Eureka – Finding a way to harvest the data in the medicinal product information

Luis Pinheiro, PharmD, MEpi, Jan Kors, PhD, Peter Rijnbeek, PhD
Erasmus University Medical Center, The Netherlands

Abstract

The product information (PI) of medicines holds valuable but unstructured information on indications and adverse drug reactions (ADRs). A database of ADRs for Centrally Authorised Products (CAPs) in the European Union exists but was manually created and curated.

An automated method to recognize and extract indications and ADRs from PIs in PDF format was developed, titled Eureka. The method uses symbolic natural language processing (NLP) and a bespoke dictionary of terms.

Performance was assessed for the ADR data extraction using the manually curated ADR database as a silver standard. The macro-averaged performance was F1 0.79, Precision 0.81 and Recall 0.77. The weighed results were: F1 0.81, Precision 0.83 and Recall 0.79.

This performance is on par with the performance of the best performing methods that have been published. Eureka holds the promise to open this data to safety monitoring and other research pipelines such as for the identification of negative controls using OHDSI's Common Evidence Model or in predictive modelling.

Research Category (please highlight or circle which category best describes your research)

Data management, data standardization

Introduction

The product information *(*PI) of medicines informs on the effective and safe use of a medicinal product. These text documents hold valuable but unstructured information, such as the indications and the adverse drug reactions (ADRs).

A research-ready database of indications and ADRs has many use cases including safety monitoring, identification of candidate predictors and clinically relevant prediction questions in predictive modelling, selection of negative controls in population level estimation, etc.

A database of ADRs for Centrally Authorised Products (CAPs) in the EU already exists. (1) It was built and curated manually making it less efficient, error-prone, hard to scale, and leading to latency problems. In fact, the last update was in 2017, and includes withdrawn products but not several hundred new products that have been authorised since then.

Others have tried to extract ADR and/or indication information from PIs (2–12), but only one used PIs in PDF format (10) and none addressed the EU PI.

The EU PI, also referred to as Summary of Product Characteristics (SmPC), is supported by a guideline (13) that states that ADRs should be coded using the Medical Dictionary for Regulatory Activities (MedDRA). (14) The specific hierarchies to be used are High Level Terms (HLTs), Preferred Terms (PTs) and Lower Level Terms (LLTs).

In this study, we aim to develop and validate an automated method for the extraction and MedDRA coding of ADRs and indications from EU SmPCs in PDF format.

Methods

An R-package called Eureka is developed that extracts MedDRA terms from sections 4.1 and 4.8 of the

SmPC in PDF format. Entity recognition and extraction was performed for all 1147 EU CAP SmPCs in PDF format. The performance for ADR data was assessed on 886 medicinal products which were recorded in the manually curated ADR database (used as a silver standard).

A dictionary, based on MedDRA HLTs, PTs and LLTs was manually created by excluding terms that do not refer to either indication or adverse drug reactions, e.g. “married” or “glucose”. This resulted in a dictionary containing 54635 terms.

Identification of indications and ADRs was performed using a symbolic NLP method. Terms in the dictionary were transformed to regular expressions which were used to detect the presence of a term in the text. Sentence tokenizing was performed using a spaCy (15) pre-trained model. Specific code was developed to address specific concerns such as terms that are conjugated, sentences that refer to underlying medical history, lexical variations of terms in MedDRA, among others.

Macro-averaged and weighed (by count of terms) F1-score, Precision and Recall were reported. False positives and false negatives were profiled.

Parsing and reading text data, entity recognition and extraction and reporting are done in R. Highlighting the extracted adverse drug reactions in the PDF is done in Python.

Results and Discussion

The macro-averaged performance was F1 0.79, Precision 0.81 and Recall 0.77. The weighed results were: F1 0.81, Precision 0.83 and Recall 0.79.

Compared with the performance of other published approaches in the past five years, where the performance was validated with only 100 PI documents, the macro-averaged F1-score is on par with the best performing method, by Pandey et al. (12)

This performance of our algorithm is very acceptable considering that using PDF files, instead of XML/HTML as done in the other algorithms, has a performance cost. Furthermore, profiling of the false negatives and positives indicates that the ADR database includes typos, roughly 6% non-MedDRA terms and several missing terms. This demonstrates the value of including Eureka in the process.

An additional advantage of Eureka is that it maps non-MedDRA terms identified by the regular expressions to standard MedDRA terms.

Symbolic NLP methods have proved at least as capable as statistical methods in extracting indications and ADRs from PIs, but complete and fully correct data requires a human-in-the-loop. Development of Eureka will require two applications: one that corrects and completes the extraction (Eureka Extract) and a browser of the data, offering analytical opportunities (Eureka Explore).

Conclusion

Eureka’s performance is on par with the best performing methods. In the upcoming months we will make further refinements of the algorithm and will make the Eureka Extract and Eureka Explore applications available for the OHDSI community.

References

1. PROTECT - Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium [Internet]. [cited 2015 Jun 19]. Available from: http://www.imi-protect.eu/about.shtml
2. Kusch MKP, Zien A, Hachenberg C, Haefeli WE, Seidling HM. Information on adverse drug reactions—Proof of principle for a structured database that allows customization of drug information. Int J Med Inform [Internet]. 2020;133(June 2019):103970. Available from: https://doi.org/10.1016/j.ijmedinf.2019.103970
3. Duke JD, Friedlin J. ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. AMIA Annu Symp Proc. 2010;2010:177–81.
4. Tiftikci M, Özgür A, He Y, Hur J. Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels. BMC Bioinformatics [Internet]. 2019;20(Suppl 21):1–9. Available from: http://dx.doi.org/10.1186/s12859-019-3195-5
5. Li Q, Deleger L, Lingren T, Zhai H, Kaiser M, Stoutenborough L, et al. Mining FDA drug labels for medical conditions. BMC Med Inform Decis Mak. 2013;13(1).
6. Fung KW, Jao CS, Demner-Fushman D. Extracting drug indication information from structured product labels using natural language processing. J Am Med Informatics Assoc. 2013;20(3):482–8.
7. Culbertson A, Fiszman M, Shin D, Rindflesch TC. Semantic processing to identify adverse drug event information from black box warnings. AMIA Annu Symp Proc. 2014;2014(6):442–8.
8. Khare R, Li J, Lu Z. LabeledIn: Cataloging labeled indications for human drugs. J Biomed Inform [Internet]. 2014;52:448–56. Available from: http://dx.doi.org/10.1016/j.jbi.2014.08.004
9. Khare R, Wei CH, Lu Z. Automatic extraction of drug indications from FDA drug labels. AMIA Annu Symp Proc. 2014;2014(i):787–94.
10. Lamy JB, Ugon A, Berthelot H. Automatic extraction of drug adverse effects from product characteristics (SPCs): A text versus table comparison. Stud Health Technol Inform. 2017;228(2):339–43.
11. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, et al. Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology. J Biomed Inform [Internet]. 2018;83(December 2017):73–86. Available from: https://doi.org/10.1016/j.jbi.2018.05.019
12. Pandey A, Kreimeyer K, Foster M, Dang O, Ly T, Wang W, et al. Adverse Event extraction from Structured Product Labels using the Event-based Text-mining of Health Electronic Records (ETHER) system. Health Informatics J. 2019;25(4):1232–43.
13. European Commission. A Guideline on Summary of Product Characterisitics [Internet]. 2009. Available from: https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf
14. MedDRA | Medical Dictionary for Regulatory Activities.
15. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017;