pharmacists with experience in reporting ADEs, as well as five undergraduate students majoring in medicine or pharmacy, to serve as annotators for studies described in chapters 3 and 4, respectively. The annotation program took more than three months and cost $400 per annotator and month. Furthermore, two or three graduate students were needed to manage the annotation schedule and ensure the quality of the annotations.

Moreover, despite the availability of many datasets in the biomedical NLP field for clinical information, more than half of the documents sources were clinical notes or biomedical scientific literature, and less than 30% of the datasets were publicly accessible and had restrictive license [137]. This might be due to the sensitive nature of clinical text.

One potential approach for addressing the issue of limited availability and high cost of annotated corpora in the biomedical domain is to use LLMs such as GPT-3 [49, 138]. Zero-shot and few-shot learning framework has been shown to be as effective as, or even outperform, traditional pretraining-finetuning frameworks on a variety of general-domain tasks [139-141]. In addition, LLMs can be adapted to various tasks through the use of prompt-based learning, also

known an in-context learning, which allows the model to be trained on a small number of examples texts, or “prompts”, without the need for large amounts of training data.

However, conflict results have been reported to the applicability of zero-shot and few-shot learning framework in the biomedical domain at the time of December 2022. One study showed that LLMs could be great zero-shot clinical information extractors through converting output of GPT-3 using a simple rule-based resolver [142]. However, other studies pointed out that few-shot or zero-shot learning framework did not achieve great performances on benchmark tasks in the biomedical NLP field [143, 144], while the same framework was comparable to or even outperformed to pretraining and fine-tuning framework in the general domain information extraction problems.

Researchers having the opinion that a current GPT-3 model is poor zero-shot or few-shot learner in the biomedical domain have pointed out that in-context learning is not feasible in the biomedical field due to long and complicated schema of annotation for information extraction [143, 144]. Also, previous studies have utilized GPT-3 pretrained on the general domain, not one pretrained on the biomedical domain like BioGPT3 [145]. Furthermore, it is also noteworthy that advancements in optimizing a zero-shot learning framework such as answer engineering [146, 147], prompt ensembling [148] and continuous prompts [149, 150] have not been applied to studies testing applicability of GPT-3 on clinical information extraction.

Appendix

7.1 Annotation Guideline for "Extraction of Comprehensive Drug Safety Information from Adverse Event Narratives Reported through Spontaneous Reporting System"

These annotation guidelines aim to define annotation task and offer guidance for annotation to consistently structure the natural language descriptions on the adverse event reports in the Korea Adverse Event Reporting System (KAERS) operated by the Korea Institute of Drug Safety & Risk Management (KIDS). This annotation guideline was updated last on August 30th, 2022.

The structured natural language data developed according to these annotation guidelines will be used to build the natural language processing (NLP) model that extracts key clinical information for drug safety evaluation from the natural language descriptions on the reports in the KAERS.

1 Word entities

In this part, the entities that comprise the KAERS natural language narrative are defined, and guidelines and notes are provided for consistently annotating those entities.

1.1 Pathological finding

1.1.1 Adverse Event (*ADE*)

ADE is referred to the distinct adverse events that are described in the narratives of the reports, which are defined as all symptoms or diagnosed diseases that occurred **after** the administration of suspected drugs. The *ADE* is a word-level entity and should be annotated together with any accompanying modifiers (e.g., chronic, acute). Any adverse event stated in the descriptions should be labeled as *ADE,* regardless of whether it actually occurred to the patient or was diagnosed as such.

- Caution: Examples like the ones below should also be annotated as *ADE* and labelled with the appropriate MedDRA code.
  - 'Lack of efficacy' (MedDRA 10014291): if a case indicates that the medication has not shown the anticipated efficacy or effectiveness after it was taken
  - 'Product quality issue' (MedDRA 10069327): if a description mentions "product quality issue", "product physical issue", "product/quality related defects", or etc.

1.1.2 Disease or Indication (*Disease*)

Disease is referred to the distinct historical disease mentioned in the in the narratives of the reports. Technically, the term "*Disease*" in *Disease* refers to all symptoms or diagnosed diseases that occurred **before** the administration of suspected drugs, and the term "Indication" is a disease that a medicine is intended to treat or prevent. However, at the annotation task using these annotation guidelines, diseases and indications should be annotated as *Disease* instead of separating these two. When an indication is annotated as *Disease,* a relation with the relevant *Drug* should be assigned.

Disease is a word-level entity, should be annotated together with any accompanying

modifiers (e.g., Type 2 diabetes). Any disease or indication stated in the descriptions should be labeled as *Disease,* regardless of whether it actually occurred to the patient or was diagnosed as such. If it is not possible to judge whether a disease and indication occurred before or after the administration of suspected drugs, it should be consistently annotated as *ADE*, rather than Disease.

1.1.3 Seriousness of Adverse Event (*ADESeriousness*)

ADESeriousness is referred to the seriousness of the adverse event that the patient who was the subject of the narratives experienced. *ADESeriousness* is annotated as one of the following for each event, based on the seriousness criteria of the adverse event provided in the ICH E2B R3 implementation guide: 1) Results in death, 2) Life threatening, 3) Caused or prolonged hospitalization, 4) Disabling or incapacitating, 5) Congenital anomaly or birth defect, 6) Other medically important condition, 7) not serious. *ADESeriousness* should be tagged at a sentence-level. A noun phrase or word can be used as *ADESeriousness* entity only when the seriousness is not stated as a sentence. For example, when it is certain that an adverse event led to hospitalization, as in the statement like 'he/she was hospitalized due to gastroenteritis.', 'hospitalized' should be annotated as both an *ADESeriousness* entity and an *EventAdmission* entity.

1.1.1 Adverse Event at the Time of Last Observation (*ADEatLastObs*)

ADEatLastObs is referred to whether the adverse event described in the narratives was maintained or resolved at the time of last observation. *ADEatLastObs* is annotated as one of the following for each event, based on the definition of outcome of adverse event at the time of last observation: 1) recovered, 2) recovering, 3) not recovered, 4) recovered with sequelae. *ADEatLastObs* should be tagged as a noun phrase or word which denotes the outcome of an adverse event at the time of

last observation. However, when the outcome of an event is 'improvement (or recovery)' followed by 'recurrence', each noun phrase or word indicating 'improvement (or recovery)' and 'recurrence' should be annotated as *ADEatLastObs,* even though 'improvement (or recovery)' is not the final outcome.

1.2 Drug

Drug is referred to the distinct drug names that are described in the narratives of the reports. When a drug name is mentioned as a generic name, it is annotated as *DrugCompound*; when it is mentioned as a product name (brand name), it is annotated as *DrugProduct*; and when it is mentioned as a drug group that shares the therapeutic target or that is categorized for other reasons without specifying a generic name or brand name, it is annotated as *DrugGroup*.

1.2.1 Drug Compound (*DrugCompound*)

DrugCompound is referred to the distinct generic names that are described in the narratives of the reports. A *DrugCompound* entity should be annotated at a word-level only. Phrases or words indicating dose, dosing interval, route of administration, or drug formulation should not be marked as *DrugCompound* because, according to these annotation guidelines, they are independent entities. When the description doesn't provide the particular drug information, e.g. 'suspected drug' or 'administered drug', it shouldn't be annotated as *DrugCompound.* As a general rule, a drug can only be annotated as *DrugCompound* if both the generic code list and the product code list include the drug name.

1.2.2 Drug Product (*DrugProduct*)

DrugProduct is referred to the distinct product names (brand names) that are described in the narratives of the reports. A *DrugProduct* entity should be annotated at a word-level only. Phrases or words indicating dose, dosing interval, route of administration, or drug formulation should not be marked as *DrugProduct* because, according to these annotation guidelines, they are independent entities. In case of 'GastidineTab [6] 150mg' or 'GranatecEyeDrops[6] 0.4%', 'GastidineTab' and 'GranatecEyeDrops' should be tagged as *DrugProduct*, while '150mg' and '0.4%' should be annotated as *Dose*. When the description doesn't provide the particular drug information, e.g. 'suspected drug' or 'administered drug', it shouldn't be annotated as *DrugProduct.* If a drug name contains a number as part of it and does not include a dose unit, the drug name including the number should be annotated as an entity, and *Dose* shouldn't be marked.

Korean oriental medicines prepared by individual hospitals or clinics of Korean medicine should be also annotated as *DrugProduct,* even if they are not registered on the generic code list or the product code list.

1.2.3 Drug Group (*DrugGroup*)

DrugGroup is referred to the distinct drug groups that shares the therapeutic target or that is categorized for other reasons (e.g. antidiabetic drugs, antihypertensive drugs, antihistamines). A *DrugGroup* entity should be annotated at a word-level only. Phrases or words indicating dose, dosing interval, route of administration, or drug formulation should not be marked as *DrugGroup* because, according to these annotation guidelines, they are independent entities. When the description doesn't provide the particular drug information, e.g. 'suspected drug' or 'administered

[6] This word is intended to be "Gastidine Tab" in English, but it was intentionally written without a space between the two words because different annotation rules were used depending on the spacing between a drug name and a word indicating the formulation, considering the nature of Korean language.

drug', it shouldn't be annotated as *DrugGroup.*

1.2.4 Others: combination therapies for cancer treatment

When a regimen name of combination therapy for cancer treatment (usually acronym formed from the drugs composing the combination) is used in the narratives of the reports, each initial letter denoting an individual generic name should be annotated separately as *DrugCompound.* (e.g. COPADM: C/O/P/AD/M - Cyclophosphamide, Oncovin (Vincristine), Prednisone, ADriamycin (Doxorubicin), Methotrexate)

1.3 Dosing Information

To assess a drug's safety, administration details as well as drug identification information are required. The KAERS natural language annotation task therefore comprises structuring the dose, dosing interval, and route of administration or drug formulation.

1.3.1 Dose (*Dose*)

Dose is referred to distinct dose of a drug administered that are described in the narratives of the reports. *Dose* contains the amount of the drug's active ingredient (e.g., '10mg', '1g'), the concentration of the injection or eye drop (e.g., '5%', '50g/L'), the frequency or cycle of dosage (e.g., '1 time', '2 cycles'), and the dosing duration (e.g., 'for a week', '1 year'). A *Dose* entity should be annotated at a word-level only, along with the numeric component (dose) and the dose unit. A word or noun phrase describing the drug amount should also be a *Dose* entity, even though it was not actually administered, in accordance with the annotation rules for the drug entities (e.g., in case of '480cc abandoned among 500cc', '480cc' and '500cc').

A number should not be tagged as *Dose* if it was solely described without a unit (e.g., a number without a unit as part of drug brand name, blending ratio of Korean oriental medicines).

1.3.2 Dosing Interval (*DosingInterval*)

DosingInterval is referred to distinct dosing intervals that are described in the narratives of the reports. *DosingInterval* categories include the frequency of dosage at a specific time (e.g., once/twice a day) and dosing intervals (e.g., every 6 hours).

A *DosingInterval* entity should be annotated at a word-level, along with the numeric component and the time unit (e.g., 'hour(s)', 'day(s)', 'week(s)').

1.3.3 Route of Administration or Drug Formulation (*RoAorFormulation*)

RoAorFormulation is referred to distinct route of administration or drug formulation that are described in the narratives of the reports. The routes of administration include oral administration, types of injection (e.g. 'iv', 'intramuscular'), other forms of administration (e.g., 'inhale', 'topical administration'). The drug formulations include specific types of formulation for oral use (e.g., 'sustained release tablet', 'SR', 'capsule'), specific types of formulation for topical use (e.g., 'ointment', 'cream', 'lotion', 'gel'), etc.

RoAorFormulation should be annotated only if the route of administration or drug formulation was described as a separate word. A drug formulation should not be annotated as an entity if it is expressed as a part of a product name (brand name) without spaces. Instead, the product name (brand name) along with the word describing the formulation should be tagged as *DrugProduct*. (e.g., 'Meropen inj' → 'Meropen': *DrugProduct* and 'inj': RoAorFormulation,

‘MeropenInj’ → ‘MeropenInj’: *DrugProduct*)

A word that requires analogical inference to determine the route of administration or drug formulation (e.g., ‘inoculation’) shouldn’t be annotated as *RoAorFormulation. RoAorFormulation* tags shouldn't be applied to drug formulations that are only utilized for particular products, such as ‘turbuhaler’ or ‘OROS (or OROS tab)’. (In this case, too, if the word defining the formulation is expressed as a part of a product name (brand name) without spaces, it should be categorized as *DrugProduct*.)

1.4 Date

Date is referred to all kinds of distinct date and time information that are described in the narratives of the reports. Date and time information means date information with the year, month, and day, and time information with the hour and minute. The date entities should be annotated at a word-level or noun phrase-level, along with the punctuation marks and the words describing the date, such as ‘6/25’ or ‘12th month[7], year of 2016’. However, if a phrase contains both date and time information such as ‘10 am on June 26th’, ‘10 am’ and ‘June 26th’ should be annotated as two entities separately. The date entities also include expressions such as ‘the next day’, ‘the same day’, and ‘unknown date’. The date entities are categorized into *DateStartorContinue, DateEnd,* and *DatePeriod*, depending on the medical events occurred at the respective date and time.

1.4.1 Date Start or Continue (DateStartorContinue), DateEnd (DateEnd)

[7] This term is intended to be "December" in English, but it was intentionally written as “12th month” to express the word meaning “month” in Korean.

DateStartorContinue, DateEnd are referred to the dates and times when a medication or an adverse event starts/continues or ends, respectively. For example, in a sentence like 'Smoflipid 20% infusion was started at 10 on May 25th to supply nutrition', 'at 10' and 'May 25th' should be tagged as *DateStartorContinue.* On the contrary, in a sentence like 'Clopidogrel administration was stopped at 18:00 evening on the same day', 'at 18:00 evening' and 'on the same day' should be tagged as *DateEnd*. A word or noun phrase containing the related dates or times should also be annotated as *DateStartorContinue* if they don't exactly mean the dates and times when a medication or an adverse event starts/continues or ends.

1.4.2 Date Period (DatePeriod)

DatePeriod is referred to the length of an adverse event, the duration of a historical disease or medication, or the passage of time since a particular date or time point (e.g., 'after 2 hours', 'since the second day'). However, a duration containing date information such as 'from January 2nd to 3rd', should be annotated as *DateStartorContinue* instead of *DatePeriod,* with the dates 'January 2nd' and '3rd' tagged separately.

1.5 Patient Information

PatientInfo is referred to distinct demographic information that are described in the narratives of the reports. Under these guidelines, patients' sexes and ages are to be structured.

1.5.1 Patient Sex (*PatientSex*)

PatientSex is referred to the patient's sex that is described in the narratives of the reports and is annotated as a noun phrase or word that denotes sex (e.g., 'female', 'male', 'M', 'F'). *PatientSex* should only be annotated if the sex is explicitly mentioned and not implied by the medical

terminology (e.g., ‘ovarian cancer’).

1.5.2 Patient Age (*PatientAge*)

PatientAge is referred to a distinct age, birth year, or birth date that are described in the narratives of the reports and is annotated as a noun phrase or word that denotes an age (e.g., ‘postnatal 8 months’, ’25 years old’, ‘born in 1968’, ‘born in July 2nd, 1988’). A word that indicates an age range (e.g., ‘in thirties’) should also be annotated as *PatientAge*, but a word that implies an age (e.g. ‘adolescent’, ‘elderly’) or a term for a disease (e.g., ‘neonatal diabetes’) shouldn’t.

1.6 Others

The other distinct entities related to drug safety information that were not previously addressed are defined in this section.

1.6.1 Test Name (*TestName*), Test Result (*TestResult*)

TestName and *TestResult* are referred to a distinct test name (e.g., ‘electrocardiogram’) and a test result (e.g., ‘normal’, ‘positive’, ‘126mmHg’, ‘lesion not identified’) that are described in the narratives of the reports. If a test name can be inferred from an expression though it was not explicitly stated, it should be annotated as *TestName*. For example, ‘BP was between 90 and 120’, ‘the body temperature was 36.8’, ‘BP’ and ‘the body temperature’ should be tagged as *TestName* and ‘between 90 and 120’ and ’36.8’ as *TestResult*, respectively. If a numeric value of a vital sign was mentioned along with a term indicating a pathological finding from the vital sign such as ‘fever’ (e.g., high fever with 39 degrees), the term should be tagged as *TestName*, and the numeric value as *TestResult*, and at the same time, the term should be tagged as *ADE*.

However, a test result itself should be annotated as *TestName* and/or *TestResult*, but not

ADE. If a changing trend such as a decrease or increase was addressed, it can be tagged as *ADE* depending on the context (e.g., 'neutrophil counts 28.2%': not annotated as *ADE*, 'neutrophil counts decreased': annotated as *ADE*)

A *TestName* entity should be tagged at a word level and *TestResult* at a noun phrase or word level, which comprises the test result's numerical value and unit. If a test result was discussed in a descriptive manner (e.g., '(in gastroscopy) no active bleeding was noted'), it should be annotated at a clause or sentence level.

The reference value (normal range) of a test should be tagged as *TestResult*, but the relation with *TestName* shouldn't be annotated.

1.6.2 Event Admission (*EventAdmission*), Event Discharge (*EventDischarge*)

EventAdmission and *EventDischarge* are reffered to a hospitalization event of a patient experiencing an adverse event that are described in the narratives of the reports. *EventAdmission* and *EventDischarge* entities should be annotated at word level.

1.6.3 Non-Drug Treatment (*NonDrugTreatment*)

NonDrugTreatment is referred to a distinct treatment other than medication (e.g., 'oral hydration') that is described in the narratives of the reports. A *NonDrugTreatment* entity should be annotated at word level. A procedure performed to diagnose a pathological condition should be tagged as *TestName*, rather than *NonDrugTreatment* (e.g., Computed Tomography or CT).

1.6.4 Action Taken with Drug (*ActionTakenwDrug*)

ActionTakenwDrug is referred to a distinct action taken with the drug that is described in the narratives of the reports. *ActionTakenwDrug* is annotated as one of the followings for each event, based on the ICH E2B R3 implementation guide: 1) Drug withdrawn, 2) Dose reduced (single dose reduced or dosing interval lengthened), 3) Dose increased (single dose increased or dosing interval shortened), 4) Dose not changed, 5) Unknown. An *ActionTakenwDrug* entity should be annotated at a noun phrase or word level.

1.6.5 WHO-UMC Results of Assessment (*WHO-UMCAssessment*)

WHO-UMCAssessment is referred to a distinct result of assessment on the causality of a drug with an adverse event based on World Health Organization-Uppsala Monitoring Centre (WHO-UMC) that is described in the narratives of the reports. *WHO-UMCAssessment* is annotated as one of the followings for each event, based on WHO-UMC's definition: 1) certain, 2) probable, 3) possible, 4) unlikely, 5) conditional/unclassified, 6) unaccessible/unclassified. (Please refer to 4. Entity Labels – C. Others – iii. action taken with drug (E2BR3) for the definition and explanation).

WHO-UMCAssessment should be annotated at a sentence level which addressed the causality assessment, but if the sentence includes two different assessments, each assessment should be annotated separately at the clause or noun phrase level.

2 Relations

This part defines the relations between the entities established above and provides the rules and additional notes to annotate the relations in a consistent manner.

2.1 ADE ↔ ADE, Disease ↔ Disease

The relations *ADE* ↔ *ADE* and *Disease* ↔ *Disease* should be annotated between the modifiers and the modified *ADE* or *Disease* separately. For example, in 'hand and foot pain', 'hand pain' and 'foot pain' should be annotated as *ADE* and the relations 'hand' ↔ 'pain' and 'foot' ↔ 'pain' should be created separately.

If a synonym is followed by an adverse event or disease enclosed in parenthesis, the terms before and after parentheses should be annotated as different entities, and the relations between the two words should be linked. (e.g., 'non-steroidal anti-inflammatory drugs (NSAIDs)': 'non-steroidal anti-inflammatory drugs' ↔ 'NSAIDs', '[Ecchymosis in Korean] (Ecchymosis)': '[Ecchymosis in Korean]' ↔ 'Ecchymosis'

Additives should be annotated as *DrugCompound*, but shouldn't be linked a relation with the drug entity.

2.2 ADE ↔ (ADESeriousness or ADEatLastObs)

The relations *ADE* ↔ *ADESeriousness* or *ADEatLastObs* should be annotated between an adverse event and the seriousness of the adverse event or the adverse event at the time of last observation.

2.3 Disease ↔ (DrugCompound, DrugProduct, or DrugGroup)

The relations *Disease* ↔ *DrugCompound, DrugProduct, or DrugGroup* should be annotated between a drug entity and the drug's indication.

2.4 (ADE, Disease, ADESeriousness, ADEatLastObs, DrugCompound, DrugProduct, DrugGroup, EventAdmission,

EventDischarge, TestName, NonDrugTreatment, or ActionTakenwDrug) ↔ (DateStartorContinue, DateEnd, or DatePeriod)

The relations above should be annotated for the beginning, ending, or maintenance of an adverse event or disease, the beginning, ending, or maintenance of drug administration, the date on which an event met the seriousness criteria, the time of the most recent AE observation, the dates on which hospitalization and discharge occurred, and the dates on which a test was performed. If a sentence or a clause tagged as *ADEatLastObs, ADESeriousness ActionTakenwDrug,* or *ADEatLastObs* includes a date, the date should be annotated as a separate entity and the relation between the date and the entity indicating the related event should be created. (e.g., 'Since the condition was improved, the patient was discharged on Aug 30th' → 'the condition was improved': *ADEatLastObs,* 'Aug 30th': *DateStartorContinue*, 'discharged': *EventDischarge,* 'Aug 30th' ↔ 'discharged' (relation))

A report date shouldn't be linked the relation with a date.

2.5 DrugProduct ↔ DrugCompound

The relation should be annotated between a product name (brand name) and its generic name. For example, for a description such as 'LoxefinTab (loxoprofen sodium) was orally taken 3 times a day from Dec 27th', a relation between 'LoxefinTab' (*DrugProduct*) and 'loxoprofen sodium' (*DrugCompound*) should be created, since 'loxoprofen sodium' is the active ingredient of 'LoxefinTab'. Besides, the other medication information such as '3 times a day' (*DosingInterval*) and 'orally taken' (*RoAorFormulation*) should be annotated the relation with 'LoxefinTab' (*DrugProduct*), separately.

2.6 (DrugCompound, DrugProduct, or DrugGroup) ↔ (Dose, DosingInterval, or RoAorFormulation)

The relations above should be annotated between a generic name, a product name, or a drug group and any information regarding drug administration.

2.7 TestName ↔ TestResult

The relations *TestName* ↔ *TestResult* should be annotated between a test name and its result.

2.8 ActionTakenwDrug ↔ (DrugCompound, DrugProduct, or DrugGroup)

The relations *ActionTakenwDrug* ↔ (*DrugCompound, DrugProduct,* or *DrugGroup*) should be annotated between a drug and the action taken with that drug.

2.9 WHO-UMCAssessment ↔ (ADE, Disease, DrugCompound, or DrugProduct)

The relations *WHO-UMCAssessment* ↔ (*ADE, Disease, DrugCompound, or DrugProduct*) should be annotated between a result of assessment on the causality of a drug with an adverse event based on WHO-UMC, the related drug, and the adverse event. If it is uncertain to which entity *WHO-UMCAssessment* is related, the relation shouldn't be linked.

3 Entity labels

This part describes the entity labels for the previously defined entities, which provide detailed

information and a medical code, and gives the rules and additional notes to annotate the entity labels in a consistent manner.

3.1 Pathological Finding

3.1.1 [pathological finding identifier] (MedDRA)

A pathological finding identifier using a MedDRA code should be labeled on *ADE* and *Disease* entities to identify pathological findings. Any entities that have been annotated as *ADE* or *Disease* must be labeled. depending on the annotator's medical knowledge and judgment, an appropriate 8- digit MedDRA code should be entered (e.g., 'Nausea': '10028813'). It is required to utilize MedDRA Version 24.0 English & Korean, and MedDRA Desktop Browser Version 4.1 can be used for the code search.

Generally, it is recommended to use the Lowest Level Term (LLT) to code the identifier. However, it is acceptable to utilize a Preferred Term (PT) if it is considered that an *ADE* or *Disease* entity may apply to more than two LLTs under that PT or if the description is insufficient to identify the LLT. The usage of higher class codes (i.e., System Organ Class (SOC), High Level Group Term, and High Level Term) is not allowed.

The annotators shouldn't assign labels to the entities based on the diagnoses made by their own medical judgment or interpretation of the statement. For example, 'irregular menstruation' should not be diagnosed and labeled based on the description 'She sometimes menstruates during her menstrual cycle and sometimes she doesn't.'

- *Disease* to be cautious: If a medication was given for a preventive purpose, such as the prophylaxis of vomiting, the identifier should be 'vomiting prophylaxis (MedDRA 10068079)', not 'vomiting'.

3.1.2 [not_occurred]

The entity label [not_occured] should be labeled on the *ADE* and *Disease* entities to indicate whether they actually occurred or not. For a detailed explanation of cases not marked as [not_occured], please see 'D. other considerations: the cases that are not labeled with [not_occured]/[not_concerned]'.

3.1.3 [seriousness]

The entity label [seriousness] should be labeled on the *ADESeriousness* entities to identify the seriousness categories of the adverse event. The entity label [seriousness] should be selected among the values listed in the table below.

If the adverse event is definitely a serious adverse event but the seriousness category was not specified (e.g., 'Serious adverse event: generalized skin eruption', 'This initial report for SAE is~'), '6_ Medically Important_Unclassified' should be marked.

Table 7.1: Entity labels for [seriousness]

The seriousness criteria of the adverse event provided in the ICH E2B R3 implementation guide
1 = Death
2 = Life Threatening
3 = Caused/Prolonged Hospitalization
4 = Disabling/Incapacitating
5 = Congenital Anomaly/Birth Defect
6 = Medically Important_Unclassified
7 = Not Serious

- Caution

1) The annotators should be aware that the severity of an adverse event (e.g., CTCAE Grade 1-4, mild/moderate/severe) and the seriousness are different concepts. Therefore, a word or noun phrase expressing the severity should not be mistakenly labeled with [seriousness]. (For 'CTCAE[8] Grade 5',
2) please see 2) following).
3) The description of 'death' cases (including 'CTCAE Grade 5') should be annotated as *AESeriouness* and labeled with [seriousness] – '1_Death'. 'Unknown cause of death' requires to have both an *AESeriouness* and *ADE* annotation.

3.1.4 [event_at_last_observation]

The entity label [event_at_last_observation] should be labeled on the *ADEatLastObs* entities to identify the category of the events at last observation provided in the ICH E2B R3 implementation guide. The entity label [event_at_last_observation] should be selected among the values listed in the table below. 'The time of last observation' refers to the latest observation in the chronological order of the narratives. Therefore, 'not recovered', rather than 'recovered', should be marked if an adverse event was recovered and then recurred.

Table 7.2: Definitions and examples for [seriousness] entity labels

The categories of the events at last observation	**Cases and Examples**

[8] CTCAE: Common Terminology Criteria for Adverse Events

1 = Recovered	The adverse reaction was not present at the time of the last observation, and the patient's health had returned to normal. e.g., 'the patient was recovered from peritonitis'
2 = Recovering	The patient had not fully recovered at the time of the last observation, but it was determined that he/she was in the process of recovering.
3 = Not recovered	At the time of the last observation, the adverse event is still observed and is not considered to be recovered or in the process of recovery. e.g., 'rash was recurred in Dec 2017'
4 = Recovered with sequelae	The adverse event is no longer observed at the time of the last observation, but the damage to the body due to the adverse event is not recovered and is observed.

3.1.5 [startcontinue_or_and]

The entity label [startcontinue_or_end] should be labeled on the *ADE* entity which has the relation annotation with the *DateStartorContinue* entity, but the adverse event was stopped/discontinued on the date. In this case, 'end' should be marked for the entity label [startcontinue_or_end].

3.2 Drug

3.2.1 [compountd identifier] & [product identifier]

The entity label [compound identifier] and [product identifier] should be labeled on the *DrugCompound* and *DrugProduct* entities to indicate the compound identifier for *DrugCompound* and the product identifier for *DrugProduct*. It is mandatory to label [compound identifier] or [product identifier] for the drug entities. The codes for [compound identifier] and [product identifier] are referred to the drug compound code list and the drug product code list prepared by the study team based on the Ministry of Food and Drug Safety's 'The active ingredient code list for the 2nd half of 2020' and 'The drug product code list for the 2nd half of 2020', respectively.

The 8-digit identifier starting with C (e.g., the integrated active ingredient code 'CM050288' for diclofenac, diclofenac sodium, diclofenac potassium) should be entered for [compound identifier] and the 10-digit identifier (e.g., the integrated product code 'C195600001' for PanpyrinTab) should be entered for [product identifier]. If a *DrugProduct* entity has a number, the code of the product name with the closest matching product name and number should be labeled.

For Korean oriental medicines prepared by individual hospitals or clinics of Korean medicine which are not registered on the code lists, C000000000 should be labeled.

3.2.2 [not_occurred] (or [not_administered])

The entity label [not_occured] should be labeled on the drug entities to indicate whether they were actually administered or not. For a detailed explanation of cases not marked as [not_occured], please see 'D. other considerations: the cases that are not labeled with [not_occured]/ [not_concerned]'.

3.2.3 [not_concerned]

The entity label [not_concerned] should be labeled on the drug entities to indicate whether they are the suspected drug or not. Based on these guidelines, historical diseases and adverse events are classified according to whether they occurred before or after the administration of suspected drugs. In order to distinguish between adverse events and historical diseases, the suspected drug should be identified when more than two distinct drugs are mentioned in the narratives of the report. Therefore, the drug(s) that the annotator used as a borderline between adverse events and historical diseases should be specified. For a detailed explanation of cases not marked as

[not_concerned], please see ‘D. other considerations: the cases that are not labeled with [not_occured]/[not_concerned]’.

3.2.4 [startcontinue_or_end]

The entity label [startcontinue_or_end] should be labeled on the drug entities which have the relation annotation with the *DateStartorContinue* entity, but the drug was stopped/discontinued on the date. In this case, ‘end’ should be marked for the entity label [startcontinue_or_end].

3.3 Others

3.3.1 [test name identifier] (based on MedDRA)

A test name identifier using a MedDRA code should be labeled on the *TestName* entities to identify test names. Any entities that have been annotated as *TestName* must be labeled with the 8-digit MedDRA code (e.g., ‘Creatine’: ‘10011328’). The annotation rules for [test name identifier] are basically the same as those for [pathological finding identifier]. However, the labels for [test name identifier] should be the LLTs or PTs under the SOC ‘Investigations’.

3.3.2 [non-drug treatment] (based on MedDRA)

A treatment identifier using a MedDRA code should be labeled on the *NonDrugTreatment* entities to identify the treatment. Any entities that have been annotated as *NonDrugTreatment* must be labeled with the 8-digit MedDRA code (e.g., ‘Percutaneous coronary intervention’: ‘10065608’). The annotation rules for [non-drug treatment] are basically the same as those for [pathological finding identifier].

3.3.3 [action taken with drug]

The entity label [action taken with drug] should be labeled on the *ActionTakenwDrug* entities to identify the category of the action taken with drug provided in the ICH E2B R3 implementation guide. Any entities that have been annotated as *ActionTakenwDrug* must be labeled with the values listed in the table below.

Table 7.3: Entity labels for [action taken with drug]

Categories of the action taken with drug provided in the ICH E2B R3 implementation guide
1 = Drug withdrawn
2 = Dose reduced
3 = Dose increased
4 = Dose not changed
5 = Unknown

3.3.4 [WHO-UMC results of assessment]

The entity label [WHO-UMC results of assessment] should be labeled on the *WHO-UMCAssessment* entities to identify the category of the causality assessment based on WHO-UMC. Any entities that have been annotated as *WHO-UMCAssessment* must be labeled. The labels should be chosen from the values listed the table below, based on the expression of the reporter or the primary source(s) of information on the narratives.

Table 7.4: Definitions and examples for [WHO-UMC results of assessment] entity labels

Causality assessments based on WHO-UMC	Case and Examples
1 = Certain	Based on the chronological order and clinical information at the time of drug administration and the time of the adverse event, it is certain that the adverse event is an adverse drug reaction (ADR) due to that drug, and it is not considered to be the condition caused by other factors.
2 = Probable	Based on the chronological order and clinical information at the time of drug administration and the time of the adverse event, it is probable that the adverse event is an ADR due to that drug, and is less likely to be the condition caused by other factors.
3 = Possible	Based on the chronological order and clinical information at the time of drug administration and the time of the adverse event, it is possible that the adverse event is an ADR due to that drug, but the possibility that the condition was caused by other factors cannot be ruled out.
4 = Unlikely	The possibility that the adverse event was caused by the drug cannot be completely ruled out, but it is more likely that other factors (such as concomitant medications or comorbidities) caused the adverse event.
5 = Conditional/Unclassified	The causality cannot be determined due to a lack of clinical information for safety assessment.
6 = Unassessable/Unclassifiable	When it is stated that the reporter and the primary source(s) of information didn't assess the causality.

3.3.5 Other considerations: the cases that are not labeled with [not_occurred] or [not_concerned]

3.3.5.1 When no entities in a document have the

[not_occured] or [not_concerned] labels

- When no drug entities in a document have the [not_occured]: The administration of each drug is regarded as having occurred.
- When no drug entities in a document have the [not_concerned]: Each drug is regarded as a suspected(concerned) drug.
- When no pathological finding entities in a document have the [not_occured]: Each adverse event/disease is regarded as having occurred.

3.3.5.2 When no entities in a document have the [not_occured] or [not_concerned] labels

- Each entity that doesn't have the [not_occured] or [not_concerned] label is regarded as having occurred or being concerned.
  - When 2 out of 10 *ADE* entities have the [not_occured] labels, the rest 8 *ADE* entities are regarded as having occurred.
  - When 1 out of 3 drug entities have the [concerned] labels, the rest 2 drug entities are regarded as not concerned.
  - When all 4 *ADE* entities don't have the [not_occured] labels and 1 out of 2 drug entities have the [occurred] labels, the 4 *ADE* entities are regarded as occurred and the drug entities without the [occurred] labels are regarded as [not_occured].

7.2 Annotation Guideline for "Extraction of Drug-Food Interactions from the Abtracts of Biomedical Articles"

'The DFI corpus' aims to develop an natural language processing model that detects a source document of scientific evidence on drug-food interaction (DFI) and extracts key-sentences describing DFI and related drug and food words. This annotation guideline was updated last on August 9, 2021.

In this annotation guideline, I defines annotation tasks and word and sentence entities annotated in the DFI corpus. Also, I provides a set of rule for annotating drug and food entities and key-sentences describing DFI or drug-drug interaction (DDI). This guideline was used for educating annotators when I created the DFI corpus. I hope that this annotation guideline helps you to understand the structure of the DFI corpus and the meaning of annotated entities for DFI extraction.

In the DFI corpus, I annotated 2270 abstracts published between January 1, 1970 and October 2, 2019 from a pre-specified medical/pharmacy journals. A detailed selection criteria for an annotated document was elaborated in Kim et al., 2021.

4 Word Entities

I annotated seven types of word entities in the DFI corpus: 'drug', 'well known target', 'drug metabolizer', 'drug transporter', 'food', 'food component', and 'ambiguous'. To prevent that some words were simultaneously recognized as both a food and a drug/drug-related molecule, I clearly defined 'food', 'drug' and 'well-known target' etc., and manually curated the word lists to make them mutually exclusive. The word lists for each entity were created based on drug information resources such as DrugBank and food database, FooDB. Annotators referred to the word lists and recognized a word as a proper entity type based on scientific reasoning about a

study design.

1.1 Drug/Drug Related Entities

'Drug', 'well known target', 'drug metabolizer', and 'drug transporter' belong to the drug/drug related entities. Annotators should refer to drug information resources such as ATC/DDD and DrugBank to determine a given word as drug/drug related entities.

ATC/DDD list contains information on new chemical entities or biologics which is ready for submission in at least one country, approved chemical entities, and approved herbal medicinal products. DrugBank is a free-to-access online database that contains information on experimental drugs as well as on FDA-approved drugs. Each entry contains several data fields including identification, properties, targets, enzymes, transporters, etc.

In this section, I provide a definition for each drug/drug related entity, which will help annotators to decide which word entity annotators should label.

1.1.1 'Drug'

Basically, 'drug' refers to any drug or its synonym included in the 'drug list' created based on ATC/DDD list and DrugBank. I defined 'drug' as a medicine or substance used for treating or preventing a disease or alleviating its symptoms in a given study design (Figure 7.1). Therefore, if insulin was administered to control blood glucose level, the word insulin should be annotated as 'drug', not 'well known target'.

Paper Title: Phase I study of temozolomide in combination with thiotepa and carboplatin with autologous hematopoietic cell rescue in patients with malignant brain tumors with minimal residual disease. (PMID: 26726947)

Searching condition: Food – Marrow / **Drug –** Carboplatin

Abstract (entity tagging)

Recurrence of malignant brain tumors results in a poor prognosis with limited treatment options. High-dose chemotherapy with autologous hematopoietic cell rescue (AHCR) has been used in patients with recurrent malignant brain tumors and has shown improved outcomes compared with standard chemotherapy. Temozolomide<drug> is standard therapy for glioblastoma and has also shown activity in patients with medulloblastoma/primitive neuro-ectodermal tumor (PNET), particularly those with recurrent disease. Temozolomide<drug> was administered twice daily on days -10 to -6, followed by thiotepa<drug> 300 mg/m(2) per day and carboplatin<drug> dosed using the Calvert formula or body surface area on days -5 to -3, with AHCR day 0. Twenty-seven patients aged 3-46 years were enrolled. Diagnoses included high-grade glioma (n=12); medulloblastoma/PNET (n=9); central nervous system (CNS) germ cell tumor (n=4); ependymoma (n=1) and spinal cord PNET (n=1). Temozolomide<drug> doses ranged from 100 mg/m(2) per day to 400 mg/m(2) per day. There were no toxic deaths. Prolonged survival was noted in several patients including those with recurrent high-grade glioma, medulloblastoma and CNS germ cell tumor. Increased doses of temozolomide<drug> are feasible with AHCR. A phase II study using temozolomide<drug>, carboplatin<drug> and thiotepa<drug> with AHCR for children with recurrent malignant brain tumors is being conducted through the Pediatric Blood and Marrow Transplant Consort

Figure 7.1: Example of annotated drug entities in an abstract

1.1.2 'Well Known Target'

'Well known target' list was created based on DrugBank (https://go.drugbank.com/targets). The list of 'well known target' consists of target entities that are known to be related to more than 5 drugs in DrugBank database (Figure 7.2). Furthermore, although a given word is not included in the 'well known target' list, I included all housekeeping genes or housekeeping gene-coded proteins expressed in the body.

Paper Title: Aegeline from Aegle marmelos stimulates glucose transport via Akt and Rac1 signaling, and contributes to a cytoskeletal rearrangement through PI3K/Rac1. (PMID: 26102565)

Searching condition: Food – Aegle marmelos **/ Drug –** Botolinum toxin

Abstract (entity tagging)

Aegeline<food component> is an alkaloidal-amide, isolated from the leaves of Aegle marmelos<food> and have shown antihyperglycemic as well as antidyslipidemic activities in the validated animal models of type 2 diabetes mellitus. Here we delineate, aegeline<food component> enhanced GLUT4<well-known target> translocation mediated 2-deoxy-glucose uptake in both time and concentration-dependent manner. 2-deoxy-glucose uptake was completely stymied by the transport inhibitors (wortmannin and genistein) in C2C12 myotubes. Pharmacological inhibition of Akt<well-known target> (also known as protein kinase B<well-known target>) and Ras<well-known target>-related C3 botulinum toxin substrate 1<well-known target> (Rac1<well-known target>) suggest that both Akt<well-known target> and Rac1<well-known target> operate aegeline<food component>-stimulated glucose transport via distinct parallel pathways. Moreover, aegeline<food component> activates p21 protein-activated kinase 1<well-known target> (PAK1<well-known target>) and cofilin<well-known target> (an actin polymerization regulator). Rac1<well-known target> inhibitor (Rac1<well-known target> inhib II) and PAK1<well-known target> inhibitor (IPA-3) completely blocked aegeline<food component>-induced phosphorylation of cofilin<well-known target> and p21 protein-activated kinase 1<well-known target> (PAK1<well-known target>). In summary, these findings suggest that aegeline<food component> stimulates the glucose transport through Akt<well-known target> and Rac1<well-known target> dependent distinct parallel pathways and have cytoskeletal roles via stimulation of the PI3-kinase-Rac1<well-known target>-PAK1<well-known target>-cofilin<well-known target> pathway in the skeletal muscle cells. Therefore, multiple targets of aegeline<food component> in the improvement of insulin sensitivity of the skeletal muscle cells may be suggested.

Figure 7.2: Example of annotated well known target entities in an abstract

1.1.3 'Drug Metabolizer'

'Drug metabolizer' list was created based on DrugBank. I defined 'drug metabolizer' as any metabolic enzyme well known for its involvement in drug metabolism in a human body. Cytochrome P450 2A13 and superoxide dismutase are examples of 'drug metabolizer'.

Paper Title: Grapefruit juice decreases the systemic availability of itraconazole capsules in healthy volunteers. (PMID: 10365642)

Searching condition: Food – Grapefruit / **Drug –** Itraconazole

Abstract (entity tagging)

The systemic availability of itraconazole capsules<drug> may be reduced secondary to elevated gastric pH and possibly by presystemic intestinal metabolism via CYP3A4<drug metabolizer>. <DFI key-sentence> Grapefruit juice<food> is acidic and an inhibitor of intestinal CYP3A4<drug metabolizer>. <DFI key-sentence> To determine the effect of grapefruit juice<food> on the systemic availability of itraconazole<drug> capsules, serum itraconazole<drug> and hydroxy-itraconazole<drug> concentrations were determined in eleven healthy volunteers studied in a randomized, two-way crossover design. Concurrent grapefruit juice<food> resulted in a 43% decrease in the mean itraconazole<drug> AUC0-48 (2507 ng x hr/mL versus 1434 ng x hr/mL, p = 0.046) and a 47% decrease in the mean hydroxy-itraconazole<drug> AUC0-72 (7264 ng x hr/mL versus 3880 ng x hr/mL, p = 0.025). Grapefruit juice also significantly increased the mean itraconazole<drug> Tmax (5.5 versus 4 hours). We conclude that concomitant grapefruit juice<food> does not enhance the systemic availability of itraconazole capsules<drug>, but rather appears to impair itraconazole<drug> absorption. Therefore, concomitant grapefruit juice<food> will not likely be useful in improving the oral availability of itraconazole<drug> capsules.

Figure 7.3: Example of annotated drug metabolizer entities in an abstract

1.1.4 'Drug Transporter'

'Drug transporter' list was created was based on DrugBank. I defined 'drug transporter' as any transport molecule involved in the transport and distribution of drugs in a human body. OATP1B1 and P-glycoprotein (Multidrug resistance protein 1) are examples of 'drug transporter'.

1.2 Food/Food Related Entities

In this section, I provide definition for 'food' and 'food component' entity. Annotators should refer to the food database, FooDB (https://foodb.ca/) to consider given word as food/food related entities.

1.2.1 'Food'

'Food' list was created based on FooDB. 'Food' refers to any word, its synonym or acronym included in the list. Thiamine, salvia, fermented milk, etc. are examples of 'food'.

1.2.2 'Food Component'

'Food component' list was created based on FooDB. I defined 'food component' as any food substance such as minerals, carbohydrate and fatty acid. I additionally included a plant from which a food substance originates, or an active ingredient of a plant as 'food component'. Oleic acid, stearic acid, Fe, etc. are examples of 'food component'.

Paper Title: Fatty acid composition of seed oil of different Sorghum bicolor varieties. (PMID: 26050001)

Searching condition: Food – Sorghum / **Drug** – Azelaic acid

Abstract (entity tagging)

In order to find out new sources of premium quality edible oil in the country, seeds of ten varieties of Sorghum bicolor<food> were initially analyzed for their total oil contents. The seed oil was later fractionated into eight fatty acids including two new saturated fatty acids. The oil contents were determined by Soxhlet method and compared with the results obtained by NMR analysis. The total oil contents in the seeds of sorghum<food> ranged from 5.0 to 8.2 % (w/w), indicating non significant difference obtained by two different techniques. The results revealed that oleic acid<food component> (31.12-48.99%), Palmitoleic acid<food component> (0.43-0.56%), linoleic acids<food component> (27.59-50.73%), linolenic acid<food component> (1.71-3.89%), stearic acid<food component> (1.09-2.59%) and palmitic acid<food component> (11.73-20.18%) was present in the seed oil of different sorghum<food> varieties when analyzed by GC-MS. It was observed that in most of the varieties polyunsaturated fatty acids (PUFA) were higher than monounsaturated fatty acids (MUFA). The two atypical SFAs, octanedioic<food component> (C8:0) and azelaic acid<food component> (C9:0) were found in some varieties. These results suggest that these S. bicolor<food> varieties could be additional sources of edible oil due to presence of clinically important saturated and high concentration of unsaturated fatty acids. A large scale production of the seed oil after refining process can contribute towards alleviation of edible oil shortage in the

Figure 7.4: Example of annotated food and food component entities in an abstract

1.3 'Ambiguous'

'Ambiguous' refers to words that are listed both on drug/food lists or words that are unclear to judge in which list (drug or food) they belong to.

5 Sentence Entities

This section provides a clear guidance and examples clarifying each sentence entity. I proposed

four types of sentence entities: ‘DFI key-sentence’, ‘Food-effect key-sentence’, ‘DDI key-sentence’, and ‘supporting sentence’.

5.1 Key-sentence

Key sentence consists of ‘DFI key-sentence’, ‘food-effect key-sentence’ and ‘DDI key-sentence’.

5.1.1 ‘DFI key-sentence’

‘DFI key-sentence’ refers to a sentence that contains any DFI about at least one entity pair, which is a combination of any drug/drug related entity and any food/food related entity. I defined DFI as below:

(1) The change of major pharmacological properties such as a total exposure to a drug, the efficacy and safety of a drug with the intake or ingestion of foods or food components

(2) When the intake or ingestion of food or food components affected an activity of drug metabolizer, drug transporter, or drug target molecules.

On the other hand, following cases are not included in ‘DFI key-sentence’.

(1) When a drug changes the effect of food or the effect of food component.

(2) If an article does not provide any direct evidence.

(3) Sentences which does not contain DFI information specified in 3.1.1, but contains information as follows:

- Research methodology (subject, test dose, analysis method, etc.)
- Dose change
- Alternative prescription information due to Food-Drug Interaction

5.1.2 'Food-effect key-sentence'

A 'food-effect key-sentence' refers to a sentence that provides information about how food intakes have an effect on the bioavailability of a drug.

5.1.3 'DDI key-sentence'

DDI is defined as a change in the effects of one drug by the presence of another drug. The effects may be an unexpected effect, a change in toxicity, treatment failure, pharmacological effect, etc.

Paper Title: Comparing the effects of Portulaca oleracea seed hydro-alcoholic extract, valsartan, and vitamin E on hemodynamic changes, oxidative stress parameters and cardiac hypertrophy in thyrotoxic rats. (PMID: 3141619)

Searching condition: Food – Portulaca oleracea / **Drug –** Levothryoxine

Abstract (entity tagging)

The present study compared the effects of Portulaca oleracea<food> (P. oleracea<food>) seed hydro-alcoholic extract, valsartan<drug>, and vitamin E<food component> on hemodynamic changes, oxidative stress markers and cardiac hypertrophy in a model of thyrotoxicosis. The hyperthyroid state was induced by intraperitoneal injection of levothyroxine<drug> (100??g/kg) for 4 weeks in male adult rats. After 2 weeks, vitamin E<food component> (20?mg/kg), valsartan<drug> (8?mg/kg), and P. oleracea<food> seed extract (400?mg/kg) were administered in three groups of thyrotoxic rats. The control group was given a daily injection of normal saline. Systolic blood pressure and heart rate were measured on three occasions with tail cuff. At the end of the fourth week, the animals were scarified and serum samples and heart tissue were collected for biochemical and histological studies. The levothyroxine<drug> increased heart rate and systolic blood pressure. A lower heart rate and reduced systolic blood pressure were observed in groups receiving valsartan<drug> and P. oleracea<food> extract. The heart weight/body weight ratio increased in groups treated with levothyroxine<drug>, but in a microscopic study, cardiomyocyte width was not different between the groups. Levothyroxine<drug> increased the level of malondyaldehide and NO metabolite but reduced the thiol concentration, superoxide dismutase<drug metabolizer>, and catalase<well-known target> activities. <DDI key-sentence> However, treatment with vitamin E and P. oleracea<food> extract increased the thiol concentration, superoxide dismutase<drug metabolizer> and catalase<well-known target> activities while decreasing malondyaldehide level. <DFI key-sentence> In addition, treatment with P. oleracea<food> extract and valsartan<drug> decreased NO metabolite level. Treatment with P. oleracea<food> extract improved levothyroxine<drug> induced oxidative stress and hemodynamic changes. These effects may be for antioxidant compone

Figure 7.5: Example of annotated DFI and DDI key-sentences in an abstract

5.2 'Supporting sentence'

A 'supporting sentence' by itself does not provide information about the occurrence of drug/food interactions but must be read in advance to understand a following key-sentence.

Paper Title: Spirulina Platensis Affects Factors Involved in Spermatogenesis and Increases Ghrelin Receptors in Testis Tissue of Rats Fed a High-Fat Diet. (PMID: 29166288)
Searching condition: Food – Spirulina / **Drug –** Testosterone

Abstract (entity tagging)

Ghrelin<well-known target> is a peptide hormone which plays important role in maintaining growth hormone release and energy homeostasis in vertebrates. Spirulina platensis<food> (SP<food>) has antioxidant and hypolipidemic effects due to its ingredients. In this study we aimed to investigate the effects of SP<food> on the testicular structure and relation between ghrelin<well-known target> and testosterone in the testis of rats fed a high fat diet (HFD). Sixty four young adult male rats were used and divided to 8 equal groups. Experimental groups received addition of 10% cholesterol<food> (CHL<food>), 43% hydrogenated vegetable oil<food> (HVO<food>) and 3% SP alone or in combination to basal diet while the control group received only basal diet. <supporting sentence> Serum ghrelin<well-known target> and testosterone levels were measured with ELISA. Receptors for ghrelin<well-known target> and androgen were detected with immunohistochemistry. For histomorphometric investigation, tubulus seminiferus, intertubular area, tubulus seminiferus lumen, Leydig cell nucleus, Sertoli cell nucleus, germ cell nucleus, spermatocyte nucleus and elongated spermatid volume densities were determined stereologically. Serum ghrelin<well-known target> level was increased especially in HVO<food> and CHL<food> combination group compared to the control while serum ghrelin<well-known target> levels were close to control levels in SP<food>-received groups. Ghrelin receptor<well-known target> level was increased in tubulus seminiferus with HVO<food>+CHL<food> administration but this effect was, however, limited in HVO<food>+CHL<food> and SP<food> challenged groups. <DFI key-sentence> HVO<food>+CHL<food> administration caused a significant decrease in Leydig cell nucleus volume density, as well as in all SP<food>-received groups, compared to the control. Significantly increased spermatocyte nucleus volume density in cholesterol<food>-receiving groups was decreased to control level with SP<food> alone and its combinations.

Figure 7.6: Example of annotated supporting sentences in an abstract

6 Relations

In this section, I provide a guidance how to annotate relation between word entities, and between word and sentence entities.

6.1 Relation between word entities

6.1.1 Synonym

Table 7.5: Relation between entities representing synonyms and definition of synonym relation

Relation	'Drug' ↔ 'Drug', 'Drug Metabolizer' ↔ 'Drug Metabolizer', 'Drug Transporter' ↔ 'Drug Transporter', 'Well Known Target' ↔ 'Well Known Target', 'Food' ↔ 'Food', 'Food Component' ↔ 'Food Component'
Definition	If the same object is expressed differently in the given abstract, such as an abbreviation or development name, annotators should tag a relation between synonyms.

6.1.2 Food and food component

Table 7.6: Relation between food and food component entities and definition of food component relation

Relation	'Food' ↔ 'Food Component'
Definition	Annotators should tag a relation between food and its food component.

6.2 Relation between word and sentence entities

6.2.1 Relation with 'DFI key-sentence'

Table 7.7: Relation between DFI key-sentence and word entities representing DFI and definition of relation

Relation	'DFI key-sentence' ↔ 'Drug', 'Food', 'Food Component', 'Ambiguous', 'Drug Metabolizer', 'Drug Transporter', 'Well Known Target'
Definition	'DFI key-sentence' must contain at least one relation between drug/drug related entity and food/food related entity. If the key word (drug/food entity) is expressed as a pronoun in the key sentence, the drug and food entity are selected as the closest noun from the sentence.

6.2.2 Relation with ‘DD key-sentence’

Table 7.8: Relation between DDI key-sentence and word entities representing DDI and definition of relation

Relation	‘DDI key-sentence’ ↔ ’Drug’, ‘Ambiguous’, ‘Drug Metabolizer’, ‘Drug Transporter’, ‘Well Known Target’
Definition	‘DDI key-sentence’ must contain at least two different drug/drug related entities. If the key word (drug entity) is expressed as a pronoun in the key sentence, the entity are selected as the closest noun from the sentence.

6.2.3 Relation with ‘Food-effect key-sentence’

Table 7.9: Relation between food-effect key-sentence and word entities representing food-effect and definition of relation

Relation	‘food-effect key-sentence’ ↔ ’Drug’, ‘Ambiguous’
Definition	‘Food-effect key-sentence’ must contain at least one relation with a drug entity. If the key word (drug entity) is expressed as a pronoun in the key sentence, the entity are selected as the closest noun from the sentence.

7 Entities & document labels

7.1 Sentence label (sentence modality)

I annotated DFI key-sentence as ‘positive’ if the study described in the abstract proved that there is an interaction between food and drug entities, and ‘negative’ if it proved there is not.

7.2 Document label (evidence-level)

In our corpus, I annotated given document with ‘evidence-level’. I proposed seven types of evidence levels: ‘clinical trial’, ‘observational study’, ‘case study’, ‘in-vivo study’, ‘in-vitro study’, ‘bioanalysis’, and ‘others’.

(1) ‘Clinical trial’: A clinical trial refers to a human trial that has evaluated the efficacy and safety of a drug or a medical procedure through randomly assigned patients regardless of whether or not the trial is blinded.

(2) ‘Observational study’: An observational study refers to a trial which has not randomly assigned patients or an analysis using existing health data such as EMR that has evaluated the efficacy and safety of a drug.

(3) ‘Case study’: A case study refers to a study that originally developed in epidemiology. In the case study, two groups within 10 patients differing in symptoms or outcome are compared retrospectively.

(4) ‘In-vivo study’: An in-vivo study refers to an experiment in which an animal or plant has been used to evaluate the efficacy or safety of a specific substance.

(5) ‘In-vitro study’: An in-vitro study refers to an experiment conducted using component of an organism extracted from animals or plants such as cells, molecules, proteins or enzymes.

(6) ‘Bioanalysis’: A bioanalysis refers to an experiment that analyzes the composition and content of foods or drugs.

(7) ‘others’: I annotated ‘others’ to studies where the above 6 types are not used or data is not directly produced from the article. Meta-analysis using clinical research results is also included in others

REFERENCES

1. WHO, *The importance of pharmacovigilance.* 2002.
2. Quan, C., et al., *Multichannel convolutional neural network for biological relation extraction.* BioMed research international, 2016. **2016**.
3. Wang, J., et al., *Adverse event detection by integrating twitter data and VAERS.* Journal of biomedical semantics, 2018. **9**(1): p. 1-10.
4. Sahu, S.K. and A. Anand, *Drug-drug interaction extraction from biomedical texts using long short-term memory network.* Journal of biomedical informatics, 2018. **86**: p. 15-24.
5. Zhang, Y., et al., *Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths.* Bioinformatics, 2018. **34**(5): p. 828-835.
6. Peng, Y., S. Yan, and Z. Lu, *Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets.* arXiv preprint arXiv:1906.05474, 2019.
7. Xue, L., et al., *mT5: A massively multilingual pre-trained text-to-text transformer.* arXiv preprint arXiv:2010.11934, 2020.
8. Clark, K., et al., *Electra: Pre-training text encoders as discriminators rather than generators.* arXiv preprint arXiv:2003.10555, 2020.
9. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
10. Bommasani, R., et al., *On the opportunities and risks of foundation models.* arXiv preprint arXiv:2108.07258, 2021.
11. Mazumder, M., et al., *DataPerf: Benchmarks for Data-Centric AI Development.* arXiv preprint arXiv:2207.10062, 2022.
12. Siun Kim, T.K., Tae Kyu Chung, Yoona Choi, YeSol Hong, Kyomin Jung, Howard Lee, *Extraction of Comprehensive Drug Safety Information from Adverse Drug Event Narratives in Spontaneous Reporting System.* (in press), 2023.
13. Kim, S., et al., *An annotated corpus from biomedical articles to construct a drug-food interaction database.* Journal of Biomedical Informatics, 2022. **126**: p. 103985.
14. Kim, S., S. Kim, and H. Lee, *A critical review of the United States regulatory pathways for determining the equivalence of efficacy between CT-P13 and original infliximab (Remicade®).* Drug Design, Development and Therapy, 2020. **14**: p. 2831.
15. Kim, S., et al., *A population pharmacokinetic-pharmacodynamic model of YH12852, a highly selective 5-hydroxytryptamine 4 receptor agonist, in healthy subjects and patients with functional constipation.* CPT: pharmacometrics & systems pharmacology, 2021. **10**(8): p. 902-913.
16. Brewer, T. and G.A. Colditz, *Postmarketing surveillance and adverse drug reactions: current perspectives and future needs.* Jama, 1999. **281**(9): p. 824-829.
17. Woodcock, J., R.E. Behrman, and G.J. Dal Pan, *Role of postmarketing surveillance in contemporary medicine.* Annual review of medicine, 2011. **62**: p. 1-10.
18. Pane, J., et al., *EU postmarket surveillance plans for medical devices.* Pharmacoepidemiology and drug safety, 2019. **28**(9): p. 1155-1165.
19. Lindquist, M. and I.R. Edwards, *The WHO Programme for International Drug Monitoring, its database, and the technical support of the Uppsala Monitoring Center.* The Journal of rheumatology, 2001. **28**(5): p. 1180-1187.
20. WHO. *The WHO Programme for International Drug Monitoring*. 2022 [cited 2022 2022, Nov 21]; Available from: https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance/health-professionals-

info/pidm.
21. FDA, *Good pharmacovigilance practices and pharmacoepidemiologic assessment*. 2019.
22. EMA, *Guideline on good pharmacovigilance practices: Module I – Pharmacovigilance systems and their quality systems*. 2012.
23. Hazell, L. and S.A. Shakir, *Under-reporting of adverse drug reactions.* Drug safety, 2006. **29**(5): p. 385-396.
24. Bergvall, T., G.N. Norén, and M. Lindquist, *vigiGrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues.* Drug safety, 2014. **37**(1): p. 65-77.
25. Lindquist, A.M., *Seeing and observing in international pharmacovigilance: achievements and prospects in worldwide drug safety*. 2003: [Uppsala]: Uppsala Monitoring Centre.
26. Rowin, E.J., et al., *Does error and adverse event reporting by physicians and nurses differ?* The Joint Commission Journal on Quality and Patient Safety, 2008. **34**(9): p. 537-545.
27. Sarker, A., et al., *Utilizing social media data for pharmacovigilance: a review.* Journal of biomedical informatics, 2015. **54**: p. 202-212.
28. Jagannatha, A., et al., *Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0).* Drug safety, 2019. **42**(1): p. 99-111.
29. Liu, S., et al., *Drug-drug interaction extraction via convolutional neural networks.* Computational and mathematical methods in medicine, 2016. **2016**.
30. Lim, S., K. Lee, and J. Kang, *Drug drug interaction extraction from the literature using a recursive neural network.* PloS one, 2018. **13**(1): p. e0190926.
31. Wu, H.-Y., et al., *An integrated pharmacokinetics ontology and corpus for text mining.* BMC bioinformatics, 2013. **14**(1): p. 1-15.
32. Radford, A., et al., *Language models are unsupervised multitask learners.* OpenAI blog, 2019. **1**(8): p. 9.
33. Liu, Y., et al., *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.
34. Raffel, C., et al., *Exploring the limits of transfer learning with a unified text-to-text transformer.* J. Mach. Learn. Res., 2020. **21**(140): p. 1-67.
35. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.
36. Mikolov, T., et al., *Advances in pre-training distributed word representations.* arXiv preprint arXiv:1712.09405, 2017.
37. Schuster, M. and K.K. Paliwal, *Bidirectional recurrent neural networks.* IEEE transactions on Signal Processing, 1997. **45**(11): p. 2673-2681.
38. Mikolov, T., W.-t. Yih, and G. Zweig. *Linguistic regularities in continuous space word representations*. in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013.
39. Battaglia, P.W., et al., *Relational inductive biases, deep learning, and graph networks.* arXiv preprint arXiv:1806.01261, 2018.
40. Tran, K., A. Bisazza, and C. Monz, *The importance of being recurrent for modeling hierarchical structure.* arXiv preprint arXiv:1803.03585, 2018.
41. Hochreiter, S., *The vanishing gradient problem during learning recurrent neural nets and problem solutions.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998. **6**(02): p. 107-116.
42. Yu, Y., et al., *A review of recurrent neural networks: LSTM cells and network architectures.* Neural computation, 2019. **31**(7): p. 1235-1270.
43. Graves, A. and J. Schmidhuber, *Framewise phoneme classification with bidirectional*

LSTM and other neural network architectures. Neural networks, 2005. **18**(5-6): p. 602-610.
44. Huang, Z., W. Xu, and K. Yu, *Bidirectional LSTM-CRF models for sequence tagging.* arXiv preprint arXiv:1508.01991, 2015.
45. Peters, M.E., et al., *Semi-supervised sequence tagging with bidirectional language models.* arXiv preprint arXiv:1705.00108, 2017.
46. Peters, M., et al., *Deep contextualized word representations. arXiv 2018.* arXiv preprint arXiv:1802.05365, 1802. **12**.
47. Sarzynska-Wawer, J., et al., *Detecting formal thought disorder by deep contextualized word representations.* Psychiatry Research, 2021. **304**: p. 114135.
48. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
49. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
50. Yang, Z., et al., *Xlnet: Generalized autoregressive pretraining for language understanding.* Advances in neural information processing systems, 2019. **32**.
51. Xia, P., S. Wu, and B. Van Durme, *Which* BERT? A survey organizing contextualized encoders.* arXiv preprint arXiv:2010.00854, 2020.
52. Sun, C., et al. *How to fine-tune bert for text classification?* in *China national conference on Chinese computational linguistics*. 2019. Springer.
53. Wolf, T., et al. *Transformers: State-of-the-art natural language processing*. in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020.
54. Gururangan, S., et al., *Don't stop pretraining: adapt language models to domains and tasks.* arXiv preprint arXiv:2004.10964, 2020.
55. Gu, Y., et al., *Domain-specific language model pretraining for biomedical natural language processing.* ACM Transactions on Computing for Healthcare (HEALTH), 2021. **3**(1): p. 1-23.
56. Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2020. **36**(4): p. 1234-1240.
57. Alsentzer, E., et al., *Publicly available clinical BERT embeddings.* arXiv preprint arXiv:1904.03323, 2019.
58. raj Kanakarajan, K., B. Kundumani, and M. Sankarasubbu. *BioELECTRA: pretrained biomedical text encoder using discriminators*. in *Proceedings of the 20th Workshop on Biomedical Language Processing*. 2021.
59. Kim, Y., et al., *A pre-trained BERT for Korean medical natural language processing.* Scientific Reports, 2022. **12**(1): p. 1-10.
60. Henry, S., et al., *2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records.* Journal of the American Medical Informatics Association, 2020. **27**(1): p. 3-12.
61. Karimi, S., et al., *Cadec: A corpus of adverse drug event annotations.* Journal of biomedical informatics, 2015. **55**: p. 73-81.
62. Roberts, K., D. Demner-Fushman, and J.M. Tonning, *Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track.* In TAC 2017, 2017.
63. Weissenbacher, D., et al., *Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018.* In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, 2018: p. 13-16.
64. Weissenbacher, D., et al., *Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019.* In Proceedings of the fourth social media mining for health

applications (# SMM4H) workshop & shared task, 2019: p. 21-30.
65. Klein, A., et al., *Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020.* In Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2020: p. 27-36.
66. Magge, A., et al. *Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021*. in *In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*. 2021.
67. Magge, A., et al., *DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter.* Journal of the American Medical Informatics Association, 2021. **28**(10): p. 2184-2192.
68. Du, J., et al., *Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning.* Journal of the American Medical Informatics Association, 2021. **28**(7): p. 1393-1400.
69. Chopard, D., et al., *Text mining of adverse events in clinical trials: Deep learning approach.* JMIR Medical Informatics, 2021. **9**(12): p. e28632.
70. Botsis, T., et al., *Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection.* Journal of the American Medical Informatics Association, 2011. **18**(5): p. 631-638.
71. WHO-UMC. *The use of the WHO-UMC system for standardized case causality assessment.* 2018; Available from: https://who-umc.org/media/164200/who-umc-causality-assessment_new-logo.pdf.
72. ICH. *E2B(R3) Electronic Transmission of Individual Case Safety Reports (ICSRs) Implementation Guide – Data Elements and Message Specification: Guidance for Industry*. 2022; Available from: https://www.fda.gov/media/81904/download.
73. Aranaz-Andrés, J.M., et al., *What makes hospitalized patients more vulnerable and increases their risk of experiencing an adverse event?* International Journal for Quality in Health Care, 2011. **23**(6): p. 705-712.
74. Sousa, P., et al., *Patient and hospital characteristics that influence incidence of adverse events in acute public hospitals in Portugal: a retrospective cohort study.* International Journal for Quality in Health Care, 2018. **30**(2): p. 132-137.
75. WHO, *The importance of pharmacovigilance*. 2002: World Health Organization.
76. Huang, Y.L., J. Moon, and J.B. Segal, *A comparison of active adverse event surveillance systems worldwide.* Drug Saf, 2014. **37**(8): p. 581-96.
77. Alomar, M., et al., *Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future.* Ther Adv Drug Saf, 2020. **11**: p. 2042098620938595.
78. KIDS. *Pharmacovigillance - Statistics on Reported ICSRs*. 2022 [cited 2022 6 May]; Available from: https://www.drugsafe.or.kr/iwt/ds/en/report/EgovICSRStatistics.do.
79. Oh, I.-S., et al., *Differential completeness of spontaneous adverse event reports among hospitals/clinics, pharmacies, consumers, and pharmaceutical companies in South Korea.* PloS one, 2019. **14**(2): p. e0212336.
80. KNARS. *Inspection and Improvement Plan for Drug Adverse Event Reporting System, Vol. 33*. 2019; Available from: https://www.nars.go.kr/fileDownload2.do?doc_id=1My351ygXsf&fileName=(%EC%9E%85%EB%B2%95%E3%86%8D%EC%A0%95%EC%B1%85%EB%B3%B4%EA%B3%A0%EC%84%9C%2033%ED%98%B8-20191219)%EC%9D%98%EC%95%BD%ED%92%88%20%EC%9D%B4%EC%83%81%EC%82%AC%EB%A1%80%20%EB%B3%B4%EA%B3%A0%EC%A0%9C%EB%8F%84%EC%9D%98%20%EC%A0%90%EA%B2%80%20%EB%B0%8F%20%EA%B0%9C%EC%84%A0%EB%B0%A9%EC%95%88.pdf.

81. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and psychological measurement, 1960. **20**(1): p. 37-46.
82. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* In the 3rd International Conference for Learning Representations, 2014.
83. Zhang, J., et al., *Why gradient clipping accelerates training: A theoretical justification for adaptivity.* arXiv preprint arXiv:1905.11881, 2019.
84. Sang, E.F. and S. Buchholz, *Introduction to the CoNLL-2000 shared task: Chunking.* In Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop, 2000.
85. Lafferty, J., A. McCallum, and F.C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the 18th International Conference on Machine Learning 2001, 2001.
86. Forney, G.D., *The viterbi algorithm.* In Proceedings of the IEEE, 1973. **61**(3): p. 268-278.
87. Soares, L.B., et al., *Matching the blanks: Distributional similarity for relation learning.* In the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
88. Bommasani, R., et al., *On the opportunities and risks of foundation models.* 2021.
89. MFDS, *Suspected Unexpected Serious Adverse Drug Reaction(SUSAR) Review*. 2020.
90. Program, N.C.I.C.T.E., *Common Terminology Criteria for Adverse Events:(CTCAE)*. 2003: Cancer Therapy Evaluation Program.
91. Li, Y., A. Jimeno Yepes, and C. Xiao, *Combining social media and FDA Adverse Event Reporting System to detect adverse drug reactions.* Drug safety, 2020. **43**(9): p. 893-903.
92. Heeley, E., et al., *Prescription-event monitoring and reporting of adverse drug reactions.* The Lancet, 2001. **358**(9296): p. 1872-1873.
93. Martin, R.M., et al., *Underreporting of suspected adverse drug reactions to newly marketed ("black triangle") drugs in general practice: observational study.* Bmj, 1998. **317**(7151): p. 119-120.
94. Alatawi, Y.M. and R.A. Hansen, *Empirical estimation of under-reporting in the US Food and Drug Administration adverse event reporting system (FAERS).* Expert opinion on drug safety, 2017. **16**(7): p. 761-767.
95. Khalili, M., et al., *Estimation of adverse drug reaction reporting in Iran: Correction for underreporting.* Pharmacoepidemiology and Drug Safety, 2021. **30**(8): p. 1101-1114.
96. Jáuregui-Garrido, B. and I. Jáuregui-Lobera, *Interactions between antihypertensive drugs and food.* Nutrición hospitalaria, 2012. **27**(6): p. 1866-1875.
97. Lau, W.C., et al., *Atorvastatin reduces the ability of clopidogrel to inhibit platelet aggregation: a new drug–drug interaction.* Circulation, 2003. **107**(1): p. 32-37.
98. Sun, X., et al., *Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss.* Entropy, 2019. **21**(1): p. 37.
99. Herrero-Zazo, M., et al., *The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions.* Journal of biomedical informatics, 2013. **46**(5): p. 914-920.
100. Wang, L.L., et al. *SUPP.AI: finding evidence for supplement-drug interactions*. in *ACL*. 2020.
101. Rosenkranz, B., P. Fasinu, and P. Bouic, *An overview of the evidence and mechanisms of herb–drug interactions.* Frontiers in pharmacology, 2012. **3**: p. 69.
102. Hamon, T., et al. *POMELO: Medline corpus with manually annotated food-drug interactions*. in *BiomedicalNLP@ RANLP*. 2017.
103. Boyce, R., G. Gardner, and H. Harkema, *Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts*, in *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. 2012, Association for Computational Linguistics: Montreal, Canada. p. 206–213.

104. Karnik, S., et al., *Extraction of drug-drug interactions using all paths graph kernel.* Proc. of the 1st Challenge task on Drug Drug Interaction Extraction, 2011: p. 83-88.
105. Collobert, R. and J. Weston. *A unified architecture for natural language processing: Deep neural networks with multitask learning*. in *Proceedings of the 25th international conference on Machine learning*. 2008.
106. Gu, Y., et al., *Domain-specific language model pretraining for biomedical natural language processing.* arXiv preprint arXiv:2007.15779, 2020.
107. NCBI. *PubMed Help: MeSH Terms [MH]. (updated March 31, 2020)*. 2005; Available from: https://www.ncbi.nlm.nih.gov/books/NBK3827/.
108. Kans, J., *Entrez direct: E-utilities on the UNIX command line*, in *Entrez Programming Utilities Help [Internet]*. 2020, National Center for Biotechnology Information (US).
109. Harrington, R.A., et al., *Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure.* BMJ open, 2019. **9**(6): p. e026652.
110. FDA, *Assessing the Effects of Food on Drugs in INDs and NDAs – Clinical Pharmacology Considerations*. 2019, Center for Drug Evaluation and Research.
111. Cejuela, J.M., et al., *tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles.* Database, 2014. **2014**.
112. Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2019. **36**(4): p. 1234-1240.
113. Huang, K., J. Altosaar, and R. Ranganath, *Clinicalbert: Modeling clinical notes and predicting hospital readmission.* arXiv preprint arXiv:1904.05342, 2019.
114. Sohn, S.-I., et al., *The ameliorative effect of 23-hydroxytormentic acid isolated from Rubus coreanus on cisplatin-induced nephrotoxicity in rats.* Biological and Pharmaceutical Bulletin, 2011. **34**(9): p. 1508-1513.
115. Kwon, M., et al., *Organic cation transporter-mediated drug–drug interaction potential between berberine and metformin.* Archives of Pharmacal Research, 2015. **38**(5): p. 849-856.
116. Misaka, S., et al., *Green tea ingestion greatly reduces plasma concentrations of nadolol in healthy subjects.* Clinical Pharmacology & Therapeutics, 2014. **95**(4): p. 432-438.
117. Uddin, M.J., et al., *Pharmacotherapy against oxidative stress in chronic kidney disease: Promising small molecule natural products targeting nrf2-ho-1 signaling.* Antioxidants, 2021. **10**(2): p. 258.
118. Nduka, S.O., M.J. Okonta, and C.O. Esimone, *Effects of Zingiber officinale on the plasma pharmacokinetics and lung penetrations of ciprofloxacin and isoniazid.* American Journal of Therapeutics, 2013. **20**(5): p. 507-513.
119. Peng, L.-Q., et al., *Pyridinium ionic liquid-based liquid-solid extraction of inorganic and organic iodine from Laminaria.* Food chemistry, 2018. **239**: p. 1075-1084.
120. Pošćić, F., et al., *Effects of cerium and titanium oxide nanoparticles in soil on the nutrient composition of barley (Hordeum vulgare L.) kernels.* International Journal of Environmental Research and Public Health, 2016. **13**(6): p. 577.
121. Dekel, Y., et al., *Formation of multimeric antibodies for self-delivery of active monomers.* Drug Delivery, 2017. **24**(1): p. 199-208.
122. Fang, K., et al., *Effects of integrated rice-frog farming on paddy field greenhouse gas emissions.* International journal of environmental research and public health, 2019. **16**(11): p. 1930.
123. Mannina, D., et al., *Reduced intensity allogeneic stem cell transplantation for younger patients with myelofibrosis.* British Journal of Haematology, 2019. **186**(3): p. 484-489.
124. Satoh, E., et al., *Black tea extract, thearubigin fraction, counteract the effects of botulinum neurotoxins in mice.* British journal of pharmacology, 2001. **132**(4): p. 797-798.

125. Shoko, T., et al., *Anti-aging potential of extracts from Sclerocarya birrea (A. Rich.) Hochst and its chemical profiling by UPLC-Q-TOF-MS.* BMC complementary and alternative medicine, 2018. **18**(1): p. 1-14.
126. Katta, N., S. Balla, and M.A. Alpert, *Does long-term furosemide therapy cause thiamine deficiency in patients with heart failure? A focused review.* The American Journal of Medicine, 2016. **129**(7): p. 753. e7-753. e11.
127. Takahashi, S. and Y. Nakashima, *Repeated and long-term treatment with physiological concentrations of resveratrol promotes NO production in vascular endothelial cells.* British journal of nutrition, 2012. **107**(6): p. 774-780.
128. Wang, X., et al., *Up-regulation of PAI-1 and down-regulation of uPA are involved in suppression of invasiveness and motility of hepatocellular carcinoma cells by a natural compound berberine.* International Journal of Molecular Sciences, 2016. **17**(4): p. 577.
129. Yoon, J.J., et al., *Oryeongsan suppressed high glucose-induced mesangial fibrosis.* BMC Complementary and Alternative Medicine, 2015. **15**(1): p. 1-11.
130. Alexandre, E.C., et al., *Chronic treatment with resveratrol improves overactive bladder in obese mice via antioxidant activity.* European journal of pharmacology, 2016. **788**: p. 29-36.
131. De Jonge, E., et al., *Impaired haemostasis by intravenous administration of a gelatin-based plasma expander in human subjects.* Thrombosis and haemostasis, 1998. **79**(02): p. 286-290.
132. van Mulligen, E.M., et al., *The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships.* Journal of Biomedical Informatics, 2012. **45**(5): p. 879-884.
133. Luo, Y.-F., W. Sun, and A. Rumshisky, *MCN: a comprehensive corpus for medical concept normalization.* Journal of biomedical informatics, 2019. **92**: p. 103132.
134. Hamon, T., et al. *POMELO: Medline corpus with manually annotated food-drug interactions*. in *Proceedings of the Biomedical NLP Workshop associated with RANLP 2017*. 2017.
135. Wu, X., J. Zhang, and H. Li, *Text-to-table: A new way of information extraction.* arXiv preprint arXiv:2109.02707, 2021.
136. Ma, Y., T. Hiraoka, and N. Okazaki, *Named entity recognition and relation extraction using enhanced table filling by contextualized representations.* Journal of Natural Language Processing, 2022. **29**(1): p. 187-223.
137. Blagec, K., et al., *Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals.* arXiv preprint arXiv:2201.07040, 2022.
138. Ouyang, L., et al., *Training language models to follow instructions with human feedback.* arXiv preprint arXiv:2203.02155, 2022.
139. Liu, X., et al., *GPT understands, too.* arXiv preprint arXiv:2103.10385, 2021.
140. Wei, J., et al., *Finetuned language models are zero-shot learners.* arXiv preprint arXiv:2109.01652, 2021.
141. Sanh, V., et al., *Multitask prompted training enables zero-shot task generalization.* arXiv preprint arXiv:2110.08207, 2021.
142. Agrawal, M., et al., *Large Language Models are Zero-Shot Clinical Information Extractors.* arXiv preprint arXiv:2205.12689, 2022.
143. Moradi, M., et al., *GPT-3 models are poor few-shot learners in the biomedical domain.* arXiv preprint arXiv:2109.02555, 2021.
144. Gutiérrez, B.J., et al., *Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again.* arXiv preprint arXiv:2203.08410, 2022.
145. Luo, R., et al., *BioGPT: generative pre-trained transformer for biomedical text generation and mining.* Briefings in Bioinformatics, 2022.
146. Jiang, Z., et al., *How can we know what language models know?* Transactions of the

Association for Computational Linguistics, 2020. **8**: p. 423-438.
147. Gao, T., A. Fisch, and D. Chen, *Making pre-trained language models better few-shot learners.* arXiv preprint arXiv:2012.15723, 2020.
148. Yuan, W., G. Neubig, and P. Liu, *Bartscore: Evaluating generated text as text generation.* Advances in Neural Information Processing Systems, 2021. **34**: p. 27263-27277.
149. Li, X.L. and P. Liang, *Prefix-tuning: Optimizing continuous prompts for generation.* arXiv preprint arXiv:2101.00190, 2021.
150. Tsimpoukelli, M., et al., *Multimodal few-shot learning with frozen language models.* Advances in Neural Information Processing Systems, 2021. **34**: p. 200-212.

초 록

약물 감시는 약물 부작용 또는 약물 안전성과 관련된 문제의 발생을 감지, 평가 및 이해하기 위한 과학적 활동이다. 그러나 약물 감시에 사용되는 의약품 안전성 정보의 보고 품질에 대한 우려가 꾸준히 제기되었으며, 해당 보고 품질을 높이기 위해서는 안전성 정보를 확보할 새로운 자료원이 필요하다. 한편 트랜스포머 아키텍처를 기반으로 사전훈련 언어모델이 등장하면서 다양한 도메인에서 자연어처리 기술 적용이 가속화되었다. 이러한 맥락에서 본 학위 논문에서는 약물 감시를 위한 다음 2가지 정보 추출 문제를 자연어처리 문제 형태로 정의하고 관련 기준 모델을 개발하였다: 1) 수동적 약물 감시 체계에 보고된 이상사례 서술자료에서 포괄적인 약물 안전성 정보를 추출한다. 2) 영문 의약학 논문 초록에서 약물-식품 상호작용 정보를 추출한다. 이를 위해 안전성 정보 추출을 위한 어노테이션 가이드라인을 개발하고 수작업으로 어노테이션을 수행하였다. 결과적으로 고품질의 자연어 학습데이터를 기반으로 사전학습 언어모델을 미세 조정함으로써 비정형 텍스트에서 임상 정보를 추출하는 강력한 자연어처리 모델 개발이 가능함을 확인하였다. 마지막으로 본 학위 논문에서는 약물감시와 관련된 임상 정보 추출을 위한 어노테이션 가이드라인을 개발할 때 고려해야 할 주의 사항에 대해 논의하였다. 본 학위 논문에서 소개한 자연어 학습데이터와 자연어처리 모델은 약물 안전성 정보의 보고 품질을 향상시키고 자료원을 확장하여 약물 감시 활동을 보조할 것으로 기대된다.

주요어: 약물 감시, 의약품 안전성 정보, 자연어처리, 정보 추출

학번: 2018-20603

감사의 글

부족함이 많은 학위 과정이지만 많은 분들의 도움과 격려 덕에 긴 시간을 마무리하는 단계까지 오게 되었습니다. 가장 먼저 치기 어린 학부생 인턴때부터 지금까지 세심히 지도해주시고 다양한 시도를 할 수 있게 격려해주신 이형기 교수님께 감사드립니다. 교수님의 믿음과 지원 덕에 독립연구자로 성장해나갈 수 있었습니다. 또한 부침이 많은 대학원생 생활을 함께 하면서 연구 내외로 경험과 고민을 나누고 제가 조금 더 '어른'이 될 수 있게 도와주신 CCADD 연구실 식구들에게도 감사드립니다. 또 공부하길 좋아하는 사람으로서의 저의 정체성을 누구보다 잘 알고 적절한 관심과 무관심을 가져준 가족들에게도 감사합니다. 학위 논문을 작성하면서 학위 과정 중에 채우지 못한 부분들을 떠올려보게 되었습니다. 제가 이제껏 배워 온 지식과 기술들로 풀 수 있는 문제, 그것들 중 적절히 포장해서 팔 수 있는 문제, 그리고 시급하게 풀어야하는 문제들이 무엇이며 얼마나 서로 떨어져 있는지 확인하고 있습니다. 앞으로도 꾸준히 노력해서 사회에 긍정적인 기여를 하는 연구자가 되고 싶습니다. 학위 과정을 격려해주신 모든 분들에게 감사드립니다.

2023년 2월,

김 시 언