Big Data Research ••• (••••) •••–•••

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Towards Human–Machine Collaboration in Creating an Evaluation Corpus for Adverse Drug Events in Discharge Summaries of Electronic Medical Records

Pei San Ang [a,*], Liza Y.P. Fan [a], Mun Yee Tham [a], Siew Har Tan [a], Sally B.L. Soh [a], Belinda P.Q. Foo [a], Celine W.P. Loke [a], Shangfeng Hu [b], Cynthia Sung [a,c]

[a] *Vigilance and Compliance Branch, Health Products Regulation Group, Health Sciences Authority, Singapore*
[b] *Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore*
[c] *Duke-NUS Medical School, Singapore*

ARTICLE INFO

Article history:
Received 28 December 2015
Received in revised form 29 March 2016
Accepted 11 April 2016
Available online xxxx

Keywords:
Adverse drug events
Discharge summaries
Narrative medical reports
Natural language processing
Text mining

ABSTRACT

Adverse drug events (ADEs) contribute significantly to morbidity and mortality in the healthcare system. The availability of digitalised hospitals' narrative clinical data offers a potentially rich resource to enhance pharmacovigilance efforts to manage potential safety issues arising from real-world use of drugs. The goal of this paper was to establish a foundation for creating an evaluation corpus by developing a set of annotation guidelines to achieve high inter-annotator agreement (IAA) and to evaluate the performance of basic entity identification tools for drugs, adverse events (AEs) and drug-AE relationships from 100 discharge summaries of a tertiary hospital in Singapore. Two teams of three annotators worked independently on text annotation using Knowtator. Three-way IAA of 86%, 70% and 49% were achieved for drugs, AEs and drug-AE relationships respectively. The performance of the machine algorithm was evaluated against annotations made by at least two annotators, with a recall of 84% and precision of 73% for drugs and a recall of 67% and precision of 53% for AEs. The high recall and precision for drug entity extraction suggests that machine pre-annotation of drugs followed by human annotation of AEs and drug-AE relationships could be a feasible approach in expediting the process of creating a larger evaluation corpus. Non-matches between machine and human annotations were examined to identify ways to further refine the algorithm. When successfully implemented, the identification of ADEs could greatly support pharmacovigilance work in characterising the magnitude and scope of ADEs and prioritising interventions to improve the drug safety.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Adverse drug events (ADEs) are a major preventable cause of morbidity, hospitalisation, and death, and result in direct costs of millions of dollars every year [1]. Pre-market clinical trials do not always reflect the conditions under which the drug will be used in routine practice. Oftentimes, these trials collect data on a few thousand subjects, hence not all ADEs can be known before a drug is approved and used by millions of patients [2].

Safety monitoring is an ongoing process throughout the product's life cycle that begins from pre-market clinical trials and continues after it is made available to patients. Drug regulatory agencies conduct various post-market surveillance activities to manage potential safety issues that arise from real-world use of drugs, and then communicate benefit-risk analyses to stakeholders such as patients, healthcare professionals (HCPs), and companies on an ongoing basis. In Singapore, the drug regulatory agency, Health Sciences Authority (HSA), receives spontaneous ADE reports primarily from HCPs. HCPs can submit an ADE report via fax, email, telephone, online via HSA's website or a module integrated in the electronic medical record (EMR) of public healthcare institutions, known as Critical Medical Information Store (CMIS) [3,4]. Using CMIS, a clinician reports the ADE to HSA directly from the patient's EMR without needing to fill out a separate ADE report. Although spontaneous reporting of ADEs by HCPs has been the cornerstone of the pharmacovigilance programme, it is well recognized that such a voluntary system is subjected to a variable and unknown

[*] Corresponding author at: Vigilance and Compliance Branch, Health Products Regulation Group, Health Sciences Authority, 11 Biopolis Way #11-01 Helios, 138667, Singapore.
E-mail address: ang_pei_san@hsa.gov.sg (P.S. Ang).

http://dx.doi.org/10.1016/j.bdr.2016.04.001
2214-5796/© 2016 Elsevier Inc. All rights reserved.

degree of under-reporting, as generally, only a fraction of all ADEs is actually reported to the regulatory agency [5].

A large volume of clinical data is captured in hospitals and outpatient clinics as part of routine clinical care. In Singapore, all public hospitals and government-supported outpatient clinics have EMR systems for clinical management of patients, which include pharmacy records, laboratory tests, and discharge summaries. Various efforts to leverage these data for ADE case identification have been explored, a few of which are described [6–9], although it is unclear to what extent these have been implemented in a clinical or regulatory setting. While spontaneous ADE reporting relies on healthcare professionals to report ADEs, other systematic ways of capturing ADE using keywords from EMRs have been suggested [10]. In order to gain a more comprehensive view of the overall ADE landscape in Singapore, one approach we are exploring is text mining of hospital discharge summaries. The ability of text mining algorithms to extract ADE information from discharge summaries could significantly enhance pharmacovigilance efforts [11–13]. Hospital discharge summaries contain rich narratives about the medical conditions of the patients, drug allergies, adverse events, laboratory investigations, procedures, treatment and outcomes. However the discharge summaries are not in a structured format, vary in content and style from clinician to clinician, and contain numerous abbreviations, spelling errors, acronyms, sentence fragments and ungrammatical constructs. Their meanings are often ambiguous depending on the context [14]. These present significant challenges to building computationally efficient and accurate bioinformatics algorithms that search discharge summaries for drugs and AEs.

This paper presents a proof-of-principle project for development of a set of guidelines for annotators and achieving consistency of annotation as a foundational step for creation of a larger evaluation corpus, construction of reference gazetteers, and evaluation of the performance of basic text mining tools to extract drugs and AEs from unstructured text of hospital discharge summaries. Pharmacovigilance staff at HSA developed guidelines for curating discharge summaries from anonymized electronic medical records, then applied those guidelines to two sets of fifty discharge summaries to measure inter-assessor agreement and compare to machine extraction of drug and AE entities. The results are guiding decisions on how to efficiently create a larger evaluation corpus for testing the performance of text mining algorithms to identify drugs suspected to be associated with specific AEs. The ultimate goal is to develop streamlined systems to identify potential drug safety signals.

2. Methods

2.1. Electronic medical records

De-identified pharmacy records and discharge summaries were obtained from the National University Hospital (NUH) in Singapore for the period January 2000 to March 2012 after ethics approval from the Domain Specific Review Board, the Institutional Review Board of the National Healthcare. NUH is a tertiary hospital with medical, surgical and dental specialities. The set of records contained 660,838 discharge summaries.

2.2. Creation of drug and AE gazetteers

A gazetteer of 4546 drug terms and 129 drug classes (such as statin, antibiotic, analgesic, NSAID) was built using the NUH's inpatient and outpatient pharmacy drug records and HSA's database of registered drugs. Many brand names included the dosage form and strength. For the purpose of creating the gazetteer, the dosage form and strength were trimmed away. For examples, Norvasc 10 mg OM was trimmed to Norvasc.

A gazetteer of AEs was constructed by combining the lowest level terms of the WHO Adverse Reaction Terminology (WHO-ART) [15] and the Medical Dictionary for Regulatory Activities (MedDRA) Terminology [16]. Both WHO-ART and MedDRA are standardised medical terminologies which are used internationally to code clinical information such as AEs, diseases and diagnoses. Four system categories in the MedDRA were excluded i.e. investigations, injury poisoning and procedural complications, social circumstances, and surgical and medical procedures as these are unlikely to be related to AEs. The AE gazetteer of combined WHO-ART and MedDRA terms contained 44,319 unique medical terms. An additional list of 2034 synonyms (such as GI for gastrointestinal, liver for hepatic) and medical abbreviations commonly used at NUH and another public sector hospital, Changi General Hospital, were included.

2.3. Text segmentation

In general, physicians write discharge summaries in distinct sections. A section title list is generally given to recognize the different sections. The first step in the algorithm was to split and classify the sections. For example, the discharge summaries typically contained an investigation section with laboratory and imaging test results. The section would start with words like investigation, initial invx, inx or IX and end with separation spacing. This section was not annotated because drug terms (such as folate and digoxin) mentioned in the investigation section were generally drug concentration measurements in the blood or tissue fluid. If there was an ADE, it likely would be mentioned in the other parts of the discharge summaries. Text pre-processing such as tokenization was employed to process the rest of the discharge summary to identify meaningful keywords [17,18].

2.4. Negation

To avoid tagging drugs which were not given to patients or AEs that did not occur, human annotators used the list in Table 1 as a guideline for negation phrases. Examples of phrases that contained drugs or drug classes that were not tagged are: “bacteria is sensitive to amoxicillin”; “spoken to family not to start warfarin”; “did not undergo chemotherapy”; “kiv to start bisphosphonates”; “discharged with standby prescription of augmentin”. Examples of AEs that were not tagged are: “test was done to rule out fracture”, “nil fever/chills/chest pain”, “no evidence of bleeding”, “no complaints of wheeze”, “BP stable no sign of sustained hypertension”.

The machine algorithm used the negation algorithm NegEx [19] to identify negation phrases. The algorithm differentiated the negation words or phrases appearing before or after an entity. Drug and AE terms found within the sentence of negation phrases were excluded.

2.5. Entity tagging by human annotators

Knowtator is a Window-based text annotation tool that allows the incorporation of domain knowledge into an annotation schema for semantic annotation. Knowtator leverages on Protégé representation system as a Protégé plugin. Two teams of three annotators manually tagged the drugs, AEs, and drug-AE relationships using Knowtator version 1.9 beta [20]. All annotators possessed a minimum qualification of a Bachelor's degree in Pharmacy, in addition to prior experience working in hospitals and reviewing voluntary ADE reports submitted to HSA. None had previous experience in text mining. Manual annotations were used to evaluate inter-annotator agreement and to benchmark machine performance.

Table 1
Guideline of negation phrases used by human annotators.

	List of negation phrases for drugs		List of negation phrases for AE
1.	d/c (this means discharge)	1.	absence of
2.	declined	2.	denied
3.	defaulted	3.	denies
4.	denied	4.	denying
5.	did not take	5.	did not exhibit
6.	did not undergo	6.	nil
7.	drugs given on previous discharge	7.	no evidence of
8.	if	8.	no sign of
9.	KIV (keep in view)	9.	no signs of
10.	not for	10.	not demonstrate
11.	not keen for	11.	not demonstrated
12.	not started	12.	patient was not
13.	not to start	13.	ruled out
14.	not to take	14.	rules out
15.	not yet given	15.	unlikely
16.	option of	16.	negative for
17.	plan	17.	no cause of
18.	refused	18.	no complaints of
19.	rejected	19.	without indication of
20.	sensitive to, resistant to, S to, R to	20.	without sign of
21.	unlikely		

Table 2
Rules for tagging drugs used by human annotators.

To tag as drug
1. when the drug was given, could have been given to patient
2. without the salt name, dose, strength and route of administration
3. tag the brand name and active ingredient as separate entities

Not to tag as drug
1. when the drug was not started or patient declined the drug
2. when it was a laboratory test for finding out bacteria sensitivity
3. terms such as amylase, blood products, plasma, platelet transfusion, dextrose, saline, health supplements, food supplement, traditional medicines and medical devices

To practice the use of Knowtator and establish a common annotation guideline, annotators initially annotated 50 discharge summaries. Afterwards, the annotators met to discuss annotation rules in order to develop consistency among the annotators (Table 2). The exercise was repeated with another set of 50 selected discharge summaries. Non-drug terms such as amylase, blood products, plasma, platelet transfusion, dextrose, saline, health supplements, food supplement, traditional medicines and medical devices were not tagged. If the drug term was mentioned as a laboratory test for finding bacterial sensitivity e.g. “bacteria is sensitive to penicillin”, the drug term would not be tagged. The active ingredient (amoxicillin) of a drug would be tagged without the name of the salt (amoxicillin sulfate). If the brand name and active ingredient were mentioned side-by-side, each term was tagged separately. For example, “panadol (paracetamol)” would be tagged as two terms i.e. “panadol” and “paracetamol”.

Table 3 lists the rules used by the annotators to tag AEs. For example, when AEs appeared with adjectives such as low, high, abnormal, aggravated, acute and chronic, the adjective was tagged along with the AE. The anatomic location was also included in the tagged phrase, with the exception of location terminology such as left, right or bilateral. Culture findings, bacterial names and histological findings were also excluded.

2.6. Entity extraction by machine

Algorithms for entity extraction were written in Java, building upon Apache UIMA. Drug and AE entity names are highly domain specific, hence the entity extraction method was keyword-based. To extract the entities with minor difference from the keywords, fuzzy searches of the drug phrases with length of more than seven characters were allowed using a Levenshtein distance with a threshold of one. Levenshtein distance is a measure of the similarity of two words [21,22]. This meant that any two words could potentially be matched with the insertion, deletion or substitution of one character. This allows for identification of misspelled words. For example, incorrect spellings such as “?fistula” and “vomitting” would be annotated as AEs. If the drug name was equal or less than seven characters, an exact match of the drug name would be carried out, though it would allow for abbreviations as contained in the drug gazetteer. The searches allowed for both upper and lower cases.

Table 3
Rules for tagging AEs used by human annotators.

	To tag as AE
1.	abnormal, atypical, low, high, elevated, increasing, decreased
2.	acute, chronic
3.	location/anatomy: upper, lower, back, leg
4.	worsening, aggravated, progression, recurrent, persistent, extensive, enhancing, frequent
5.	smaller, bigger
6.	dry, productive
7.	issue, disorders, symptoms, features, episodes
8.	drug-induced
9.	location of AE e.g. facial pain, injection site pain, chest pain, muscle pain (with the exception of skin related AEs)
General terms:	
1.	ADR, ADE, AR, AE
2.	drug reaction
3.	allergy, drug allergy

	Not to tag as AE
1.	left, right, bilateral
2.	mild, moderate, severe, serious
3.	small, big
4.	suspected
5.	no culture, no bacterial names, no histological findings
6.	location for skin related AEs e.g. maculopapular rash, scaly rash, skin eruption, pruritus, urticarial
7.	values of lab results e.g. ALT 300

2.7. Proximity finder

Drug terms throughout the discharge summaries were tagged, with the exception of the investigations section. AE phrases found only within three sentences from any drug term were included, based on the assumption that drug-AE relationships would be found close to the drug terms. The boundaries of three sentences were defined as sentence breaks with full stop, paragraph breaks and bullet points.

2.8. Relation classification

Certain words and phrases suggested potential drug-AE relationships such as “allergic to”, “associated with”, “likely to,” “secondary”, and “switched” (Table 4). In the phrase “previously on donepezil, discontinued due to GI side effects, switched to memantine”, the words “discontinued” and “switched” in proximity to the drug “donepezil” indicate a relationship to the AE “GI side effects”, and the drug “donepezil” and “GI side effects” were tagged as a drug-AE relationship. Similarly, in the phrase “started to develop (sic) rash all over body and face a/w upper lip swelling (4 days after starting augmentin)”, the AE relationship of “augmentin” to “rash” was tagged. For cases when multiple terms were mentioned such as “allergic to ibuprofen (angioedema)”, both “allergic ⟷ ibuprofen” and “angioedema ⟷ ibuprofen” were tagged as drug-AE relationships.

Please cite this article in press as: P.S. Ang et al., Towards Human–Machine Collaboration in Creating an Evaluation Corpus for Adverse Drug Events in Discharge Summaries of Electronic Medical Records, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2016.04.001

Table 4
Trigger phrases to suggest presence of drug-AE relationship.

	Words suggesting of drug-AE relationship
1.	a result of
2.	allergic to
3.	allergy to
4.	associated with
5.	caused by
6.	developed + changed
7.	due to
8.	held off in view of
9.	incorrect timing of dosage
10.	interaction with medications
11.	likely to
12.	possibly due to
13.	secondary
14.	stopped/switched/discontinued

2.9. Experiments to evaluate inter-annotator agreement

After reaching agreement on an annotation guideline, annotators worked independently on tagging another two sets of 50 discharge summaries each, randomly selected from the period January 2011 and December 2011. In test set 1, three annotators in group A were tasked to tag 50 discharge summaries for drugs only, and the three annotators in group B tagged the same 50 discharge summaries which had been pre-annotated for drug entities by the machine algorithm. Discharge summaries without drug pre-annotation were referred to as “plain records”. Annotators working on pre-annotated records were tasked with correcting any misidentified drugs and tagging any drugs missed by the machine algorithm, while leaving correctly annotated drugs unchanged. In test set 2, the roles of the two groups were switched. Tagging was done without any consultation or discussion with another annotator. The number of minutes taken to annotate the set of plain records, as well as the number of minutes to review and correct the set of machine pre-annotated records were recorded. The amount of time for the machine to prepare the pre-annotated records was also noted. Results of each set of annotated discharge summaries were analysed for each group to determine the inter-annotator agreement. Knowtator calculates a three-way inter-annotator agreement (IAA) based on class match, which does not require having the exact boundaries as long as the annotations overlap. Class match three-way IAA was defined as the number of class matches divided by the sum of matches and non-matches.

After drug annotation exercises were completed, three annotators reviewed the results and reached consensus on the drug entities in the 100 discharge summaries. The consensus set was then used for the next stage of simultaneously annotating AEs and drug-AE relationships that occurred within three sentences of a drug name. Group A tagged AEs and drug-AE relationships in test set 1 while group B worked on test set 2.

2.10. Comparison of human and machine annotation

Machine extractions for drugs and AEs were compared to results of entities tagged by at least two annotators to calculate recall and precision. A key use for the text mining algorithm is to flag out discharge summaries that are likely to contain an entity of interest for subsequent review and verification by clinicians. Therefore, we set a higher target for recall at 80% than the target for precision at 60% as pre-specified goals.

Table 5
Annotation of drugs in set 1 and 2

	Set 1 (50 records)	Tag drugs				
		Plain records		Machine pre-annotated		3-way IAA
		Drugs tagged	Time (min)	Drugs tagged	Time (min)	
Group A	Annotator 1	306	155	–		88%
	Annotator 2	285	99	–		
	Annotator 3	307	147	–		
Group B	Annotator 4	–		305	150	86%
	Annotator 5	–		290	120	
	Annotator 6	–		273	134	
	Set 2 (50 records)	Tag drugs				
		Plain records		Machine pre-annotated		3-way IAA
		Drugs tagged	Time (min)	Drugs tagged	Time (min)	
Group A	Annotator 1	–		234	145	82%
	Annotator 2	–		197	73	
	Annotator 3	–		234	116	
Group B	Annotator 4	231	135	–		84%
	Annotator 5	223	150	–		
	Annotator 6	204	127	–		

3. Results

3.1. Inter-annotator agreement

Overall, we achieved a three-way IAA of 88% and 84% for drugs on plain records and 86% and 82% on pre-annotated records for Set 1 and Set 2, respectively (Table 5). The average time taken to annotate drugs in 50 plain records was 136 minutes (2 hours 16 minutes), while the average time taken to correct machine pre-annotated records was 123 minutes (2 hours 3 minutes).

3.2. Comparison of human and machine annotations

To assess machine annotation performance, precision and recall of entity extraction were compared to human annotation results based on matching at least two annotators. For drug annotation, recall was 84% and precision was 73% (Table 6), thereby meeting the pre-specified target goals. Examples that the machine incorrectly tagged as drug entities were “tension” and “moderate”. The drug gazetteer contains drug brands “Tensilon” and “Modecate”, hence the fuzzy matching algorithm incorrectly tagged those words as drugs. To improve machine performance further, the algorithm could be modified to disallow tagging of normal English words as drugs. Another misclassification was not tagging “anti tetanus IV” as a drug. Instead, during the stage of AE tagging, “tetanus” was misclassified as an AE. Examples that the machine algorithm missed tagging were brand names (e.g. Madopar), drug abbreviations for a drug (e.g. piptazo) or class of drugs (e.g. OHGAs for oral hypoglycemic agents) that were not contained in the drug gazetteer, or typographical errors with non-alphanumeric characters like “)” and “/”. The drug gazetteer can be updated with the additional drug brand names and abbreviations, and the fuzzy matching can be modified to accept the other characters as allowable substitutions.

For AE annotation, recall was 67% and precision was 53% (Table 7) based on matching to at least two annotators. Examples that the machine incorrectly tagged as AEs were the use of “diabetic” in the phrase “Metformin withheld – S/B diabetic nurse”, and “anxiety” in the phrase “Agreeable to take medication and f/u with anxiety clinic”, when these words were used as adjec-

Table 6
Evaluation of machine performance for drugs.

Set 1 and 2 combined	Annotated by 2 humans	Not annotated by 2 humans	Precision
Annotated by machine	444	165	73%
Not annotated by machine	85		
Recall	84%		

Table 7
Evaluation of machine performance for AEs.

Set 1 and 2 combined	Annotated by 2 humans	Not annotated by 2 humans	Precision
Annotated by machine	291	262	53%
Not annotated by machine	145		
Recall	67%		

Table 8
Inter-annotator agreement (IAA) by class matcher.

Set 1 and 2 combined			
Type	IAA	Matches	Non-matches
Drug (plain records)	86%	1341	215
AE	70%	951	416
Drug-AE relationship	49%	66	70

tives for nurse and clinic, respectively. The machine algorithm also mistakenly tagged individual words like “high” in the phrase “MRI brain (21/7/11) high parietal lesion”, where the word “high” was a MedDRA term to mean euphoric mood. The term “BGIT” for bleeding from the gastrointestinal tract in the phrase “Not for anti-coagulation or antiplt for now in view of recent BGIT” was untagged although BGIT was in the AE gazetter. This was because “Not” led the sentence, and this was considered a negative description. Some AEs found by human annotators were just beyond the three sentence limit for searching for machine searching of AEs. Other AEs that the machine missed were abbreviations such as LOW for loss of weight, PR bleed for peri-rectal bleed, HHNK for hyperglycemic hyperosmolar nonketotic and certain phrases for abnormal laboratory values (e.g. elevated troponin, raised ESR), which were not contained in the AE gazetter. More medical abbreviations can be added to the AE gazetteer by use of other curated medical abbreviation lists. However we are mindful of the choice of abbreviations to be added to the AE gazetteer as the machine could wrongly detect the abbreviations without considering the context. Common laboratory tests that are indicative of AEs in phrases combined with adjacent words indicative of abnormal values can be implemented in revised machine algorithms. The distance between AE and drug may also need to be optimized.

The overall three-way IAA for drugs and AEs were 86% and 70% respectively. However the three-way IAA for drug-AE relationships was only 49% (Table 8).

4. Discussion

Evaluation of the performance of text mining algorithms for identification of potential AEs from discharge summaries requires a set of human annotated records that can serve as a benchmark or evaluation corpora [12]. To lay the foundation for creating such a benchmark, we initially evaluated two sets of 50 discharge summaries in order to develop a guideline for annotation of drugs, AEs and drug-AE relationships to achieve greater consistency among annotators. After developing the guideline, two additional sets of 50 discharge summaries were independently evaluated by six annotators, grouped into two groups of three annotators (Fig. 1). Both groups achieved high human IAA of ∼85% for drug entity tagging. The machine algorithm for drug entity extraction achieved a recall of 84% (Table 6) and precision of 73% meeting the pre-specified target goal of 80% for recall and 60% for precision. Annotation of plain records or machine pre-annotated records attained similar IAA, thus machine pre-annotation neither improved nor degraded the consistency of tagging. Machine annotation of each set of 50 records took less than three minutes compared to an average of 2 hours 16 minutes for each human annotator to tag plain records for drugs only (Table 5). The average time taken for each human annotator to review and correct machine pre-annotated records for drug entities was 2 hours and 3 minutes, which was not significantly less than the time taken to annotate plain records. By contrast, machine annotation took less than 3 minutes per set. Given the similarity of the recall of the machine algorithm and human IAA, it may be reasonable going forward to use machine annotated records for drugs without correction as the starting point for AEs and drug-AE tagging to reduce the tedium of creating a large benchmark set.

A limitation of using machine pre-annotated records for drug identification is that annotators only search for AEs within three sentences of a tagged drug. Errors from the machine annotation of drugs would be carried over in subsequent searches for AEs and drug-AE relationships. Bias also would be introduced in the evaluation of the performance of machine algorithms. However, these may be acceptable trade-offs for the benefit of expediting the creation of a benchmark set of discharge summaries. Further improvements in machine algorithms to increase recall and precision for drug entity extraction would reduce the impact of the errors and bias introduced from machine pre-annotations of drug entities.

The three-way IAA for AEs was considerably lower than for drugs at only 70% (Table 8). One possible reason could be because a drug is usually mentioned in a single word while AEs often occur as multiple descriptive terms. The recall and precision using the machine algorithm for AE annotation were only 67% and 53%, lower than the pre-specified targets. Since human annotators had only a modest IAA of 70% for identifying an AE, it was not surprising that the machine algorithm was also subpar. Future work needs to focus on developing clearer guidelines for human annotators for tagging AE entities to increase IAA. To increase recall and precision of machine algorithms, revisions as described under results (Section 3.2) can be implemented and tested. Until the machine algorithm achieves the desired targets for AEs tagging and the three-way IAA of drug-AE relationship identification is increased, it is not meaningful to evaluate the machine performance of drug-AE relationship identification.

The schematic of the evaluations conducted in this proof-of-principle exercise is shown in Fig. 1. Based on the results from these small trials of annotations, a reasonable approach to creating a benchmark set would be to use machine algorithms to pre-annotate drugs in the discharge summaries, and then have two human annotators tag for AEs and drug-AE relationships in proximity of the identified drugs. Evaluation of non-matches between human annotators and the current machine algorithm have provided additional insight into modifications of the algorithm to achieve higher recall and precision.

5. Future research directions

Classical natural language processing models are trained on well written text. Accuracy of tagging drops significantly with unstructured text, abundant abbreviations, misspellings and ungrammatical constructs. While our study showed the feasibility of our methods, more work is needed to establish a drug safety surveillance mechanism. Our plan for the future is to work on improved text mining algorithms for entity and relationship extraction.

Please cite this article in press as: P.S. Ang et al., Towards Human–Machine Collaboration in Creating an Evaluation Corpus for Adverse Drug Events in Discharge Summaries of Electronic Medical Records, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2016.04.001

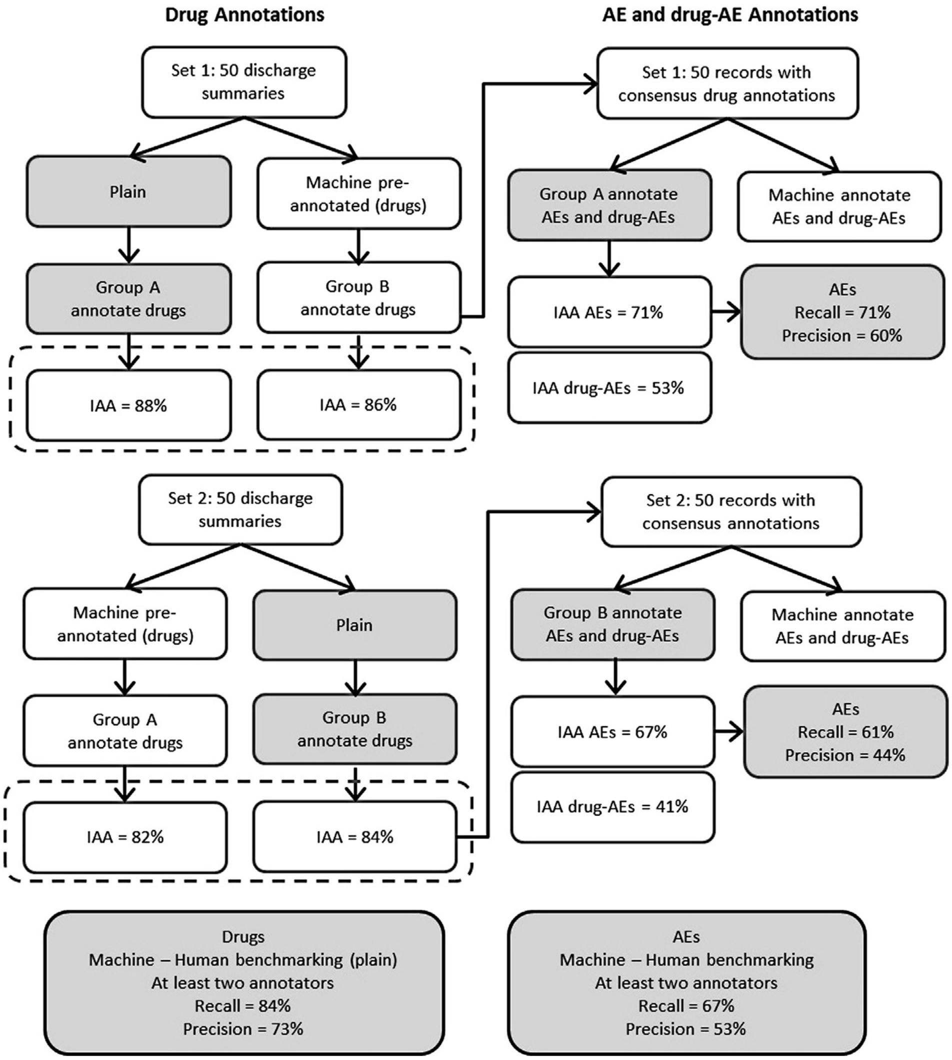
Drug Annotations
AE and drug-AE Annotations
Set 1: 50 discharge summaries
Plain
Machine pre-annotated (drugs)
Group A annotate drugs
Group B annotate drugs
IAA = 88%
IAA = 86%
Set 1: 50 records with consensus drug annotations
Group A annotate AEs and drug-AEs
Machine annotate AEs and drug-AEs
IAA AEs = 71%
AEs Recall = 71% Precision = 60%
IAA drug-AEs = 53%
Set 2: 50 discharge summaries
Machine pre-annotated (drugs)
Plain
Group A annotate drugs
Group B annotate drugs
IAA = 82%
IAA = 84%
Set 2: 50 records with consensus annotations
Group B annotate AEs and drug-AEs
Machine annotate AEs and drug-AEs
IAA AEs = 67%
AEs Recall = 61% Precision = 44%
IAA drug-AEs = 41%
Drugs Machine – Human benchmarking (plain) At least two annotators Recall = 84% Precision = 73%
AEs Machine – Human benchmarking At least two annotators Recall = 67% Precision = 53%

Fig. 1. Workflow of drug, AE and drug-AE relationship annotations.

1. Improve inter-annotator agreement by refining the guidelines on annotation.
2. Develop an evaluation corpus to benchmark the performance of the text mining algorithms.
3. Expand the reference gazetteer by extracting AE terms from approved drug labels and other medical abbreviation lists.
4. Investigate deep machine learning strategies, such as those recently described in the review articles [23–25], to iteratively improve recognition of patterns that are likely to be ADEs.
5. Evaluate algorithm performance on discharge summaries from other hospitals that have different EMR technology platforms and clinician practices for writing discharge summaries to assess the applicability of the algorithms in the broader health care system.
6. Explore text-mining strategies, such as frequent itemset mining, that would be suited for discovery of previously unknown drug-AE relationships.

Since machine pre-annotation did not cause a drop in the effectiveness of identifying AEs but can bring up the efficiency of AEs identification, our future effort will also emphasize the use of visualization to support human–machine collaboration. Potential drug-AE pairs that are discovered by a machine algorithm will be represented on a 3D plot in an interactive visual manner as shown in Fig. 2. The x-axis in the 3D plot represents the separation between drug and AE on the hospital summary that they are found, while the y and z-axes represent the strength of the trigger and negation phrases, respectively, that are associated with the drug-AE pairs in the summary. The size of the dot represents the number of hospital summaries that cluster together in similar region of the 3D plot. By providing this visual interface, reviewers can focus on the region of the plot that represents a high probability of containing valid drug-AE pairs and then retrieve the relevant hospital summaries for human review. This type of graphical visualisation can greatly enhance a human's ability to quickly spot potential drug-AE cases from a huge number of hospital discharge summaries.

Ultimately, we seek to obtain a more accurate estimate of the incidence and pattern of ADEs, as this would allow drug regulatory authorities and hospitals to prioritise interventions to reduce these events, such as through dosage adjustments or by avoiding certain drug combinations. As ADEs contribute significantly to morbidity and mortality in the health care system, progress towards successfully identifying and predicting risk factors for ADEs from discharge summaries would have a significant impact on public health.

Acknowledgements

The authors would like to thank Geraldine John, whose research for an MPH thesis prepared a foundation for conceptualisation of

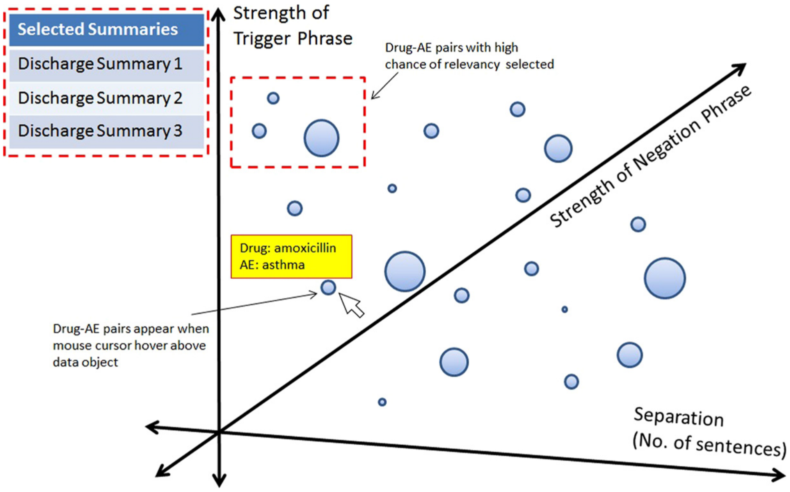
Selected Summaries
Discharge Summary 1
Discharge Summary 2
Discharge Summary 3
Strength of Trigger Phrase
Drug-AE pairs with high chance of relevancy selected
Strength of Negation Phrase
Drug: amoxicillin
AE: asthma
Drug-AE pairs appear when mouse cursor hover above data object
Separation (No. of sentences)

Fig. 2. An interactive visual approach to support human–machine collaboration.

the project, and Meiqun Hu for her contributions to the text mining algorithms.

Funding

The Biomedical Research Council funded this study through a grant for the Surveillance and Pharmacogenomics Initiative for Adverse Drug Reactions (BMRC SPF Ref No: SPF2014/001).

Ethics approval

The National Healthcare Group Domain Specific Review Board approved this study.

References

[1] M. Pirmohamed, S. James, S. Meakin, C. Green, A.K. Scott, T.J. Walley, K. Farrar, B.K. Park, A.M. Breckenridge, Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients, BMJ, Br. Med. J. 329 (7456) (2004) 15–19.

[2] Committee on safety of drugs, Report for 1969 and 1970. HMSO, London, 1971.

[3] W. Lim, Development of medical informatics in Singapore – keeping pace with healthcare challenges, in: Asia Pacific Association for Medical Informatics Meeting, Taipei, Taiwan, 2006.

[4] Y. Koh, A. Lim, L. Tan, P.S. Ang, S.H. Tan, D. Toh, C.L. Chan, in: Pharmacovigilance in Singapore – Harnessing IT and Genomics to Detect Safety Signals, Scrip Regulatory Affairs, 2012.

[5] L. Hazell, S.A. Shakir, Under-reporting of adverse drug reactions: a systematic review, Drug Safety 29 (5) (2006) 385–396.

[6] G. Melton, G. Hripcsak, Automated detection of adverse events using natural language processing of discharge summaries, J. Am. Med. Inform. Assoc. 12 (4) (2005) 448–557.

[7] M.Y. Park, D. Yoon, K. Lee, S.Y. Kang, I. Park, S.H. Lee, W. Kim, H.J. Kam, Y.H. Lee, J.H. Kim, R.W. Park, A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database, Pharmacoepidemiol. Drug Saf. 20 (6) (2011) 598–607.

[8] E. Ramirez, A.J. Carcas, A.M. Borobia, S.H. Lei, E. Piñana, S. Fudio, J. Frias, A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients, Clin. Pharmacol. Ther. 87 (1) (2010) 74–86.

[9] X. Wang, G. Hripcsak, M. Markatou, C. Friedman, Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study, J. Am. Med. Inform. Assoc. 16 (3) (2009) 328–337.

[10] P.C. Waller, S.J. Evans, A model for the future conduct of pharmacovigilance, Pharmacoepidemiol. Drug Saf. 12 (1) (2003) 17–29.

[11] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, JAMA 309 (13) (2013) 1351–1352.

[12] H. Gurulingappa, A. Mateen-Rajput, L. Toldo, Extraction of potential adverse drug events from medical case reports, J. Biomed. Semant. 3 (1) (2012) 15.

[13] L. Celi, E. Moseley, C. Moses, P. Ryan, M. Somai, D. Stone, K. Tang, From pharmacovigilance to clinical care optimisation, Big Data 2 (3) (2014) 134–141.

[14] R. Eriksson, P.B. Jensen, S. Frankild, L.J. Jensen, S. Brunak, Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text, J. Am. Med. Inform. Assoc. 20 (5) (2013) 947–953.

[15] WHO Adverse Reaction Terminology (WHO-ART), Uppsala Monitoring Centre, World Health Organization Collaborating Centre for International Drug Monitoring http://www.umc-products.com/DynPage.aspx?id=73589&mn1=1107&mn2=1664.

[16] Medical Dictionary for Regulatory Activities (MedDRA) Terminology, International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

[17] J.C. Reynar, Topic Segmentation: Algorithms and Applications, Computer and Information Science, University of Pennsylvania, 1998.

[18] Y.Y. Choi, Advances in domain independent linear text segmentation, in: Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, http://www.aclweb.org/anthology/A00-2004.

[19] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310.

[20] Knowtator. Mozilla Public License Version 1.1 http://knowtator.sourceforge.net/.

[21] G. Navarro, A guided tour to approximate string matching, ACM Comput. Surv. 33 (1) (2001) 31–88.

[22] R.A. Wagner, M.J. Fischer, The string-to-string correction problem, J. ACM 21 (1) (1974) 168–173.

[23] T. Huang, L. Lan, X. Fang, P. An, J. Min, F. Wang, Promises and challenges of big data computing in health sciences, Big Data Res. 2 (1) (2015) 2–11.

[24] H. Gurulingappa, J. Fluck, M. Hofmann-Apitius, L. Toldo, Identification of adverse drug event assertive sentences in medical case reports, in: The First International Workshop on Knowledge Discovery in Health Care and Medicine, KDHCM'11, Online Proceedings: Athens, Greece, September 9, 2011.

[25] S. Visweswaran, P. Hanbury, M. Saul, G. Cooper, Detecting adverse drug events in discharge summaries using variations on the simple Bayes model, in: AMIA Annu. Symp. Proc., Washington, DC, 2003.