**Annotation Guideline for the DFI corpus**

Siun Kim, Yoona Choi, Jung-Hyun Won, and Howard Lee

Center for Convergence Approaches in Drug Development, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Korea

(Last updated on August 9, 2021)

TABLE OF CONTENTS

[1. Introduction 2](#_Toc70598415)

[2. Word Entities 2](#_Toc70598416)

[2.1. Drug/Drug Related Entities 3](#_Toc70598417)

[2.1.1. ‘Drug’ 3](#_Toc70598418)

[2.1.2. ‘Well Known Target’ 4](#_Toc70598419)

[2.1.3. ‘Drug Metabolizer’ 5](#_Toc70598420)

[2.1.4. ‘Drug Transporter’ 6](#_Toc70598421)

[2.2. Food/Food Related Entities 6](#_Toc70598422)

[2.2.1. ‘Food’ 6](#_Toc70598423)

[2.2.2. ‘Food Component’ 6](#_Toc70598424)

[2.3. ‘Ambiguous’ 7](#_Toc70598425)

[3. Sentence Entities 7](#_Toc70598426)

[3.1. Key-Sentence 7](#_Toc70598427)

[3.1.1. ‘DFI Key-Sentence’ 7](#_Toc70598428)

[3.1.2. ‘Food-Effect Key-Sentence’ 8](#_Toc70598429)

[3.1.3 ‘DDI Key-Sentence’ 8](#_Toc70598430)

[3.2. ‘Supporting Sentence’ 9](#_Toc70598431)

[4. Relations 10](#_Toc70598432)

[4.1. Relation between Word Entities 10](#_Toc70598433)

[4.1.1. Synonyms 10](#_Toc70598434)

[4.1.2. Food and Food Component 10](#_Toc70598435)

[4.2. Relation between Word and Sentence Entities. 11](#_Toc70598436)

[4.2.1. Relation in ‘DFI-Key Sentence’ 11](#_Toc70598437)

[4.2.1. Relation in ‘DDI-Key Sentence’ 11](#_Toc70598438)

[4.2.2. Relation in ‘Food-Effect Key-Sentence’ 11](#_Toc70598439)

[5. Entities Labels 11](#_Toc70598440)

[5.1. Sentence Label (Sentence Modality) 11](#_Toc70598441)

[5.2. Document Label (Evidence-Level) 12](#_Toc70598442)

# Introduction

‘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.

In this annotation guideline, we defines annotation tasks and word and sentence entities annotated in the DFI corpus. Also, we 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 we created the DFI corpus. We 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, we 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.

# Word Entities

We 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, we 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.

## 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.

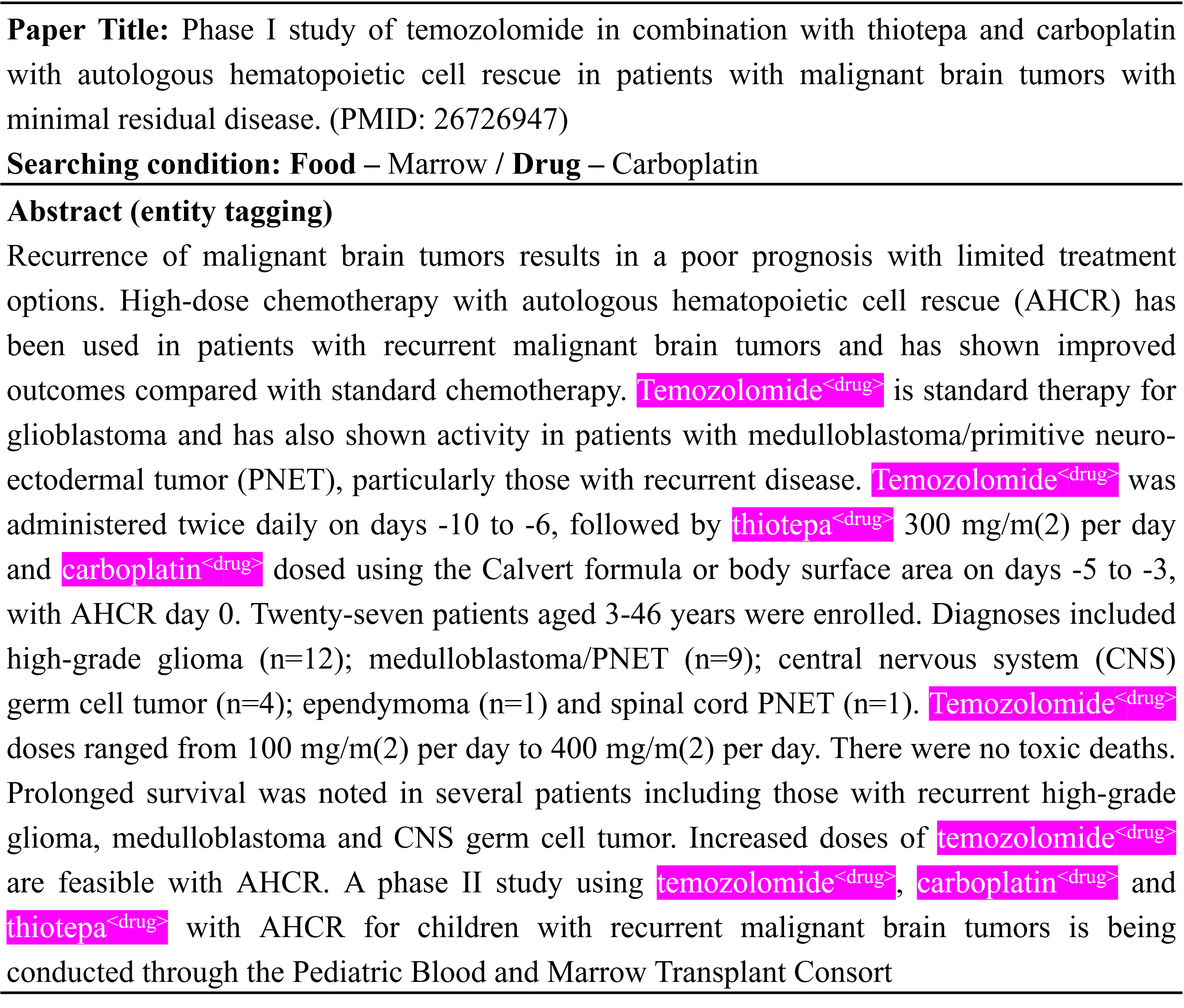
DrugBank is a free-to-access online database that contains information on experimental drugs as well as on FDA-approved drugs. Each entry contains several data fields including identification, properties, targets, enzymes, transporters, etc.

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

### ‘Drug’

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

Figure 1. Example of annotated drug entities in an abstract



### ‘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. Furthermore, although a given word is not included in the ‘well known target’ list, we included all housekeeping genes or housekeeping gene-coded proteins expressed in the body.

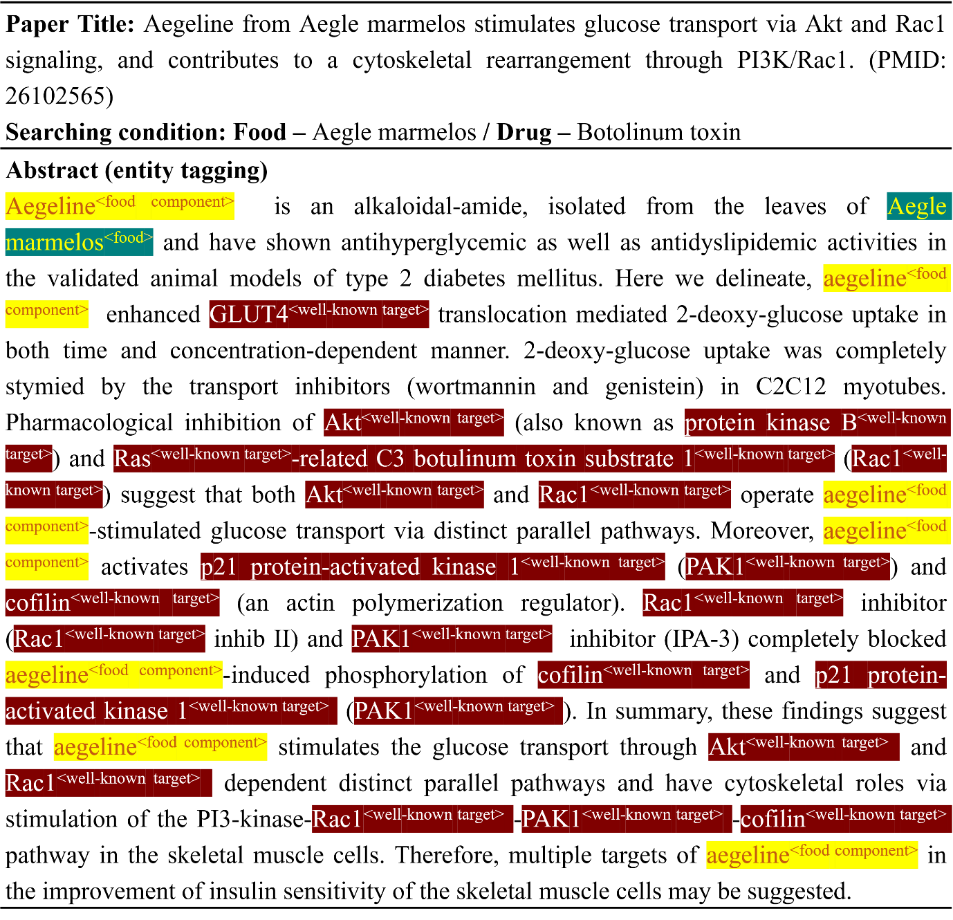


Figure 2. Example of annotated well known target entities in an abstract

### ‘Drug Metabolizer’

‘Drug metabolizer’ list was created based on DrugBank. We 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’.

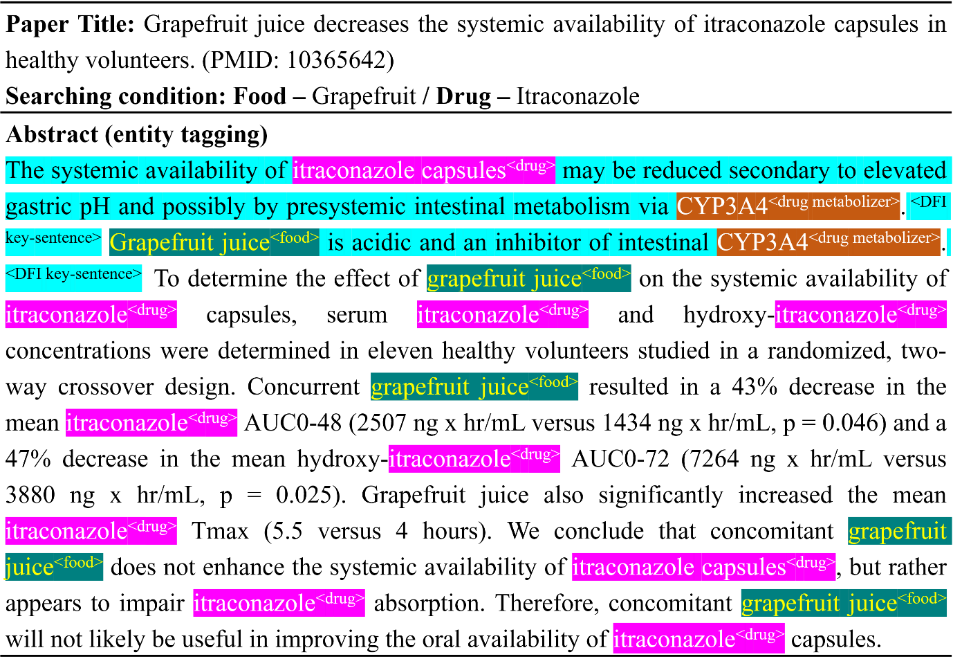


Figure 3. Example of annotated drug metabolizer entities in an abstract

### ‘Drug Transporter’

‘Drug transporter’ list was created was based on DrugBank. We 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’.

## Food/Food Related Entities

In this section, we 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.

### ‘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’.

### ‘Food Component’

‘Food component’ list was created based on FooDB. We defined ‘food component’ as any food substance such as minerals, carbohydrate and fatty acid. We 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’.

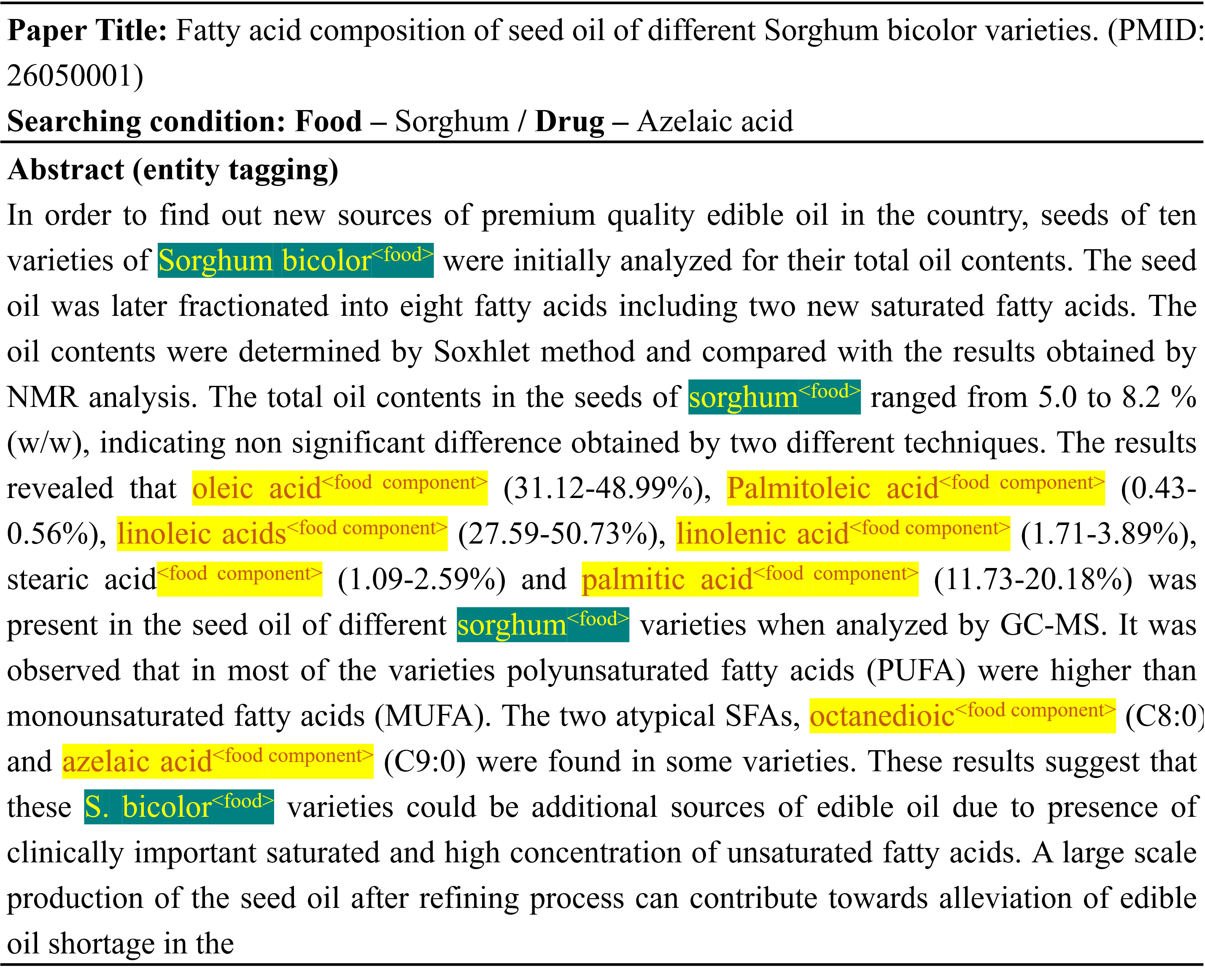


Figure 4. Example of annotated food and food component entities in an abstract

## ‘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.

# Sentence Entities

This section provides a clear guidance and examples clarifying each sentence entity. We proposed four types of sentence entities: ‘DFI key-sentence’, ‘Food-effect key-sentence’, ‘DDI key-sentence’, and ‘supporting sentence’.

## Key-Sentence

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

### ‘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. We 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

### ‘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.

### ‘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.

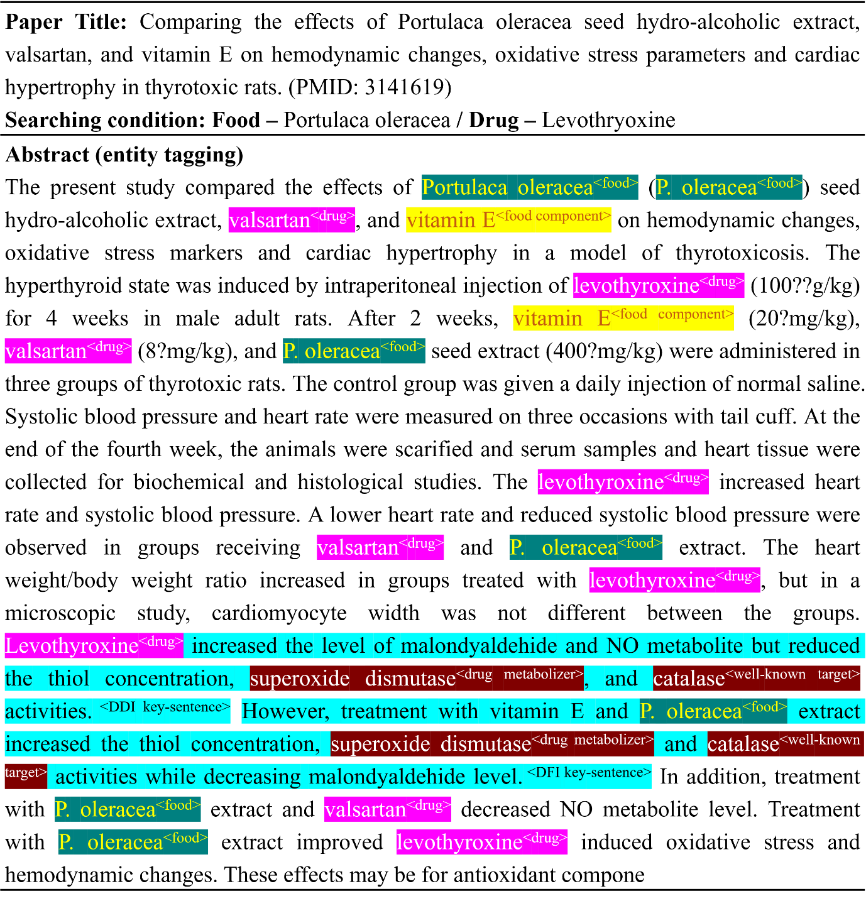
****

Figure 5. Example of annotated DFI and DDI key-sentences in an abstract

## ‘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.

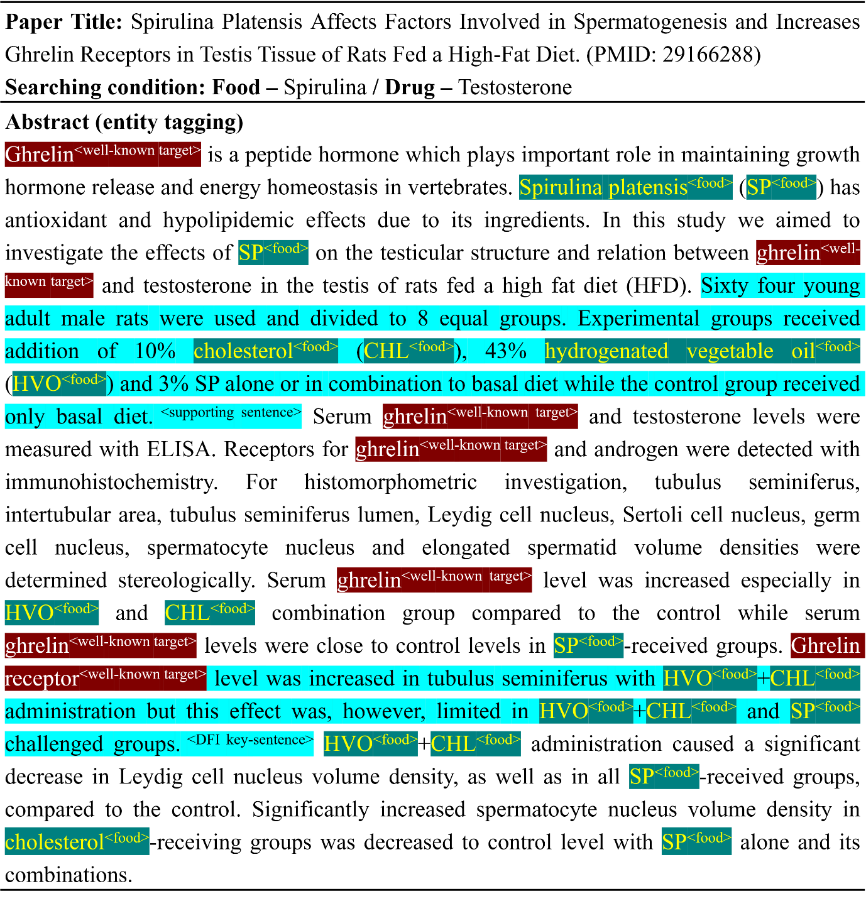


Figure 6.Example of annotated supporting sentences in an abstract

# Relations

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

## Relation between Word Entities

### Synonyms

Table 1. 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. |

### Food and Food Component

Table 2. 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. |

## Relation between Word and Sentence Entities.

### Relation in ‘DFI Key-Sentence’

Table 3. 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’ |
| Annotation Guideline | ‘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. |

### Relation in ‘DDI-Key Sentence’

Table 4. 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’ |
| Annotation Guideline | ‘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. |

### 

### Relation in ‘Food-Effect Key-Sentence’

Table 5. Relation between food-effect key-sentence and word entities representing food-effect and definition of relation

|  |  |
| --- | --- |
| Relation | ‘food-effect key-sentence’ ↔ ’Drug’, ‘Ambiguous’ |
| Annotation Guideline | ‘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. |

# Entities & Document Labels

## Sentence Label (Sentence Modality)

We 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.

## Document Label (Evidence-Level)

In our corpus, we annotated given document with ‘evidence-level’. We 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’: We 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