Predicting the presence of drug-adverse event pairs in discharge summaries

Tan Hui Xing[1], Tham Mun Yee[1], Ang Pei San[1], Belinda Foo[1], Sally Soh[1], Desmond Teo[1], Tan Siew Har[1], Tang Yixuan[2], Yang Jisong[2], Ling Zheng Jye[3]
Anthony Tung[2], Cynthia Sung[1,4], Sreemanee Raaj Dorajoo[1]

[1]Vigilance & Compliance Branch, Health Sciences Authority, [2]Department of Computer Science, School of Computing, National University of Singapore, [3]National University Health System, [4]Health Services and Systems Research, Duke-NUS Medical School

For more information, please contact: sreemanee_dorajoo@hsa.gov.sg

OBJECTIVE

To compare rule-based versus machine learning algorithms in their abilities to detect drug-adverse event (AE) pairs as documented in discharge summaries, as a means of enhancing post-market surveillance of approved medications.

INTRODUCTION

Hospital discharge summaries offer a potentially rich resource to enhance pharmacovigilance efforts to evaluate drug safety in real-world clinical practice. However, it is infeasible for experts to read through all discharge summaries to find cases of drug-adverse event (AE) relations.[1]

This work presents a comparison of our previously published rule-based algorithm, named REAP (**Re**adpeer for **A**ctive **P**harmaco-vigilance), against a novel machine learning approach to automatically extract segments of text that contain drug-AE relationships.[2]

METHODOLOGY

Rule-based algorithm development

- NLP pipeline developed to extract drug and AE names based on a list of customized dictionaries, fuzzy logic (including Soundex) and negation detection (Fig.1)
- A set of expert-derived rules based on specific trigger phrases are carefully designed to identify candidate drug-AE pairs (Fig. 2)
- The customised Readpeer interface allows pharmaco-vigilance (PV) experts to annotate and label the rule-based algorithm output

Machine learning algorithm development

- Using 90% of the annotated data (n=1692), we built models and tested the best performing ones on the remaining 10% (n=188) as a form of validation.
- Term-frequency-inverse document frequency (TF-IDF) and word2vec were used to vectorize the text before training the models using k-nearest neighbour (kNN), Naïve-Bayes (NB), Stochastic Gradient Descent (SGD) and Random Forest (RF) algorithms.

METHODOLOGY

Figure 1. Workflow for rule-based algorithm development

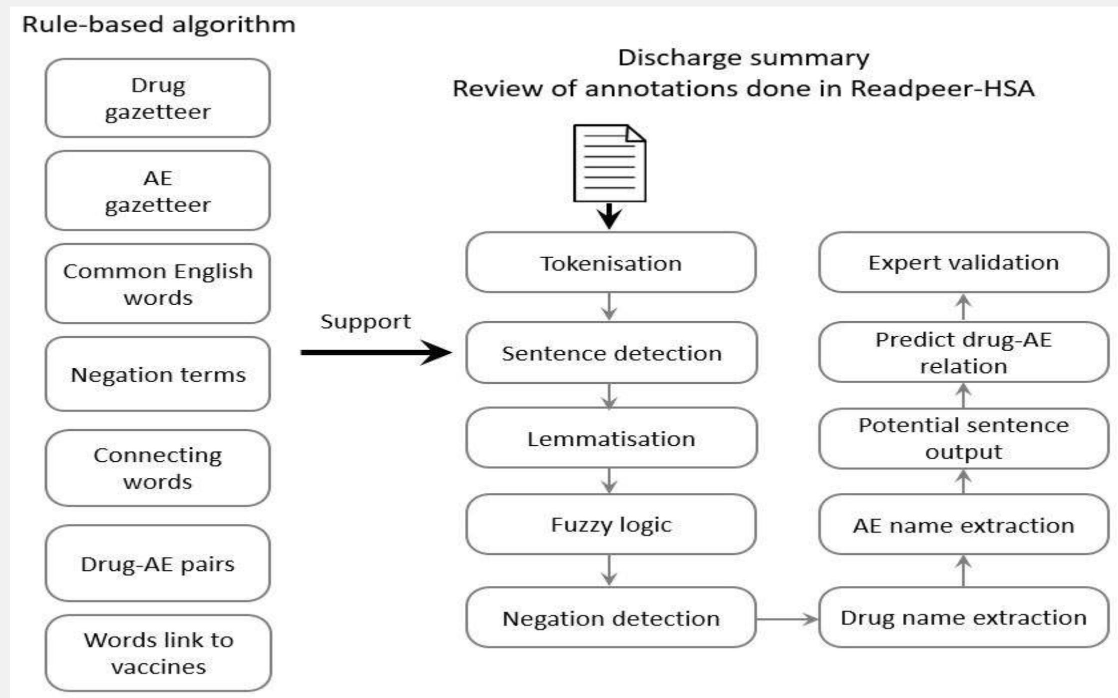
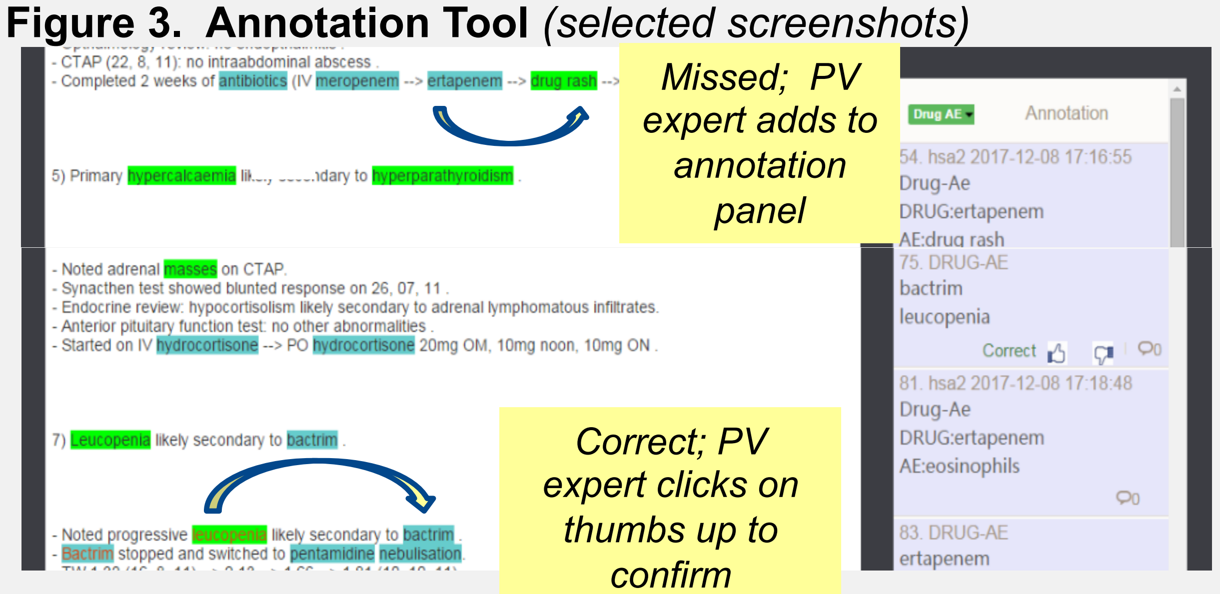


Figure 2. Selected Drug-AE Relationship Rules

No	Relation Rule Group	Phrase Set	Examples
1	Drug *Cause* AE	Cause: {caused, induced, resulted in, …}	Isoniazid induced DILI
2	AE *AttributeTo* drug	AttributeTo: {attributed to, due to, secondary to…}	Hypoglycemia due to glimepiride
3	*AllergyTo* drug	AllergyTo: {da to, allergic to, …}	Allergic to penicillin
4	Drug *StopAfter* AE, word distance (drug,AE) < 12	StopAfter: {stop, held off, discontinued,…}	Simvastatin discontinued after leg muscles became painful
5	…	…	…

Figure 3. Annotation Tool *(selected screenshots)*



RESULTS & DISCUSSION

Optimal vectorization methods prior to machine learning

Training Phase (n=1692)			
Vectorization method	Average Precision	Average Recall	Average F-score
TF-IDF	0.778	0.704	0.738
Word2vec	**0.840**	**0.718**	**0.772**

Word2vec word embeddings generated models with a higher average precision and recall compared to TF-IDF. Therefore, all validation phase models were built using word2vec.

Optimal vectorization methods prior to machine learning

Validation Phase (n=188)			
	Precision	Recall	F-score
Rule-based algorithm	*0.757*	*0.586*	*0.661*
k-Nearest neighbour	0.780	0.780	0.780
Naïve Bayes	0.820	0.690	0.750
Stochastic Gradient Descent	0.820	0.660	0.740
Random Forest	**0.830**	**0.750**	**0.790**

CONCLUSION

- Machine learning approaches appear to be better at detecting drug-AE pairs in discharge summaries than expert-derived rule-based algorithm.

ACKNOWLEDGEMENTS

This work was supported by a Strategic Positioning Fund grant from the Biomedical Research Council of the Agency for Science, Technology and Research of Singapore (SPF2014/001).

SELECTED REFERENCES

1.G.B. Melton et al. Automated detection of adverse events using natural language processing of discharge summaries. JAMIA **12** (4) (2005) 448-45.
2.T Yixuan et al. Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer *IJMI* **128** (2019) 62-70.