Journal of Biomedical Informatics 55 (2015) 73–81

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

CADEC: A corpus of adverse drug event annotations

Sarvnaz Karimi *, Alejandro Metke-Jimenez, Madonna Kemp, Chen Wang

CSIRO, Australia

ARTICLE INFO

Article history:
Received 24 October 2014
Revised 9 February 2015
Accepted 20 March 2015
Available online 27 March 2015

Keywords:
Adverse drug reaction
Medical forum
SNOMED CT
MedDRA
Annotated corpus
Drug safety
Social media
Information extraction
Consumer reviews

ABSTRACT

CSIRO Adverse Drug Event Corpus (CADEC) is a new rich annotated corpus of medical forum posts on patient-reported Adverse Drug Events (ADEs). The corpus is sourced from posts on social media, and contains text that is largely written in colloquial language and often deviates from formal English grammar and punctuation rules. Annotations contain mentions of concepts such as drugs, adverse effects, symptoms, and diseases linked to their corresponding concepts in controlled vocabularies, i.e., SNOMED Clinical Terms and MedDRA. The quality of the annotations is ensured by annotation guidelines, multi-stage annotations, measuring inter-annotator agreement, and final review of the annotations by a clinical terminologist. This corpus is useful for studies in the area of information extraction, or more generally text mining, from social media to detect possible adverse drug reactions from direct patient reports. The corpus is publicly available at https://data.csiro.au.[1]

1. Introduction

Due to limitations in clinical trials, not all the potential side effects of medications are discovered prior to the drug going to market [3]. Adverse drug reactions that remain unknown create major concerns in public health [7]. They are responsible for thousands of incidents of death or serious injury, as well as millions of hospitalisations. Consequently, they cost billions of dollars to the healthcare systems around the world [1,14,27].

Postmarket surveillance, also known as pharmacovigilance, plays an important role in identifying those adverse drug side effects that are left undetected while a drug is in the market [3,6]. The traditional surveillance practice, enforced by regulatory agencies such as the Food and Drug Administration (FDA) in the US and the Therapeutic Goods Administration (TGA) in Australia, is to collect volunteer reports of adverse drug side effects, investigate the reports, and issue safety signals if a drug is suspected to cause an adverse effect.

More recently, active surveillance has been studied where different data sources are automatically monitored for reports of possible adverse reactions; FDA's Sentinel Initiative is an example of active monitoring. One of the information sources that could be actively monitored is medical forums where consumers discuss their first-hand experience with medications.

A human annotated corpus is a valuable resource in the development and evaluation of text-mining methods reliant on machine learning. Such annotations need to account for medications and adverse effects, as well as patient condition and demographic data. Such a corpus can be expensive to construct but once created will serve multiple studies in the advancement of their text-mining algorithms. Therefore, we developed a corpus of medical forum posts taken from AskaPatient[2] which collects ratings and reviews of medications from their consumers. These posts were annotated for entities such as the names of the drugs consumed, and their adverse effects. Additionally, these annotated entities were linked to controlled vocabularies: the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), AMT (The Australian Medicines Terminology) and MedDRA (The Medical Dictionary for Regulatory Activities).

To our knowledge, our corpus is the first richly annotated and publicly available corpus of *medical forum posts* that can be applied to text mining tasks related to pharmacovigilance.

2. Related work

We review the existing relevant corpora and their specifications, as well as a brief overview of methodologies proposed in the literature to extract drug adverse effect information from social media.

* Corresponding author.
E-mail address: sarvnaz.karimi@csiro.au (S. Karimi).

[1] The data can be used for research purposes only, under the CSIRO data licence.

[2] http://www.askapatient.com/.

http://dx.doi.org/10.1016/j.jbi.2015.03.010
1532-0464/

2.1. Relevant corpora

Literature reports on the corpora that are annotated to facilitate information extraction from biomedical literature, electronic health records, and other textual data. Below, we introduce some of these corpora together with their specifications such as annotated entities and their data sources, as well as stating the differences to our corpus. We only review those datasets that share either an entity of interest with our corpus, or are closely related to text mining in the pharmacovigilance area.

Leaman et al. [17] developed the Arizona Disease Corpus (AZDC) where biomedical literature (MEDLINE abstracts) were annotated for mentions of diseases. The disease mentions were then normalised to their corresponding UMLS concepts. They also demonstrated that disease mentions have similar characteristics to entities such as genes and proteins. That is, they have name variants, one name can refer to entities of different semantic types, and are prone to complex syntactic structures. These characteristics can lead to ambiguity in automatically recognising disease mentions in text. In 2014, an extension of AZDC called the NCBI disease corpus [9] was introduced which annotates disease mentions in PubMed abstracts using a web-based tool called PubTator [31]. These mentions were then linked to Medical Subject Headings (MeSH) or Online Mendelian Inheritance in Man (OMIM) standard terminologies. The annotations were done by 14 annotators each annotating four to five batches of 30 abstracts with each batch receiving at least two annotations. Annotators were given detailed guidelines.

Roberts et al. [26] annotated deceased cancer patient records for a large number of entities such as intervention, drug or device, and condition (full list in Table 1, second row). Their data, which was used in a shared task, was created based on carefully designed guidelines finalised after an initial set of annotations. They hired 25 annotators, and had each document annotated twice in order to calculate inter-annotator agreement. Their definition of entities was generally inclusive and inspired from the UMLS concepts. For example, they annotated drugs and medical devices under the same entity. Their *condition* entity type also covered a range of concepts such as symptoms, complications, and injuries.

Gurulingappa et al. [10] described a corpus of annotated MEDLINE abstracts for human diseases and adverse effects. The abstracts were found by querying PubMed for "disease OR adverse effect" and then a subset of 400 abstracts was randomly chosen to be annotated. Two annotators were instructed to identify mentions of disease and adverse effects based on their context. The final corpus contains 813 mentions of adverse effects and 1428 mentions of disease. The corpus is publicly available with the annotations in the IOB (Inside, Outside, Beginning) format.

Gurulingappa et al. [11] created a corpus for extracting information related to drug safety from medical case reports in MEDLINE. The MEDLINE abstracts were chosen randomly from the pool of abstracts returned by querying PubMed with the MeSH (Medical Subject Headings) terms "drug therapy" and "adverse effect". The corpus was annotated by three annotators for the mentions of drugs (brand names, trivial names, abbreviations), adverse effects, dosage (quantity and frequency), as well as the relationships among these entities. Adverse effects of a certain drug covered a range of signs, symptoms, diseases, disorders, abnormalities, organ damage and even death caused by that drug.

In their annotations, Gurulingappa et al. [11] excluded names of medical devices or hospital chemicals. Also, they only annotated drug name mentions in relation to an adverse event. The final corpus included those sentences from the abstracts that had at least one mention of an adverse effect.

Van Mulligen et al. [30] report a publicly available annotated corpus of biomedical literature where instances of drugs, disorders, genes, and the relationships among the identified entities are annotated. Data for the corpus was collected by querying PubMed. Search strategies are listed in their paper. They initially annotated the corpus using a Named Entity Recogniser (NER) and then asked their annotators to correct the automatic annotations.

Deleger et al. [8] created an annotated corpus of clinical records. The corpus was created with two different purposes: (1) a de-identification task for which the data was annotated for personal health information such as patient's age or email address; and (2) annotation of entities related to the medication, such as medication name and type, as well as disease and symptoms. Annotations in this step were based on SNOMED CT and UMLS concepts. Annotators were instructed to annotate entities if there existed a corresponding SNOMED CT or UMLS concept. Deleger et al. [8] annotated a corpus of clinical notes and the FDA drug labels using two annotators. A guideline was developed and updated after a trial annotation step. They calculated inter-annotator agreement to ensure only documents with full agreement were included in the final corpus.

One line of work in the area of pharmacovigilance is studying the drug-drug interactions (DDIs) that lead to adverse drug reactions. Herrero-Zazo et al. [13] created a corpus that supports text-mining in this area. Data for the corpus was sourced from DrugBank [32] and MEDLINE abstracts. Annotations followed carefully written guidelines by two pharmacists. The data was initially annotated by MetaMap to identify mentions of biomedical entities and then passed to the annotators for further curation and correction. Inter-annotator agreements are reported separately for the entities and relationships. This corpus has been used in the SemEval[3] 2013 shared task. Our corpus is different as it does not look at the drug-drug interactions and instead focuses on reported adverse effects of a single drug.

We summarise the existing corpora in Table 1. The second column specifies the origin of the dataset. The type of data and the corpus size are listed in the third and forth columns. The annotated entities and relationships among these entities are listed in the final column of the table. Note that often the same entity type, for example drug, differs in its definitions across different corpora.

All the existing corpora are created on the basis of biomedical literature, e.g., from MEDLINE abstracts. The CADEC corpus, which we introduce further in the following sections, is the only corpus that is sourced from consumer reports on social media, introducing its unique linguistic characteristics and processing challenges.

2.2. Drug adverse effect mining from social media

Mining signals of adverse drug reactions from social media has been studied since 2010. Leaman et al. [18] mined patients' comments on a medical forum called DailyStrength[4] to find mentions of adverse drug events. Their data was annotated for adverse effect, beneficial effect, indication, and other. They used a lexicon that combines COSTART[5] and a few other resources to extract adverse effect information from patient comments using a sliding window approach.

Chee et al. [4] applied classifiers to identify drugs that have potential for becoming part of the watchlist of the FDA. They used patients posts on Health and Wellness Yahoo! Groups.

Benton et al. [2] extracted potential adverse effects from a number of different breast cancer forums, such as breastcancer.org, using frequency counts of terms in a controlled vocabulary. They

[3] SemEval (Semantic Evaluation) is a shared task for evaluations of semantic analysis systems on a shared dataset and agreed evaluation metrics.

[4] www.dailystrength.org.

[5] Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) was developed by the FDA for coding of adverse drug reaction in post-market reports. It is now replaced by MedDRA.

Table 1
Specifications of the existing relevant corpora and CADEC.

Corpus	Origin	Type	Size	Entities/relationships
Leaman et al. [17]	MEDLINE	Literature	749 abstracts (2784 sentences)	Disease
Roberts et al. [26]	Royal Marsden Hospital	Cancer patient records	150 documents (50 clinical narratives, 50 histopathology reports, 50 imaging reports)	Condition, intervention, investigation, result, drug or device, locus, negation signal, laterality signal, sub-location signal, and relations
Gurulingappa et al. [10]	MEDLINE	Literature	400 abstracts	Disease, adverse effect
Gurulingappa et al. [11]	MEDLINE	Medical case reports	2972 abstracts (4272 sentences)	Drug, adverse effect, dosage, relationships among these entities
Deleger et al. [8]	Cincinnati children's hospital, ClinicalTrials.gov, DailyMed	FDA drug labels, clinical trial announcements, clinical notes	3503 clinical notes for personal health information, 1655 for disease and disorder	Medication name, medication type, date, dosage, duration, form, frequency, route, status change, strength, modifier, disease/disorder, sign/symptom, and personal health information (age, date, email, etc.)
Van Mulligen et al. [30]	MEDLINE	Literature	300 abstracts	Drug, disorder, gene, relationships among these entities
Herrero-Zazo et al. [13]	MEDLINE and DrugBank database	Literature and drug information	233 MEDLINE abstracts and 792 text from DrugBank	Pharmacological substance (drug generic name, brand, drug group, and active substances not approved for human use), four types of DDI relationships
Dogan et al. [9]	PubMed abstracts	Literature	793 abstracts (6881 sentences)	Disease mentions
Our corpus (CADEC)	AskaPatient	Medical forum	1253 posts (7398 sentences)	Drug, adverse effect, disease, symptom, finding

then used association rule mining to establish the relationship between the matching terms. Association rule mining is a data-mining approach popular for mining adverse effects from regulatory and administrative databases. Yang et al. [33] studied signal detection from a medical forum called MedHelp[6] by extending the existing association rule mining algorithms by adding *interestingness* and *impressiveness* metrics. To find mentions of adverse drug effects in the text, they used a sliding window and a consumer controlled vocabulary to match the terms.

Liu and Chen [21] implemented a system called AZDrugMiner. Data was collected using a crawler from Diabetes Online Community.[7] To find mentions of adverse effects and related information or relationships such as drug-adverse effect, they first used MetaMap which maps text to UMLS concepts, and then extracted relations using co-occurrence analysis.

A full review of these techniques can be found in [16]. All of these studies are evaluated on different privately held corpora, which makes the comparison of their effectiveness difficult. The provision of a public corpus such as CADEC will facilitate the comparison and evaluation of these mining techniques.

3. Corpus material

The data for the CADEC corpus was sourced from a medical forum called AskaPatient, which is dedicated to consumer reviews on medications. Patients can rate the medication by filling a detailed form on a specific drug based on their brand name, e.g., Tamiflu. This form requests satisfaction rate, reason for taking the drug, dosage and frequency, duration of taking the drug, side effects experienced in free-text form, comments in free text form, as well patient demographics including age and gender. Not all the information in the rating form is mandatory. Also, there is a one review per drug policy in the website where consumers are asked not to enter multiple reviews unless they want to update their previous post.

AskaPatient provided us with consumer posts on the following 12 drugs: Voltaren (Diclofenac Sodium), Cataflam (Diclofenac Potassium), Voltaren-XR (Diclofenac Sodium), Arthrotec (Diclofenac Sodium; Misoprostol), Pennsaid (Diclofenac Sodium), Solaraze (Diclofenac Sodium), Flector (Diclofenac Epolamine), Cambia (Diclofenac Potassium), Zipsor (Diclofenac Potassium), Diclofenac Sodium, Diclofenac Potassium, and Lipitor (Atorvastatin Calcium). We divided these medications into two categories: Diclofenac, which includes those medications with Diclofenac in their active ingredient, and Lipitor. As per the rating form explained above, these posts contain patient demographics, a satisfaction rating on the medication from 1 (low) to 5 (high), reason for taking the medication, how it was administered, patient comments on the effectiveness of the drug and if any side effects were experienced. A sample post for Voltaren is shown in Table 2. In CADEC, we only annotate and provide the free text sections of each post. The language of all these posts is English. Most of the posts are written in colloquial language and do not follow formal English grammar and punctuation rules. They largely report the patients personal experience, however, sometimes conditions of a family member were reported.

Statistics on the posts that are used to create the corpus are shown in Table 3. Statistics are listed for the entire corpus, as well as each drug category (Diclofenac and Lipitor) separately. Number of posts for the original data was 1321, however, 71 posts did not contain any text. These posts were excluded from the final corpus. The length of the posts and their average size in sentences and words are also reported in Table 3, rows four to seven. Lipitor has a substantially higher number of posts in a similar time span than Diclofenac with each post being longer on average. Gender of the reporting consumers are almost equally divided between men and women, with 42 posts missing gender information. Age range of the patients was from 17 to 84 years old, with average age being 52.

4. Annotation

We annotated the corpus in two main stages: (1) *entity identification*, and (2) *terminology association*, also known as *normalisation*, to link the identified entities to controlled vocabularies.

Below we explain the annotation guidelines, the tools used during the annotation, and the annotation process.

4.1. Guidelines

Annotation guidelines are a method of controlling the quality of the corpus construction, with a number of suggestions provided in

[6] www.medhelp.org/.
[7] http://community.diabetes.org.

Table 2
A sample post on Voltaren in AskaPatient.com.

Rating	Reason	Side effects	Comments	Sex	Age	Duration/dosage	Date added
4	Osteoarthritis of the hip	It helps relieve chronic pain but over time, causes intestinal pain and bleeding. I had symptoms similar to diverticulitis: blood in stool, pain	I would be cautious about paying attention to cramping, intestinal/stomach pain which can lead to very serious conditions	M	63	1.5 years 100MG ER 1X D	4/12/2013

Table 3
Statistics on the data used in CADEC.

	Corpus	Diclofenac	Lipitor
No. posts	1321	264	1057
No. posts with text	1250	250	1000
No. sentences	7632	1263	6369
Avg. post length (sentence)	6	5	6
No. words	101,486	16,778	84,708
Avg. post length (word)	81	67	85
Time span	January 2001–September 2013	February 2002–August 2013	January 2001–September 2013
Gender	F 662 (50.1%) M 617 (49.9%)	F 181 (68.6%) M 76 (28.8%)	F 481 (45.6%) M 541 (51.2%)
Age range	17–84	17–78	19–84
Avg. age	52	47	54

the literature on how to devise guidelines for annotators [5,12,19]. We adapted some of the suggestions by Zazo et al. [12] in creating the guidelines. Annotations were done at the sentence level. No entity that spanned over sentences was annotated. Annotations could be discontinuous within the same sentence. Duplicate entities within one sentence were annotated independently, that is, all the occurrences of the same entity were annotated. Generic mentions of an entity were not annotated, for example, the term *side effect*. Embedded entities were not separately annotated. That is, if part of an entity contained another entity, only the main entity was annotated. For example, if the post mentions *muscle pain* as an adverse side effect, *pain* is not separately annotated as another entity. Co-referential/anaphoric references were not annotated. Leading prepositions, qualifiers, or possessive adjectives were excluded to promote more consistent spans. For example, *arthritis* was annotated instead of *my arthritis*.

Specifically, for the first stage of the annotations we defined the entities of interest as below. These entities and their definitions were modifications of entities proposed by Karimi et al. [15].

Drug Mentions of the name of a medicine or drug are annotated with *drug*. Drug classes, such as Nonsteroidal anti-inflammatory drugs (NSAIDS), were excluded. Medical devices were also excluded. For example, in the sentence "I must be addicted to Diclofenac", "Diclofenac" is annotated with *Drug*.

ADR Mentions of adverse drug reactions that, according to the text, are clearly associated with a drug are annotated with an *ADR* label. For example, in the sentence "Sometimes causes drowsiness.", "drowsiness" is an adverse effect. All the necessary context for an ADR concept is annotated. For example, if the sentence said "I experience acute stomach pain...", then "acute stomach pain" is annotated not just "stomach pain", whereas in a sentence such as "I felt blank like a blank piece of paper", only "felt blank" is annotated.

Disease This entity specifies the reason for taking the drug. Patients may mention the name of a disease for which they take the medicine. If it is a specific disease name, it is tagged with *Disease*. For example, in the sentence "...after 3 years of having Ativan keep the anxiety & aggression in check...", both "anxiety" and "aggression" are annotated (separately) as *Disease*.

Symptom This entity specifies a reason for taking the drug. Patients may mention symptoms of a disease that led to them taking a drug. For example, in the sentence, "My heart was racing and ...", a symptom "heart racing" is highlighted.

Finding A clinical finding is any adverse side effect, disease or symptom that was not directly experienced by the reporting patient, or any other clinical concept that could fall in any of these categories but the annotator is not clear as to which one it belongs.

These definitions were finalised after receiving feedback from the annotators and experts in the field, including a pharmacist. A pilot annotation task was set up to annotate a small number of posts and the annotation setting was modified based on the feedback.

For the second stage of the annotations, terminology association, the following guidelines were used. Details are explained in Section 4.3.

SNOMED CT (SCT) Any span of text annotated with any tag other than *Drug* should be mapped to the corresponding SNOMED CT concept from the *Clinical Finding* hierarchy. If no concept exists then assign the tag *concept_less*.

AMT Any span of text annotated with *Drug* should be mapped to the corresponding AMT concept. If no matching concept exists then assign the tag *concept_less*.

MedDRA Any span of text annotated with *ADR* should be mapped to the corresponding MedDRA term.

4.2. Entity identification

The first stage of the annotation process was identifying mentions of the entities of interest, e.g., adverse reactions, in the forum posts. We used Brat—a web-based text annotation tool developed in the University of Tokyo [28]—to set up the annotations for this stage.

Forum posts on the drugs in the Diclofenac category were annotated by four medical students. A 22% fraction of the Lipitor posts were annotated by the medical students and the rest by two computer scientists. Three of the authors screened the annotations and corrected clear mistakes. For example, if the annotators had missed parts of a word (pai instead of pain) during annotation, we fixed the span. All these annotations were further reviewed by a clinical terminologist during the normalisation stage.

Fig. 1 shows examples of annotations using Brat. These examples show the language diversity of how different patients express their conditions. In example (a), Trazodone is misspelled as

Trazadone by the patient. Also there were two drug tags *thyroid* and *testostrone* (misspelled as testonrone) which were not drug names and should not have been annotated. Such cases were fixed in the reviews of the tags. Example (b) contains a colloquial term *charley horse* for painful spasms or cramps. It also is a complicated case of annotation where there are multiple discontinuous tags that share a common term. Example (c) shows a post written in all capital letters, where a list of adverse events are given without any punctuations in between. These examples emphasise the necessity of development of language processing techniques that are both capable of handling irregular and colloquial text, and handling medical concepts in such data.

The forum posts were divided evenly between the annotators, except for 55 documents that were given to all the annotators for the purpose of calculating the inter-annotator agreement. Two metrics were used for this calculation: *strict agreement* and *relaxed agreement*, as described by Metke-Jimenez et al. [22]. Both of these metrics are based on the average of the pair-wise agreement between the annotators as

$$agreement(i,j) = \frac{match(A_i, A_j, \alpha, \beta)}{max(n_{A_i}, n_{A_j})},$$

where A_i represents the set of annotations by annotator i, A_j represents the set of annotations by annotator j, n_{A_i} is the size of set A_i and n_{A_j} is the size of set A_j. $match(A_i, A_j, \alpha, \beta)$ is a function that counts the number of matching tags. The *match* function has two binary parameters: span strictness α and tag strictness β. Both these parameters can be either *strict* or *relaxed*. If span matching is strict, then the annotations being compared must match exactly. Consider the sentence "I experienced increased muscle tension". If one annotator annotates the text fragment "muscle tension" and another annotates the text fragment "increased muscle tension" then the *match* function with strict span matching will return no matches. If span matching is configured to be relaxed, then annotations that overlap will be counted as a match, with the restriction that each annotation can only be matched to one other annotation. For tag strictness, if both annotators annotate the same text fragment, for example "muscle tension", but one of them uses the tag ADR, and the other one uses the tag Symptom, then the function will return a valid match only if the tag strictness is relaxed.

Table 4 shows the inter-annotator agreement using different configurations of the agreement metric. When span and annotation settings were both relaxed, the average agreement for Diclofenac was approximately 78% and for Lipitor it was 95%. Note that agreements are for four annotators for Diclofenac and two for Lipitor. Therefore, we cannot directly compare these agreements among the two drugs.

4.3. Terminology association

While annotating entities in a text corpus is valuable in itself, linking these entities to standard terminologies provides another level of information on the corpus. Such a process is referred to as *normalisation* by other researchers, e.g., Pradhan et al. [25]. An example of such mapping in the corpora reviewed in Section 2 is Arizona Disease Corpus [17] in which disease names are mapped to their corresponding UMLS concepts.

To normalise the CADEC entities, a clinical terminologist reviewed the entities identified in the previous stage to map them to their representative concept in SNOMED CT, AMT, and MedDRA. During this process the entities that may have been wrongly annotated, were corrected.

4.3.1. Linking to SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms) is a clinical terminology that provides codes, synonyms and definitions of clinical terms, and can be accessed through the UMLS Metathesaurus.

The main benefit of using a standard vocabulary to normalise the terminology used in the forums is bridging the gap between the language of lay people and medical experts. It has been argued that coding clinical documents, such as clinical records, with SNOMED benefits statistical data collection by providing unambiguous, formal, standard terms describing clinically important information [29]. We therefore chose SNOMED CT as the target controlled vocabulary to map the entities.

The entities were classified into the following categories: ADR, Drug, Disease, Finding, and Symptom. All entities from each category with the exception of *Drug* were mapped to SNOMED CT-AU (SCT-AU) v20140531 by a clinical terminologist utilising the CSIRO Snapper tool [23]. Most entities were mapped in a one to one manner; however, a one to many mapping approach was undertaken in situations where the entity could be best described utilising more than one SCT concept. For example the entity "severe back pain" was mapped to the SCT-AU concepts 76948002 (Severe pain) and 161891005 (Backache). Table 5 lists some examples from the CADEC entities (last column, Original Entity) mapped to one or more concept(s) in SNOMED CT.

4.3.2. Linking to AMT

The Australian Medicines Terminology (AMT), developed and maintained by NeHTA Australia, is a terminology designed to describe and unambiguously identify medicines available in the Australian healthcare. Its intended use is specifically within software applications utilised in Australian healthcare environments. Only products registered with the TGA for the purposes of treating patients are defined within the AMT [24].

The entities from the Drug category were mapped to AMT V2.56 in a one to one manner. For the most part, drugs were mapped to *trade product* as mostly they were described in the text entries by their trade name. However, due to the international nature of the collection, some drugs which are readily available in the other countries are not available in Australia. For example, Advicor which is a combination of Niacin and Lovastatin is not present in AMT in either trade version or generic product so was therefore assigned the value *concept_less*. If drugs could not be found within the AMT utilising the trade name the generic form of the drug was then searched and if present utilised as the equivalent map. For example the drug Aciphex was mapped to its generic form 21296011000036107 (Rabeprazole) or the drug Cataflam was mapped to 21288011000036105 (Diclofenac). AMT is not structured to include concepts for *drug classes* so entities that were described as antibiotics or statins, for example, were assigned the value *concept_less*. Some drugs were too ambiguous to assign a concept from AMT, such as benadryl; AMT contains many versions of benadryl so it was not possible to definitively assign a mapping. Table 5 shows the mappings assigned to Cataflam and Demerol.

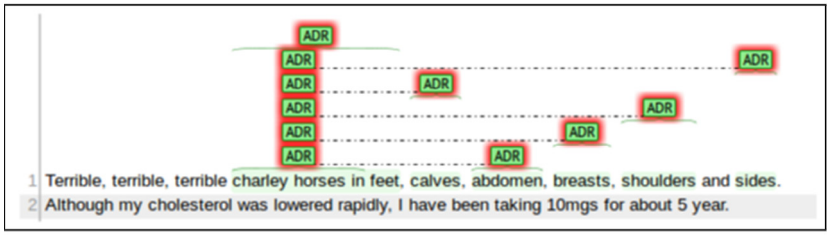
4.3.3. Linking to MedDRA

MedDRA[8] (Medical Dictionary for Regulatory Activities) is the standard thesaurus used by the pharmaceutical industry and regulatory agencies such as the FDA. It contains vocabulary used for adverse drug reactions structured in a hierarchy. The Lowest Level Term (LLT) is the most specific terminology that expresses an individuals condition such as "feeling queasy". One level up is Preferred Terms (PT), which together with LLTs are often used in

[8] http://www.meddra.org/.

(a) Drug Trazodone misspelled

(b) Multiple discontinious tags overlapping



(c) All capital letters and no punctuation used

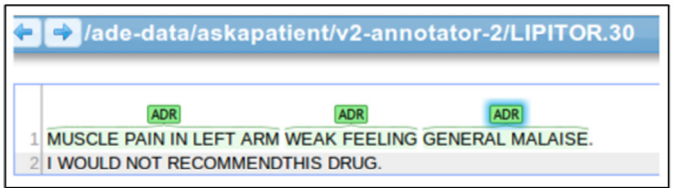


Fig. 1. Examples of entity annotation. (a) An example of drug misspellings; (b) An example of complicated annotations; and (c) An example of irregular text.

Table 4
Average pair-wise agreement (percentage) between annotators.

Span α	Tag β	Agreement	
		Diclofenac	Lipitor
Strict	Strict	46.6	74.2
Strict	Relaxed	49.1	81.6
Relaxed	Strict	68.7	85.1
Relaxed	Relaxed	77.9	94.8

Table 5
SNOMED CT and AMT normalisation examples.

SCT code	Concept	Original entity
76948002	Severe pain	Extreme pain in both shoulders
45326000	Shoulder pain	
284140004	Unable to move arm	Unable to reach the top of my head with my hands
70733008	Limitation of joint movement	Movement is restricted and it is impossible to make a fist
76948002	Severe pain	Extreme pain
271599002	Feeling content	Mellows me out
68962001	Myalgia	Muscle pain
AMT code		
21288011000036105	Diclofenac	Cataflam
34839011000036106	Pethidine	Demerol
concept_less	–	Felt much older than I was

reports of adverse drug events both by pharmaceutical companies and regulators. Upper levels in MedDRA hierarchy are more generic.

We used MedDRA V16.0 to annotate the concepts identified in the SNOMED CT mapping stage. These concepts were annotated at the LLT level, to capture the specific terms that express the patients' conditions. Table 6 shows examples of MedDRA normalisations from the CADEC corpus. The last column, Original Entity, is the entity identified in the first stage of annotation, and the first column is the corresponding MedDRA concept identifier. We list other information that can be retrieved from MedDRA given the MedDRA ID regarding that entity in the second and third columns. The second to last row is an example of one entity being mapped to two MedDRA IDs.

If an entity was mapped to more than one SCT concept, all the SCT concepts were then mapped to MedDRA. For example "severe back pain" which was mapped to two SCT concepts Severe pain and Backache, these concepts were then mapped to the MedDRA LLT concepts 10003993 (Backache) and 10033371 (Pain). However, it should be noted that in this particular case due to the classificatory nature of MedDRA the second map is superfluous as it does not add further definition.

4.3.4. Normalisation challenges

While mapping of free text entries such as these certainly gives an added layer to the knowledge bank, it can occasionally be somewhat subjective. Pain is a good example of such subjectivity. An entity such as "pain so bad I thought I was going to die", can be coded either as *severe pain* or *excruciating pain*. In terms of

Table 6
MedDRA normalisation examples.

MedDRA ID	Preferred name	Classified as	Original entity
10040617	Shoulder pain	Musculoskeletal pain	Extreme pain in both shoulders
10033407	Pain hunger	Abdominal pain upper	Hunger pangs
10038742	Restless legs	Restless legs syndrome	Legs are restless
10043890	Tiredness	Fatigue	Very tired
10069830	Unable to eat	Aphagia	Couldn't eat or drink
10069830	Unable to eat	Aphagia	Can't eat normal
10013781	Dry mouth	9 classes, e.g., Oral dryness	Dry mouth
10003068	Aptyalism	4 classes, e.g., Asialia	
concept_less	–	–	Lungs feel heavy

assigning a MedDRA concept this particular example of subjectivity is not such an issue as MedDRA being classificatory in nature, is only interested in the *pain*. Other examples such as "I couldn't get out of bed" were assigned a *concept_less* value due to the uncertainty of what the person was trying to convey. It is unclear if they meant "I am so tired I can't get out of bed" or "I didn't feel like getting out of bed" or "I am physically incapable of getting out of bed due to physical impairments". The SCT concept which refers to "getting out of bed" is utilised to describe a person's physical ability to "get out of/or off the bed", i.e. if these persons are able to sit up by themselves, swing their legs to the side and get to a standing position. MedDRA, however, does not have a term for "unable to get out of bed".

There were some entities that were parsed such that they missed subtle context. For example "Memory loss/ability to concentrate" was parsed as "memory loss" and "ability to concentrate". However it is more likely that the entry was meant to convey "memory loss" and "loss of ability to concentrate" hence the latter was mapped to a concept equivalent of "loss of ability to concentrate".

5. Corpus statistics

In the first stage of the annotation, 64 posts (5.1%) did not receive any entity annotation and therefore did not require normalisation. After the normalisation stage, a further 78 posts (total of 142 posts or 11.4%) received no MedDRA annotations. An example of a post that did not receive any entity annotation was a post on Lipitor with the content: *I did not experience any of the myriad of possible side effects of the drug*.

We created a consolidated corpus in which for the portion of the corpus that received multiple annotations, only one set based on random choice of an annotator, was used for calculating the statistics. A better solution for a corpus that is annotated by multiple annotators would have been following the centroid method by Lewin et al. [20].

Table 7 lists the frequency of the annotated entities in the entire corpus as well as in each drug category (Lipitor and Diclofenac) separately. A total of 9111 entities were identified. ADRs comprised 69.3% of the total number of entities, followed by drugs (1800 or 19.8%). From all the 9111 annotated entities, only 39.4% (3591 entities) were unique; people generally reported similar reactions.

The number of entities annotated as Symptom was larger for Diclofenac than Lipitor (239 compared to 38) even though the number of Diclofenac posts was much smaller. The reason being, often patients mentioned *pain* as their symptom or the reason for taking the medication, whereas for Lipitor/Atorvastatin the symptoms related to Hypercholesterolemia (high cholesterol) were not mentioned in the posts as often.

Table 7
Number of entities annotated in the entire corpus and each of the drug categories, as well as the unique number of entities per each category.

Entity	Corpus		Diclofenac		Lipitor	
	All	Unique	All	Unique	All	Unique
ADR	6318	2713	888	508	5445	2316
Disease	283	162	59	42	226	129
Drug	1800	321	246	113	1542	222
Symptom	275	130	239	103	38	30
Finding	435	265	41	34	394	238
All	9111	3591	1473	800	7645	2935

Table 8
Number of continuous and discontinuous tags of each type.

	ADRs	Diseases	Symptoms	Findings	Drugs	All
Continuous	5318	280	255	397	1797	8047
Discontinuous (overlapping)	918	2	13	34	2	969
Discontinuous (non-overlapping)	82	1	7	4	1	95
Total	6318	283	275	435	1800	9111

Table 9
Most frequent entity values for each of the drug categories with their frequencies in brackets.

Entity	Top 5 for Diclofenac	Top 5 for Lipitor
ADR	Diarrhea (28), nausea (25), vaginal bleeding (17), cramps (16), dizziness (14)	Pain (185), fatigue (84), depression (83), muscle pain (82), memory loss (62)
Disease	Arthritis (10), endometriosis (3), migraines (2), bi-polar (2), plantar fasciitis (2), ibs (2), post traumaic osteoarthritis (1), lower lumbar arthritis (1)	Heart attack (14), arthritis (13), diabetes (8), fibromyalgia (8), ms (6)
Drug	Arthrotec (53), voltaren (25), celebrex (9), cataflam (6), advil (6), ibuprofen (6), aleve (5)	Lipitor (1034), zocor (44), pravachol (22), crestor (20), coq10 (19)
Finding	Stroke (2), menopause (2), heart attack (2), stomach trouble (1), arthritis (1)	Arthritis (17), high cholesterol (15), heart attack (13), stress (8), pain (7)
Symptom	Pain (89), knee pain (8), back pain (6), inflammation (4), agony (3), lower back pain (3)	Pain (3), inflamation (2), menopause (2), low bp and pulse (1), anxiety (1)

Table 8 reports the number of continuous and discontinuous tags that were annotated during the entity annotation step. Continuous tags represent a continuous set of words, where as discontinuous tags are broken into multiple spans within a sentence. Examples of such discontinuous tags are shown in Fig. 1 (b).

We then looked at the entities that were annotated with any of the labels. Table 9 lists the most frequent entity values for each of the five types of entities (Drug, ADR, Disease, Symptom, and Finding) for the two drug categories. Given the two drug categories in the corpus treat very different conditions, their entity values for almost all the entity categories were different as well. For example, there was a clear distinction between the sets of drugs mentioned for Diclofenac and Lipitor. However, the drugs mentioned in the posts were often alternative or complementary drugs taken by the patients, and they rarely mentioned other medications that they were taking while on the drug that was the subject of their post.

Annotations from the normalisation stage were also analysed with statistics shown in Table 10. The final set contained 1655 concepts from AMT, from which only 123 (7.4%) were unique. In other

Table 10
Statistics on entities normalised with AMT, SNOMED CT (SCT) and MedDRA.

	AMT	SCT	MedDRA
No. annotated concepts	1655	7259	6569
No. unique concepts	123	924	686
Mean per post	1.2	5.4	5.2
Maximum no. concepts per post	17	42	34
1st most freq. concept	Lipitor (1073)	Pain (371)	Pain (485)
2nd most freq. concept	Arthrotec (62)	Myalgia (288)	Myalgia (311)
3rd most freq. concept	Zocor (48)	Severe pain (196)	Arthralgia (311)

words, the same drug names (mostly Lipitor) were repeated across all the posts. On average, there was 1.2 drug name concepts per post linked to AMT. In an extreme case, we had a post with 14 drug brand names which linked to AMT. Similar statistics are reported for SCT and MedDRA in Table 10. We also report the three most frequently linked concepts together with their frequencies across the corpus for each terminology at the bottom of Table 10.

6. Lessons learnt

In the process of creating the CADEC corpus, a number of decisions were made with regard to what should be annotated, rules associated with this and the tools best suited to the completion of the exercise. The main web-based tool used was Brat. It provided a number of advantages: creating character offset based annotations in plain text files that were easy to process, as well as an easy to set up and annotate interface. However, one shortcoming of this tool was automatic provision of a default tag in case an annotator highlights a fraction of the text but forgets to select the suitable tag. In our case, Brat had chosen the tag *Drug* for such cases. We therefore introduced an additional step to review the annotations and modify the incorrect tags.

7. CADEC limitations

CADEC imposes the following limitations that could affect the studies that use this dataset:

Data type Social media text is often noisy and contains inaccurate, incomplete, or even false information.

Data source There are likely limitations inherited from our data source: AskaPatient. There might be a specific group of consumers that use AskaPatient to review their medications. There might also be implications inherited from the data structure used in AskaPatient to collect the reviews.

Data size The size of the corpus, 1253 posts, is limited which is not representative of all the drugs and consumer reviews existing on the web.

Limited coverage of drug types and adverse effects CADEC only includes drugs that contained Diclofenac and Atorvastatin in their active ingredients. Therefore, the range of adverse events in the corpus are those of 12 drugs that are present in the corpus, and adverse effects specific to other drug types are absent. Even for the drugs covered in the dataset, the coverage of rare adverse events known as, idiosyncratic drug reactions,[9] is limited.

Lack of drug–drug interactions and drug overdose annotations The information provided in the reports by the consumers are often focused on one specific medication. Therefore, information on other drugs that are administered by the consumers are often missing. This causes lack of information on potential drug-drug interactions that have caused the reported adverse reactions. The same applies for the dosage and frequency of medication usage. Whether the patient has experienced adverse effects due to overdose or not is not known due to the nature of these reports. We have not annotated the corpus for drug-drug interactions or intoxication due to overdose.

Annotator errors Despite our efforts for precise annotations, the dataset may contain human errors in identifying the concepts or linking them to the medical terminologies. Annotation errors will affect the evaluation of any automated system that uses CADEC as gold standard.

8. Conclusions and future work

We created an open access corpus named CADEC (CSIRO Adverse Drug Event Corpus) for researchers of text mining for pharmacovigilance. The corpus is composed of medication consumer posts from a medical forum, AskaPatient, annotated with concepts such as drug names, adverse reactions, diseases, and symptoms. These concepts are linked to their corresponding concepts in controlled vocabularies. The CADEC corpus provides opportunities for researchers in a number of areas to (1) develop and evaluate systems that automatically extract adverse drug events from layperson reports; (2) develop systems that extract medications from free-text; (3) develop systems that automatically map free-text to SNOMED CT or MedDRA, because it contains a rich mapping of informal terminology to formal, as expressed in SNOMED CT and MedDRA; and (4) employ variants in expressing one medical condition by laypeople, or spellings of drug names, that are captured in CADEC, in addition to the Consumer Health Vocabulary (CHV) in dictionary-based text mining algorithms.

The challenges of annotating data from medical forums which comprise irregular text and colloquial language were also discussed. Most notably, in such data, text spans are often ambiguous, creating challenges for standard text processing techniques.

We are in the process of annotating the corpus with relationships as well as including other entities such as medication dosage and frequency.

Acknowledgements

AskaPatient kindly provided the data used in this study for research purposes only. The CADEC corpus is under CSIRO Data Licence. This licence allows users to use the data for non-commercial purposes with appropriate attribution.

Ethics approval for this project was obtained from the CSIRO ethics committee which classified the work as low risk (CSIRO Ecosciences #07613).

We would like to thank the University of Queensland medical students, Timothy Sladden, Thomas Souchen, Warren Brown, and Digvijay Khangarot, who contributed to the development of the guidelines and annotations of the CADEC corpus.

References

[1] ACSQHC, National safety and quality health service standards, 2012.
[2] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. Leonard, J. Holmes, Identifying potential adverse effects using the web: a new approach to medical hypothesis generation, J. Biomed. Inform. 44 (6) (2011) 989–996.

[9] Idiosyncratic drug reactions are also known as Type B side-effects and are seen very rarely among individuals that their immune system react to certain medications even in small dosage.

[3] J. Berlin, S. Glasser, S. Ellenberg, Adverse event detection in drug development: recommendations and obligations beyond phase 3, Am. J. Public Health 98 (8) (2008) 1366–1371.
[4] B. Chee, R. Berlin, B. Schatz, Predicting adverse drug events from personal health messages, in: AMIA Annual Symposium Proceedings, Washington, DC, 2011, pp. 217–226.
[5] K. Cohen, P. Ogren, L. Fox, L. Hunter, Corpus design for biomedical natural language processing, in: The ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Detroit, Michigan, 2005, pp. 38–45.
[6] P. Coloma, G. Trifiró, V. Patadia, M. Sturkenboom, Postmarketing safety surveillance, Drug Safety 36 (3) (2013) 183–197.
[7] J. Couzin, Drug safety: gaps in the safety net, Science 307 (5707) (2005) 196–198.
[8] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, I. Solti, Building gold standard corpora for medical natural language processing tasks, in: AMIA Annual Symposium, Washington, DC, 2012, pp. 144–153.
[9] R. Dogan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.
[10] H. Gurulingappa, R. Klinger, M. Hofmann-Apitius, J. Fluck, An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature, in: The 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (7th edition of the Language Resources and Evaluation Conference), 2010, pp. 15–22.
[11] H. Gurulingappa, A. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, Biomed. Inform. 45 (5) (2012) 885–892.
[12] M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, Annotation issues in pharmacological texts, Proc. – Social Behav. Sci. 95 (2013) 211–219.
[13] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions, Biomed. Inform. 46 (5) (2013) 914–920.
[14] B. Hug, C. Keohane, D. Seger, C. Yoon, D. Bates, The costs of adverse drug events in community hospitals, Joint Commission J. Quality Patient Safety 38 (3) (2012) 120–126.
[15] S. Karimi, S. Kim, L. Cavedon, Drug side-effects: What do patient forums reveal? in: The 2nd International Workshop On Web Science and Information Exchange in the Medical Web, Glasgow, UK, 2011, pp. 14–15.
[16] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, Text and data mining techniques in adverse drug reaction detection, ACM Comput. Surveys (2015) in press.
[17] R. Leaman, C. Miller, G. Gonzalez, Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark, in: The 3rd International Symposium on Languages in Biology and Medicine, Jeju Island, South Korea, 2009, pp. 82–89.
[18] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks, in: Workshop On Biomedical Natural Language Processing, Uppsala, Sweden, 2010, pp. 117–125.
[19] G. Leech, Corpus annotation schemes, Literary Linguist. Comput. 8 (4) (1993) 275–281.
[20] I. Lewin, S. Kafkas, D. Rebholz-Schuhmann, Centroids: gold standards with distributional variations, in: The Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012, pp. 3894–3900.
[21] X. Liu, H. Chen, AZDrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums, in: The 2013 International Conference On Smart Health, Beijing, China, 2013, pp. 134–150.
[22] A. Metke-Jimenez, S. Karimi, C. Paris, Evaluation of text processing algorithms for adverse drug event extraction from social media, in: The First International Workshop on Social Media Retrieval and Analysis, Gold Coast, Australia, 2014, pp. 15–20.
[23] J. Michel, M. Lawley, A. Chu, J. Barned, Mapping the Queensland health iPharmacy medication file to the Australian medicines terminology using Snapper, Stud. Health Technol. Inform. 168 (2010) 104–116.
[24] NEHTA, Australian Medicines Terminology v3 model–common v1.4, Tech. rep. EP-1825:2014, National E-Health Transition Authority, 2014.
[25] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, G. Savova, Semeval 2014 task 7: analysis of clinical text, in: The 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 2014, pp. 54–62.
[26] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, A. Setzer, Building a semantically annotated corpus of clinical texts, Biomed. Inform. 42 (5) (2009) 950–966.
[27] E. Roughead, S. Semple, Medication safety in acute care in Australia: where are we now? Part 1: A review of the extent and causes of medication problems 2002–2008, Australia New Zealand Health Policy 6 (1) (2009) 18.
[28] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for NLP-assisted text annotation, in: The Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107.
[29] D. Truran, P. Saad, M. Zhang, K. Innes, SNOMED CT and its place in health information management practice, Health Inform. Manage. 39 (2) (2010) 37–39.
[30] E. Van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. Kors, L. Furlong, The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships, Biomed. Inform. 45 (5) (2012) 879–884.
[31] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucl. Acids Res. 41 (W1) (2013) 518–522.
[32] D. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, Drug Bank: a comprehensive resource for in silico drug discovery and exploration, Nucl. Acids Res. 34 (2006) 668–672.
[33] C. Yang, L. Jiang, H. Yang, X. Tang, Detecting signals of adverse drug reactions from health consumer contributed content in social media, in: ACM SIGKDD Workshop On Health Informatics, Beijing, China, 2012.