Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6769–6776
Marseille, 11–16 May 2020

Annotation of Adverse Drug Reactions in Patients' Weblogs

Yuki Arase[†], Tomoyuki Kajiwara[‡], Chenhui Chu[‡]
[†]Graduate School of Information Science and Technology, Osaka University
Yamada-oka 1-5, Suita, Osaka, Japan
[‡]Institute for Datability Science, Osaka University
Yamada-oka 2-8, Suita, Osaka, Japan
arase@ist.osaka-u.ac.jp, kajiwara@ids.osaka-u.ac.jp, chu@ids.osaka-u.ac.jp

Abstract

Adverse drug reactions are a severe problem that significantly degrade quality of life, or even threaten the life of patients. Patient-generated texts available on the web have been gaining attention as a promising source of information in this regard. While previous studies annotated such patient-generated content, they only reported on limited information, such as whether a text described an adverse drug reaction or not. Further, they only annotated short texts of a few sentences crawled from online forums and social networking services. The dataset we present in this paper is unique for the richness of annotated information, including detailed descriptions of drug reactions with full context. We crawled patient's weblog articles shared on an online patient-networking platform and annotated the effects of drugs therein reported. We identified spans describing drug reactions and assigned labels for related drug names, standard codes for the symptoms of the reactions, and types of effects. As a first dataset, we annotated 677 drug reactions with these detailed labels based on 169 weblog articles by Japanese lung cancer patients. Our annotation dataset is made publicly available for further research on the detection of adverse drug reactions and more broadly, on patient-generated text processing.

Keywords: Patient-generated text, effect of drugs, adverse drug reaction

1. Introduction

Drugs are one of the primary treatment options, and can have both therapeutic and side effects. Among these side effects, those causing undesired and harmful reactions are called adverse effects. Such adverse effects are a severe problem for patients, since they significantly degrade their quality of life and make the therapeutic approach unacceptable. They are particularly painful for patients who need long-term treatment, such as cancer patients. Therefore, information on adverse drug reactions is essential for medical practitioners to prescribe drugs that maximise the therapeutic effects while minimising the adverse drug reactions as much as possible. Moreover, it is also crucial for pharmaceutical science to understand the mechanisms that cause adverse drug reactions and develop new drugs.

There have been two primary sources of information for adverse drug reactions: reports of clinical trials and post-marketing surveillance. Reports of clinical trials provide detailed information of adverse drug reactions observed under a carefully controlled setting. SIDER (Kuhn et al., 2015)[1] is a database of drugs and their adverse effects, which was created by processing these reports of clinical trials. Post-marketing surveillance, on the other hand, is managed by governments to collect information of adverse drug reactions from the public to monitor for new adverse events that could not be observed during clinical trials. Attendance to the surveillance is completely voluntary; medical practitioners and patients who have taken a drug submit the adverse drug reactions through an online form. Since it is essentially a formal report to governments, descriptions of adverse drug reactions tend to be an objective summary, containing, *e.g.*, just names of symptoms. Furthermore, voluntary-based data collection is hard to scale up. Collected information is made public as a database for further research. In Japan, the Pharmaceuticals and Medical Devices Agency releases such a database.[2]

In both databases, descriptions of adverse drug reactions are formal and objective due to the way their data is collected. These objective descriptions are reliable; however, they lack information about how patients feel and suffer from adverse drug reactions. For example, although medical practitioners know what "glossitis" is, it is not easy to understand how that symptom manifested and how patient suffering manifested. Such subjective descriptions of adverse effects are essential for medical practitioners to understand the mechanisms of adverse effects, as well as to smoothly communicate with their patients.

To collect subjective descriptions about the effects of drugs, we focused on a third source of information—patient weblogs. There is a world-wide trend of sharing fight experiences against diseases on the internet. A typical example is PatientsLikeMe.[3] Patients write about their real experiences in their weblog articles, exchange comments, and share information that they curated, to help each other fight and survive diseases. Because of their nature, patient weblogs provide lively descriptions of their physical conditions. Previous studies confirmed the usefulness of such patient self-reporting notes for finding information on adverse effects (Leaman et al., 2010; Nikfarjam and Gonzalez, 2011; Yang et al., 2012). Nikfarjam et al. (2015) annotated adverse effects on posts mined from a health-related online forum and Twitter,[4] where shortness of content is typical. In contrast with their dataset, we targeted weblogs,

[1] http://sideeffects.embl.de/

[2] https://www.pmda.go.jp/safety/info-services/drugs/adr-info/suspected-adr/0006.html (in Japanese)

[3] https://www.patientslikeme.com/

[4] https://twitter.com/

タグリッソ服用丸２か月が経ちました。今のところの副作用は、相変わらずの舌炎症。はじめは、米や小麦料理がザラザラした舌触りで不味いって感じでした。 (It has been two months since the start of Tagrisso. Its adverse effect is glossitis, as I have always had it during these months. It started as a weird sense that made me feel grains of rice and noodles as sandy.)
Drug name: タグリッソ (Tagrisso) 1. Reaction: 舌炎症 (glossitis) ICD-10 code: K140 Effect-type: Adverse-effect Positive 2. Reaction: 米や小麦料理がザラザラした舌触りで不味い (a weird sense that made me feel grains of rice and noodles as sandy) ICD-10 code: R432 Effect-type: Adverse-effect Positive

Table 1: Running example of drug effect annotation on a patient's weblog article (English translations are in parentheses). We identify drug names and their effects as reported in the article, and we assign labels for the corresponding ICD-10 codes and their effect types.

本当にギリギリまで後発薬 ゲフィチニブ を使わせて下さいました (We appreciate that the doctor allowed to keep using the generic Gefitinib until the very last minute.)
ムコスタが レバミピド 錠に、とか、なんでみんなこの時期に変更しはんねやろ (Why do drugs change their names in this season of the year, like Mucosta changed to Rebamipide tablet.)
主治医はネットでお名前を検索すると、ゲフィチニブ（イレッサ）の論文を書いておられるみたいで、どうやらイレッサの専門家みたいです (I found my doctor's paper on Gefitinib (Iressa) on the internet. He seems to be an expert on Iressa.)

Table 2: Examples of drug name identification (English translations are in parentheses). Underlined phrases are merely modifying the drug names highlighted in yellow, and thus excluded from annotation targets.

which provide richer context and detailed descriptions of adverse effects. Such context is crucial, because a certain drug might cause varying reactions in different sets of patients. We crawled patients' public weblogs from TOBYO,[5] a Japanese initiative of PatientsLikeMe, and reported the effects of drugs described therein. In particular, we annotated spans describing drug reactions, related drug names, corresponding ICD-10 (the 10th edition of International Statistical Classification of Diseases and Related Health Problems created by the World Health Organisation) codes of reactions, and types of effects. To the best of our knowledge, this is the first dataset that provides annotations of drug effects described in detail with rich contexts.

Table 1 shows an example of our annotation results, where a patient wrote about the adverse effect of Tagrisso. It not only provides the name of the symptom (glossitis), but also a subjective and lively description of glossitis as "a weird sense that made me feel grains of rice and noodles as sandy." We have created an annotation dataset of 169 articles that identifies 677 drug reactions. Our dataset is available at our web site[6] for future research on patient-generated text processing.

2. Annotation Scheme

We designed a three-step annotation process, as shown in Fig. 1. First, annotators carefully read through the entire texts. Once they have understood the contents of the texts,

1. they identify the drug names (Sec. 2.1.) and
2. mark the corresponding spans that describe drug reactions due to the identified drugs (Sec. 2.2.).
3. Finally, annotators label the spans to assign the related drug names, the ICD-10 codes for the reaction, and types of effects (Sec. 2.3.).

We documented all the standards of annotations and provided them to the annotators as a guideline. The guideline was immediately updated whenever annotators raised questions about ambiguous cases.

2.1. Drug Name Identification

Table 2 shows examples of drug name identification. To achieve a high agreement for drug name identification, our guideline includes an instruction to exclude modifying expressions for a drug name, such as "generic" and "tablet", as in the first and second examples in Table 2. We regard both trade names and medicinal substances as drug names. Patients sometimes write both of them, putting one in parentheses, such as the third example in Table 2. In such cases, the guideline provides an instruction to annotate both of them as a single span of one drug name. The guideline also instructs annotators to make sure that an identified span is a drug name by searching the web whenever they are unsure.

2.2. Drug Effect Identification

After identification of the drug name, annotators detect spans that describe the drug reactions: patients' subjective observation of physical and/or mental conditions and medical examination results. The granularity of descriptions is diverse, from simply indicating symptom names to describing them quite in detail, as shown in the first and second adverse effects in Table 1, respectively. Annotators identify both types of descriptions as drug reactions.

Since one of the purposes of patients who write weblogs is to share information, their articles often include effects of drugs that patients obtain from the web, such as reports of clinical trials. We exclude such descriptions to anno-

[5] https://www.tobyo.jp/
[6] https://yukiar.github.io/adr-jp/

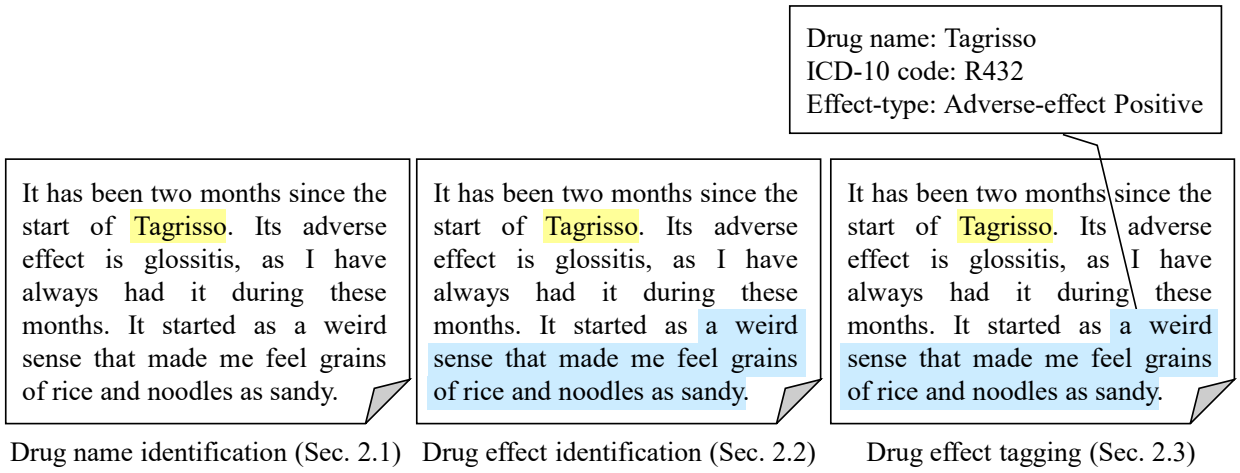

Figure 1: Our annotation process consists of three steps. Annotators identify drug names (Sec. 2.1.) and spans describing reactions to the identified drugs (Sec. 2.2.), and assign labels (Sec. 2.3.).

tate only self-reported drug reactions.[7] Further, we ignore descriptions of their conditions that are irrelevant to drug effects, such as bad conditions due to a cold.

For self-reporting descriptions of drug reactions, our guideline instructs the annotators to identify necessary and sufficient spans reporting effects of drugs. Span annotation is inevitably ambiguous. To improve the agreement of annotations, we set the following criteria to determine a span.

- A span can be a word, phrase, sentence, or multiple sentences.
- For noun phrases such as "eruption" and "muscular pain", expressions to indicate their existence and nonexistence should be part of a span.
- A span should exclude modality expressions and time information.

Note that we annotate spans related not only to actual observation of expected effects, but also to nonoccurrence of effects that have instead been predicted to appear.

2.3. Drug Effect Tagging

Finally, annotators assign labels to identified spans of drug reactions. They label the spans to assign (a) drug names related to the reactions, (b) standardised codes for the reactions, and (c) types of the effects.

For (a), annotators copy and paste the drug names identified at the first step (Sec. 2.1.).

As a standardised code for (b), we use ICD-10, which was developed for systematically recording, analysing, and comparing diseases and death cases across nations and areas. Our guideline suggests finding an appropriate ICD-10 code at MANBYO-SEARCH,[8] which provides a search engine that allows searching an ICD-10 code by querying a drug reaction described in the corresponding span. In case MANBYO-SEARCH returns multiple codes as search results, the annotators are requested to conduct further web search to decide the most plausible code for the span. Our guideline encourages the annotators to assign codes to as many spans as possible. It requests to modify queries whenever MANBYO-SEARCH does not find any appropriate codes. In case a corresponding ICD-10 code does not exist, annotators are allowed to assign an "N/A" label.

For (c), we regarded the effects of drugs reported in patients' weblogs as either target effects or adverse effects. Side effects of drugs do not always cause adverse drug reactions. Certain side effects of drugs are beneficial for different therapeutic purposes, which results in drug repositioning. It is uncommon that medical practitioners explain to patients whether the drugs they prescribe are repositioned. Therefore, we do not distinguish between target effects and repositioned effects, and regard both of them as target effects. Our guideline instructs the annotators to label one category by choosing from the following:

Target-effect Positive Targeted therapeutic effects successfully exhibited.

Target-effect Negative Expected therapeutic effects did not exhibit.

Adverse-effect Positive Adverse effects exhibited as predicted by medical practitioners.

Adverse-effect Negative Predicted adverse effects did not exhibit.

Sometimes patients reported that their conditions were unchanged when taking prescribed drugs. For such cases, annotators are asked to decide to assign either "Target-effect Positive" or "Target-effect Negative" labels from the context. If a drug was prescribed to maintain the patients' current conditions, its effect is considered as positive. On the other hand, if the drug aimed to improve the patients' conditions, the effect is considered as negative. If a span has either "Target-effect Negative" or "Adverse-effect Negative," the guideline instructs to assign the ICD-10 codes of their

[7] Our annotation dataset contains weblog articles written by the nursing family. In this study, we do not distinguish them from patient self-reporting articles, regarding them as equally reliable.

[8] https://www.episodebank.com/manbyo/

positive counterparts, due to nonexistence of codes for reactions or symptoms that did not manifest.

Since symptoms and conditions written in weblogs are patient self-reports, it is not possible to distinguish whether a therapeutic or adverse effect did not exhibit, or whether they exhibited but were not reported. This is a limitation of our dataset; however, the rich context available in weblogs still represents precious information to understand drug effects in real lives.

2.4. Many-to-Many Correspondences between Drugs and Effects

Patients often take multiple drugs: primary drugs for their target effects and supporting drugs for controlling the adverse effects of the primary drugs. Besides, primary and supporting drugs can be multiple. When patients have taken multiple drugs for a similar purpose, it is difficult to spot which drug had a specific effect. Furthermore, adverse effects may occur due to a combination of drugs. Therefore, our guideline instructs to label a span describing effects with all possible drugs, so that correspondences between drugs and effects are many-to-many.

3. Annotation Settings

3.1. Patients' Weblog Collection and Filtering

We crawled weblogs from TOBYO, which are open to the public. TOBYO is a platform for information and experience sharing for patients, and represents the Japanese version of PatientsLikeMe. As initial dataset, we targeted lung cancer because of its impact on Japanese society. Among cancers, which are the leading cause of death in Japan, lung cancer is one of the most common and has the highest mortality rate. Weblogs shared at TOBYO are assigned tags that represent diseases. We crawled all the weblogs tagged as lung cancer, which resulted in 472 weblogs with 117k articles.[9]

In contrast with clinical records written by medical practitioners, patient weblogs contain a wide variety of topics; not only records of progress of their treatment and physical conditions, but also memories of their daily lives. Moreover, a certain percentage of weblogs aim at sharing curated information, *e.g.*, citations from medical papers and trial reports, as well as copies of drug package inserts. Hence, we first selected weblog articles that were likely to describe drugs that patients had really taken and their reactions.

We selected articles that satisfied the following conditions. Articles should:

- contain one to five drug names in our dictionary,
- be at least 140 characters long, and
- do not have an embedded URL.

We used a dictionary of Japanese drug names, which provides 27k entries. Our collaborator created the dictionary by processing drug package inserts to collect variants of drug names.[10] Table 3 shows statistics of the selected articles. Since these weblog articles are written in a colloquial style with various medical terms, tokenisers (which are generally trained with news articles in a general domain) perform poorly. Hence, we used a character as a unit for statistics. Among the filtered $4,147$ articles, we annotated 500 samples picked randomly.

Number of articles	4,147
Average number of lines in an article	27
Average number of characters in an article	693

Table 3: Statistics of crawled articles after filtering

To reduce annotator burden of reading texts that were unlikely to describe drug reactions, we extracted a sentence containing at least one drug name and the following 9 sentences as the annotation target, assuming that patients first list drugs they take and then discuss their effects. Compared to the dataset annotated by Nikfarjam et al. (2015), which has only 1.5 sentences for one entry, our dataset provides much richer context.

3.2. Annotator Profiles

Because the annotation target is weblogs written by patients, we considered expertise in the medical and pharmaceutical science as not necessary for annotation. We recruited three professionals who had rich experiences in annotation tasks within various domains. One of the annotators had a half-year experience in annotating medical documents.

3.3. Annotation Procedure

Three annotators independently annotated the same 500 articles, *i.e.*, the cumulative total number of annotated articles was $1,500$. We used Microsoft Word as annotation tool, which was familiar to annotators. The annotators were encouraged to raise questions when they encountered ambiguous cases where it was difficult to make a decision. One of the authors was responsible for handling these questions. The guideline was immediately updated whenever necessary then shared with annotators to maintain a consistent standard.

4. Analysis of Annotation Results

Over the cumulative total of $1,500$ annotated articles, the number of identified drug names was 108, and the number of identified ICD-10 codes was 104. In this section, we discuss the quality and consolidate the annotation results.

4.1. Agreement Rates of Annotations

To examine the agreement level of the annotations, we formatted the annotation results with the character-level Inside–Outside–Beginning (IOB) tagging scheme (Ramshaw and Marcus, 1995) and calculated Fleiss' kappa. The IOB tagging scheme is common for evaluating named entity recognition, which is a sequential tagging task of named

[9] Crawling was conducted in Sept. 2019.

[10] The dictionary of Japanese drug names will be released at our web site.

	Fleiss' kappa	# of unique tags
IOB (span only)	0.635	3
IOB+drug name	0.615	287
IOB+ICD-10	0.565	211
IOB+effect type	0.589	59
IOB+all labels	0.520	1,020

Table 4: Agreement rates of annotations.

Number of articles	169
Number of annotated drug reactions	677
Number of unique drugs	87
Number of unique ICD-10 codes	78
Average number of sentences in an article	8.1
Average number of characters in an article	328.5
Average number of drug reactions in an article	2.5
Average number of drugs related to a drug reaction	1.6
Average number of characters in the description of a drug effect	10.0

Table 5: Statistics of consolidated annotations.

Target-effect Positive	118
Target-effect Negative	83
Adverse-effect Positive	396
Adverse-effect Negative codes	80

Table 6: Frequency of types of effects

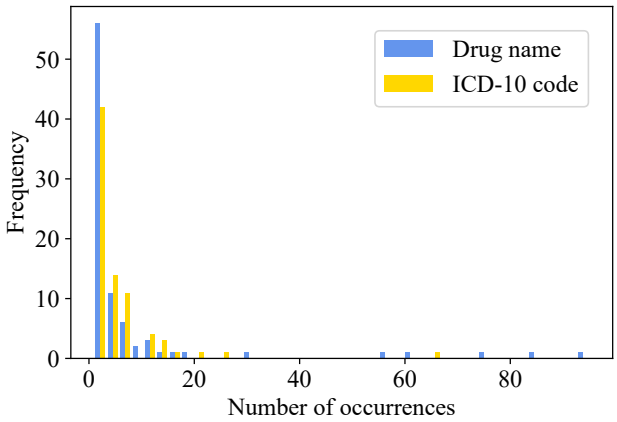

Figure 2: Histograms for drug names and ICD-10 codes.

entities, similar to our span identification of drug reactions. Specifically, every character in a sentence was tagged with "B" if it was the beginning of a span, "I" if it was inside the span, and "O" if it was outside of the span. Since we have three kinds of labels for each span of drug effect, *i.e.*, a drug name, the ICD-10 code, and the effect type, we combined these labels with IOB tags.

A span may have multiple sets of labels when the corresponding drug reaction relates to multiple drugs. We regarded these sets of labels assigned to the span as a single annotation, *i.e.*, concatenated each type of label as a single label. In other words, we examined the agreement rate of an exact match of all possible annotation labels. The portion of annotated spans in an article was much smaller than that of unannotated spans. To avoid that unannotated spans had a dominant effect on the agreement rates, we excluded sentences that had no annotated span.

Table 4 shows the Fleiss' kappa values and the number of unique tags for each combination of IOB tags and annotation labels. For the simplest case that identified only spans describing drug effects (the first row in Table 4), the Fleiss' kappa value reached 0.635. When we combined IOB tags and labels, the kappa values were still as high as 0.615 for drug names, 0.565 for ICD-10 codes, and 0.589 for effect types. Even when we combined all the label types that resulted in 1,020 unique tags, the kappa value was 0.520. These results show that the annotators produced reliable annotations with high agreements.

4.2. Annotation Consolidation

Sec. 4.1. showed that the annotators produced annotations with high agreements. We further improved the quality of annotation by consolidating our results.

First, for each article, we identified sets of exactly matching labels where at least two annotators agreed on the whole set of drug name, ICD-10 code, and effect type. We discarded sets of labels produced by only one annotator. For these agreed sets of labels, we took the longest span attached to the exactly matching sets in order to complement the ambiguity in span identification. As a result, we obtained 677 labelled spans of drug reactions on 169 weblog articles as summarised in Table 5. On average, an article had 8 sentences with 329 characters, which provide 2.5 descriptions of drug reactions with a length of 10 characters . The average number of related drugs for a drug reaction is 1.6.

Table 6 shows frequencies of the types of effects. Patients report nearly four times more adverse effects than therapeutic effects. On the contrary, they report the nonoccurrence of expected effects with similar frequency for therapeutic and adverse effects. Fig. 2 shows histograms for drug names and ICD-10 codes. As expected, the frequencies of both drug names and ICD-10 codes are highly skewed.

The consolidation process reduced the number of weblog articles from 500 to 169. Most of the unannotated articles were curated information of drugs or articles irrelevant to drug reactions, such as records of everyday affairs.

Table 7 shows some examples of the consolidated annotations. The first example describes the adverse effect of cilostazol, which aimed to control the adverse effect of the primary drug Tagrisso. Although this example looks simple, it requires determination of which drugs (Tagrisso and cilostazol) caused the adverse effect described in a free-form sentence by considering the context. The second example has many-to-many correspondences of drugs (irinotecan and carboplatin) and their adverse effects consisting of increased urine output and redness on the cheeks. In addition, it requires distinguishing that the patient did not have reactions of diarrhea and dizziness. The third example is the most complex case. It describes that two drugs, docetaxel and ramucirumab, had adverse effects consisting of muscular pain and oral discomfort, while their target effect

タグリッソが耐性となり残念な気持ちはあるものの、副作用対策で服用しているシロスタゾールが終わるのは嬉しいんです。...少し動くだけでドキドキしていて、今は走るどころか早歩きも無理です。それもあと少しの辛抱だ。 (While I'm disappointed to have gained resistance to Tagrisso, glad to stop cilostazol prescribed to inhibit the adverse effects of Tagrisso. ... My heart gets pounding by just light physical activities, which prohibits me from walking fast, no way to run. However, it's almost over.)

Drug name: シロスタゾール (cilostazol)

1. Reaction: 少し動くだけでドキドキしていて、今は走るどころか早歩きも無理 (My heart gets pounding by just light physical activities, which prohibits me from walking fast, no way to run)
   ICD-10 code: R000
   Effect-type: Adverse-effect Positive

イリノテカン、カルボプラチンの副作用は水分コントロールが崩れるそうで、下痢になりやすいそうですが、私の場合尿量倍増ですね。 ... あと、なぜか今回変わりダネ副作用、のぼせてる感覚ないんですけどほっぺがまっかっか。この半年で400個くらいのリンゴ消費してるからりんごの呪いかなあ。 (I have heard that irinotecan and carboplatin have an adverse effect that disturbs body water control, which causes diarrhoea. In my case, these drugs caused an increase in urine output. ... I'm experiencing a rare adverse effect of apple cheeks although I don't feel dizzy. Is it a curse of the 400 apples that I consumed in the past half a year!?)

Drug name: イリノテカン (irinotecan), カルボプラチン (carboplatin)

1. Reaction: 尿量倍増 (increase in urine output)
   ICD-10 code: N/A
   Effect-type: Adverse-effect Positive
2. Reaction: ほっぺがまっかっか (apple cheeks)
   ICD-10 code: R232
   Effect-type: Adverse-effect Positive

今のドセタキセル+ラムシルマブの副作用は、軽い筋肉痛と、刺激物を食べると、口の中がヒリヒリするぐらいです。... ジーラスタの副作用の腰痛は軽度。頭痛は相変わらずで、起きて数時間が調子が悪いです。お昼前になると、調子が良くなり、簡単に家事をやって、好きなことをしてます。... 副作用は悪化することなく、脳転移の症状も出るこなく過ごせて良かった。 (As for the current adverse effects of docetaxel+ramucirumab, I only have light muscular pain and oral discomfort that makes the inside of my mouth feel like burning when I eat spicy foods. ... A light lower back pain due to the adverse effect of G-Lasta. I still have a headache, which makes me feel unwell for a few hours after getting up. I usually feel better around noon, so do easy housework and spend some time for my hobby. ... I'm happy that these adverse effects are under control and having no symptoms due to brain metastasis so far.)

Drug name: ドセタキセル (docetaxel), ラムシルマブ (ramucirumab)

1. Reaction: 軽い筋肉痛 (light muscular pain)
   ICD-10 code: M7919
   Effect-type: Adverse-effect Positive
2. Reaction: 刺激物を食べると、口の中がヒリヒリする (oral discomfort that makes the inside of my mouth feel like burning when I eat spicy foods)
   ICD-10 code: N/A
   Effect-type: Adverse-effect Positive
3. Reaction: 脳転移の症状も出るこなく (no symptoms due to brain metastasis)
   ICD-10 code: C793
   Effect-type: Target-effect Positive

Drug name: ジーラスタ (G-Lasta)

1. Reaction: 腰痛 (lower back pain)
   ICD-10 code: M5456
   Effect-type: Adverse-effect Positive

Drug name: ドセタキセル (docetaxel), ラムシルマブ (ramucirumab), ジーラスタ (G-Lasta)

1. Reaction: 頭痛 (headache)
   ICD-10 code: R51
   Effect-type: Adverse-effect Positive

Table 7: Examples of consolidated annotations (English translations are in parentheses).

of preventing symptoms of brain metastasis had exhibited successfully. The patient also took a third drug, G-Lasta, which caused lower back pain. It is hard to identify which of these three drugs led to the adverse effect of headache. Hence, the annotators regarded all of them as related drugs.

5. Related Work

5.1. Annotation Datasets for Adverse Effect Detection

Annotated datasets are the basis for powerful statistical and machine learning approaches, which are promising for adverse effect detection. There are a few annotated datasets created by previous studies. Thompson et al. (2018) annotated drug effects and their interactions as described in abstracts of scientific papers. For annotations of patient-generated texts, the Social Media Mining for Health Applications Shared Task[11] provides binary labelling of tweets to indicate the existence of adverse drug reactions and labelling of spans of the reactions as standard codes. Our dataset annotates more abundant information, *i.e.*, related drug names and effect types.

The dataset created by Nikfarjam et al. (2015) is most relevant to our study. This dataset identifies spans describing drug reactions on posts submitted to a health-related online forum as well as Twitter. It also assigns labels for related drug names, types of effects, and corresponding concept IDs defined in the Unified Medical Language System. The significant difference from our dataset is the description style for the drug reactions, which is caused by the difference in the data sources. Since Nikfarjam et al. (2015) aimed to examine whether adverse effects were collectable from the web, they targeted online forums and Twitter, which provide abundant short posts. On the other hand, we aim to collect rich descriptions of each adverse effect from weblog articles.

5.2. Challenges in Adverse Effect Mining

A challenge in adverse effect mining is that terminologies are different from those in general domains. To identify spans describing adverse drug reactions, previous studies used lexicons in the medical domain (Leaman et al., 2010; Yang et al., 2012), lexical patterns identified by association rule mining (Nikfarjam and Gonzalez, 2011), and word embedding (Nikfarjam et al., 2015). Differences in style of the texts is another issue. People write posts to forums and Twitter in a colloquial style, which is significantly different from the formal style used to define medical concepts. Previous studies used pattern mining (Stilo et al., 2013) and applied deep neural networks (Limsopatham and Collier, 2016) to map colloquial expressions in posts onto formal writing used in medical concepts. Our dataset poses both challenges, as it annotates weblogs written in colloquial style, where adverse drug reactions are described not only as phrases, but also as sentences or even longer spans.

Recent studies show that a pre-trained model for text representation achieves impressive performances on various downstream tasks, such as automatic question answering and natural language inference (Devlin et al., 2019). Alsentzer et al. (2019) adapted the pre-trained model to the medical domain. Further, sentiments in texts are useful clues to identify adverse effects (Sarker and Gonzalez, 2015; Wu et al., 2018; Alhuzali and Ananiadou, 2019), which is reasonable because adverse effects are firmly negative events for patients. These approaches are promising for adverse effect mining on patient-generated texts, from online forums to weblogs including our dataset.

6. Potential Task Designs with Our Dataset

Our annotation dataset is a valuable resource to advance research on the automatic detection of drug effects. More broadly, it is useful for research on knowledge extraction from patient-generated texts. Various types of tasks can be designed using our dataset. To name a few:

1. Automatic linking of texts describing drug reactions to concept IDs in a medical ontology
2. Prediction of effect type given an article and drug names discussed in the article
3. Span identification describing drug effects given an article
4. Span identification and labelling of drug names, corresponding ICD-10 codes, and effect types

The first task automatically identifies spans describing drug effects and maps onto standard codes or IDs defined in a medical ontology (Mullenbach et al., 2018; Limsopatham and Collier, 2016). The second task is a variant of the aspect-based sentiment analysis in the domain of drug reaction detection. The third task is a kind of sequential labelling problems, but with a number of labels as large as related drug names. The fourth task is an advanced sequential labelling problem with multiple types of labels with dependent relations. This task conforms to a more practical setting on drug effect mining.

As discussed in Sec. 4.2., our annotation dataset consists of challenging examples that (a) have many-to-many correspondences between drugs and drug reactions, (b) require to distinguish drugs related to a specific reaction from other drugs prescribed together, and (c) require to label not only phrases but also sentences or even longer spans. Therefore, our dataset is also useful as an advanced dataset for existing aspect-based sentiment analysis and sequential labelling problems.

7. Conclusion and Future Work

We annotated patient weblog articles crawled from the patient-networking platform with the aim of collecting the effects of drugs with rich and detailed descriptions. As an initial trial, we targeted weblogs of Japanese lung cancer patients, because of its significant impact on Japanese society. Our dataset identifies 677 drug reactions on 169 weblog articles labelled as related drug names, ICD-10 codes, and types of effects.

We are expanding the dataset to annotate $2,000$ more weblog articles. To give more contexts of drug reactions, we are annotating the entire weblog articles for the next version. We will also label the standardised medical terminology defined in the Medical Dictionary for Regulatory Activities (MedDRA) for drug reactions,[12] which is the international medical terminology developed under the auspices

[11] https://healthlanguageprocessing.org/smm4h-sharedtask-2020/

[12] https://www.meddra.org/

of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Our dataset will be publicly available for future research. Furthermore, we are developing methods to detect drug reactions from patient-generated texts automatically. In the future, we will annotate more context information on drug taking, including prescribed amounts, dosage, time of taking drugs, and previous drug records.

Acknowledgements

We thank Kazuki Ashihara for his contribution to annotation as well as valuable discussions with us. This work was supported by JST AIP-PRISM Grant Number JPMJCR18Y1, Japan.

Bibliographical References

Alhuzali, H. and Ananiadou, S. (2019). Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP) & Shared Task*, pages 339–347, August.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the Clinical Natural Language Processing Workshop*, pages 72–78, June.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, June.

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2015). The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, October.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., and Gonzalez, G. (2010). Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP) & Shared Task*, pages 117–125, July.

Limsopatham, N. and Collier, N. (2016). Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1014–1023, August.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1101–1111, June.

Nikfarjam, A. and Gonzalez, G. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 1019–1026, October.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, March.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Proceedings of the Workshop on Very Large Corpora*, pages 82–94.

Sarker, A. and Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207, February.

Stilo, G., De Vincenzi, M., Tozzi, A. E., and Velardi, P. (2013). Automated learning of everyday patients' language for medical blogs analytics. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 640–648, September.

Thompson, P., Daikou, S., Ueno, K., Batista-Navarro, R., Tsujii, J., and Ananiadou, S. (2018). Annotation and detection of drug effects in text for pharmacovigilance. *Journal of Cheminformatics*, 10(1), August.

Wu, C., Wu, F., Liu, J., Wu, S., Huang, Y., and Xie, X. (2018). Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In *Proceedings of the Social Media Mining for Health Applications Workshop (SMM4H) & Shared Task*, pages 34–37, October.

Yang, C. C., Yang, H., Jiang, L., and Zhang, M. (2012). Social media mining for drug safety signal detection. In *Proceedings of the International Workshop on Smart Health and Wellbeing (SHB)*, pages 33–40, October.