Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-339

Automatic Extraction of Drug Adverse Effects from Product Characteristics (SPCs): A Text Versus Table Comparison

Jean-Baptiste LAMY[1], Adrien UGON and Hélène BERTHELOT
LIMICS, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny, France,
INSERM UMRS 1142, UPMC Université Paris 6, Sorbonne Universités, Paris

Abstract. Background: Potential adverse effects (AEs) of drugs are described in their summary of product characteristics (SPCs), a textual document. Automatic extraction of AEs from SPCs is useful for detecting AEs and for building drug databases. However, this task is difficult because each AE is associated with a frequency that must be extracted and the presentation of AEs in SPCs is heterogeneous, consisting of plain text and tables in many different formats. Methods: We propose a taxonomy for the presentation of AEs in SPCs. We set up natural language processing (NLP) and table parsing methods for extracting AEs from texts and tables of any format, and evaluate them on 10 SPCs. Results: Automatic extraction performed better on tables than on texts. Conclusion: Tables should be recommended for the presentation of the AEs section of the SPCs.

Keywords. Adverse Effects, Natural Language Processing, SPCs

1. Introduction

Drugs can greatly improve patient health, but sometimes cause adverse effects (AEs). The potential AEs that a given drug can cause are described in the Summary of Product Characteristics (SPCs) of the drug. AEs are included in drug databases that vary widely in their quality and their ability to address specific questions [1]. There are very few evaluations of drug databases and most compared several databases between themselves, but not to a *gold standard* [2]. Resources in languages other than English, such as French, are also very limited. It is thus often preferable to refer to the original SPC rather than to rely on a database.

SPCs can be read by experts but also analyzed by computers, using natural language processing (NLP) methods [3]. However, extracting AEs from SPCs is complex for two reasons: (1) the nature of the AEs must be extracted, typically using MedDRA terms (Medical Dictionary for Regulatory Activities), as well as their associated frequencies (*i.e.* very rare, rare, infrequent, frequent, very frequent or unknown) and System Organ Class (SOC), which is required to distinguish some terms (*e.g.* MedDRA has two "vertigo" terms, in two different SOCs); (2) the presentation of AEs is extremely heterogeneous (see proposed taxonomy in Figure 1). Some SPCs describe AEs using plain or semi-structured text, whereas others use tables or a

[1] Corresponding Author: jean-baptiste.lamy@univ-paris13.fr.

combination of both. Some are in HTML and others in PDF format. There is no standard table format, but we distinguished two main formats: *table with per-row frequencies* (*i.e.* the frequency is found in the same row as the AE) and *table with per-column frequencies* (the frequency is found in the same column as the AE). Each category has many variants: with SOCs displayed as full-width rows or in an additional row or column with some vertical and/or horizontal lines omitted.

We describe NLP methods for extracting AEs from SPCs and compare extraction from tables with extraction from text. We conclude by providing recommendations for the presentation of the AE section in SPCs.

Unstructured text :

This drug may cause vascular disorders. Rare cases of thrombophlebitis and very rare cases of venous thrombosis have been observed when taking this drug.

Semi-structured text :

Vascular disorders

Rare :	Thrombophlebitis
Very rare :	Venous thrombosis

Tables (per-row frequencies) :

SOC	AE	Frequency
Vascular disorders	Thrombophlebitis	Rare
	Venous thrombosis	Very rare

AE	Frequency
Vascular disorders	
Thrombophlebitis	Rare
Venous thrombosis	Very rare

Frequency	AE
Vascular disorders	
Rare	Thrombophlebitis
Very rare	Venous thrombosis

Tables (per-column frequencies) :

SOC	Very rare	Rare	Frequent
Vascular disorders	Venous thrombosis	Thrombophlebitis	

Very rare	Rare	Frequent
Vascular disorders		
Venous thrombosis	Thrombophlebitis	

Figure 1. Taxonomy of the various presentations of AEs in SPCs. The same two AEs are displayed in four different presentations, with several variants of the tables.

2. Materials and methods

2.1. Materials

We selected all drugs from the Thériaque French drug database and removed generic drugs, withdrawn drugs, and drugs available only in the hospital. We kept a single dosage for each brand name. We obtained 2607 drugs and downloaded their SPCs from the French repository (*base de données publique des médicaments*, public drug database, http://base-donnees-publique.medicaments.gouv.fr/). Among these, 297 were in PDF and 2310 in HTML formats.

2.2. Natural language processing and indexing

We set up a MedDRA indexing chain, using MedDRA 17.1 (French translation) with PyMedTermino [4], the French version of the SnowBall lemmatiser from the NLTK Python module (http://www.nltk.org/) and the Enchant spellchecker

(https://github.com/AbiWord/enchant) enriched with the list of all words present in MedDRA. We extended the indexer with a list of 201 synonyms (*e.g.* "of mouth" is a synonym of "oral") determined manually during preliminary work.

Many sentences (or parts of sentences) in plain text AE descriptions include MedDRA terms, but these terms are not used to refer to drug AEs but rather to patient comorbidities (*e.g.* "diabetes" in "patients with diabetes are at risk of <AE>"), the drug indication (*e.g.* "in the indication of hypertension, this drug may cause <AE>"), or they described AEs limited to a specific population (*e.g.* pediatric population). We thus defined a set of *negative patterns* to remove such sentences. The following patterns were used: "treatment of...", "indication of...", "patient suffering from...", "patient with...", "on...", "in..." (translated from French).

2.3. Table extraction

Extracting tables from HTML files is easy, but not from PDF files. We used a modified version of Pdf-table-extract (https://github.com/ashima/pdf-table-extract). This tool relies on shape-recognition methods for finding vertical and horizontal lines in PDF files and then determines the intersection of the lines and the cell coordinates. The text of each cell can then be extracted. We extended Pdf-table-extract to (a) detect vertical/horizontal white spaces as row/column separators (mandatory because some lines are often omitted in SPCs), and (b) generate pseudo-HTML as output. The pseudo-HTML produced can then be parsed as plain HTML SPCs.

We then wrote a specific table-parsing algorithm following these steps: (1) index the table's content for finding all SOCs (*e.g.* vascular disorders), AEs (*e.g.* thrombophlebitis), and frequency terms (*e.g.* rare); (2) determine whether the table has per-row frequencies (if we found more than one frequency term in a given column) or per-column frequencies (otherwise); (3) associate a SOC and a frequency with each AE; the SOC is the first one located before the AE (in a row-first order) and the frequency is given by the row (per-row frequencies table) or the column (otherwise).

Table 1. The distribution of the 2607 SPCs according to the presentation used for describing AEs. The last column shows the mean number of AEs per SPC in each category.

	Number of SPCs		**AEs per SPC**
AEs described as text	1662	(63.7%)	24.3
AEs described in a table, including:	739	(28.3%)	52.6
table (per-row frequencies)	373	(14.3%)	51.1
table (per-column frequencies)	366	(14.0%)	54.1
No AEs found	196	(7.5%)	0.0
Non-analyzable SPCs	10	(0.4%)	-
Total	**2607**	**(100%)**	**30.5**

Table 2. Results of the evaluation of 10 SPCs. The expected number of AEs (according to the expert), the precision, recall and F-measure, and the percentage of AE frequencies that were correctly extracted for each presentation of AEs are shown.

	Expected	**Precision**	**Recall**	**F**	**Freq.**
AEs described as text	69	25%	38%	0.30	54%
AEs described in a table, including:	297	87%	81%	0.84	97%
table (per-row frequencies)	145	86%	84%	0.85	95%
table (per-column frequencies)	152	89%	78%	0.83	99%

2.4. Evaluation methods

The indexer was run on the 2607 SPCs to quantify the proportion of each AE presentation format. Ten SPCs were randomly chosen: four with text, three with per-row frequency tables and three with per-column frequency tables. An expert pharmacist (HB) manually extracted the AEs from the SPCs and coded them in MedDRA. The AEs extracted by the indexer were compared to those extracted by the expert. We considered only the AEs presented in tables for SPCs describing AEs in tables, as recent AEs are sometimes added in the text but the table is not updated. Many orthographic variants and synonyms are present in MedDRA. The comparison was thus performed at the PT level (Preferred Term). Finally, we computed recall, precision and F-measure using usual formula and considering the expert coding as a gold standard. We also computed the percentage of frequencies correctly extracted.

3. Results

With the indexer, we classified the 2607 SPCs according to the format used to present AEs (Table 1). AEs are described using text in 64% of SPCs and with tables in 28%. Approximately half of the tables have per-row frequencies and the other half per-column frequencies. Thus, there is no preferred table format. The mean number of AEs is higher in SPCs using tables, *i.e.* tables tend to be preferred for drugs with many AEs.

We evaluated the indexer on 10 SPCs (Cetrotide®, Panretin®, Niflugel®, Kalinox®, Picato®, Comtan®, Optiray®, Plasmalyte Viaflo®, Circadin®, Permixon®) (Table 2). The indexer gave good results for tables (F-measure = 0.84), but the results were poor for plain text (F = 0.30). The same difference was observed for the determination of the frequencies associated with the AEs: with tables, the extraction of the frequencies was almost perfect (97%), whereas it was less good with texts (54%). Manual analysis of the errors showed that the list of synonyms in MedDRA was not exhaustive, even after we completed it manually. Additionally, some tables were not well-formed in the HTML files (*e.g.* inconsistent use of the colspan HTML attribute). Many problems were encountered with text. Several very general terms were recognized as AEs whereas they were sometimes not (*e.g.* "gas" was interpreted as flatulence because it corresponds to a MedDRA LLT, but may also refer to an aerosol drug). Some recent SPCs have very long texts describing AEs and clinical trials. It was not easy to determine whether the described effects were simple observations during the trial or can be generalized. AEs were difficult to distinguish from indications in some sentences (*e.g.* in "For gouty arthritis, the adverse events rate was 0.2%", it is not clear whether gouty arthritis is an AE or a potential indication of the drug). The last two problems were also encountered by the expert, although she was able to solve them.

4. Discussion and conclusion

Here, we proposed a taxonomy for the presentation of AEs in SPCs and designed a specific indexer for extracting AEs from text and tables in various formats. We show that automatic extraction performed much better with tables than text. Tables were also easier to analyze by experts. The indexer we used was simple. It could be improved with additional terminologies and more complex NLP methods. However, it is

improbable that it would change the text *vs* table performance ratio. A study showed that simple dictionary-based NLP methods were as efficient as more complex methods for many medical applications [5]. In the literature, NLP has been widely applied to clinical narratives [6] but not to SPCs. F-MTI [7] achieved good performance for AE extraction (precision 77.0%, recall 59.4%, mixing texts and tables) but without extracting the frequencies. Better results were obtained for the contraindications [8] and interactions [9] sections of the SPC, possibly because the presentation is less variable than that of AEs. European guidelines for SPCs [10] do not provide guidance for presenting of AEs. The HL7 Structured Product Labeling standard [11] recommends using SNOMEDCT but has no guidance for tables. Based on this study and the expertise we acquired during research projects, we recommend: (1) using tables for presenting AEs in SPCs; any format can be used as there is no common format and it is possible to design a table-parsing algorithm that accepts all formats, (2) when the table spreads over several pages, repeat headers, (3) systematically updating the tables describing AEs when new AEs are added to the SPCs (currently, new AEs are often added in text only), (4) adding tables to SPCs of old drugs that do not have AEs presented in a table.

In conclusion, our indexing tool can be used for semi-automatic extraction of AEs presented in tables. Future work should focus on better evaluation and comparison of the automatic extraction results to existing drug databases.

Acknowledgement

This work was funded by the VIIIP project of the French drug agency (ANSM, *Agence Nationale de Sécurité du Médicament et des produits de santé*, AAP-2012-013).

References

[1] Clauson KA, Marsh WA, Polen HH, Seamon MJ, Ortiz BI. Clinical decision support tools: analysis of online drug information databases. *BMC medical informatics and decision making*. 2007;7:7.

[2] Biarez O, Sarrut B, Doreau CG, Etienne J. Comparison and evaluation of nine bibliographic databases concerning adverse drug reactions. DICP : *the annals of pharmacotherapy*. 1991;**25**(10):1062–5.

[3] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;**18**(5):544–51.

[4] Lamy JB, Venot A, Duclos C. PyMedTermino: an open-source generic API for advanced terminology services. *Stud Health Technol Inform*. 2015;210:924–928.

[5] Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *J Am Med Inform Assoc*. 2015;**22**(1):121–31.

[6] Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*. 2012;**3**(1):15.

[7] Pereira S, Plaisantin B, Korchia M, Rozanes N, Serrot E, Joubert M, et al. Automatic construction of dictionaries, application to product characteristics indexing. *Stud Health Technol Inform*. 2009;150:512–6.

[8] Rubrichi S, Quaglini S, Spengler A, Russo P, Gallinari P. A system for the extraction and representation of summary of product characteristics content. *Artif Intell Med*. 2013;**57**(2):145–54.

[9] Rubrichi S, Quaglini S. Summary of Product Characteristics content extraction for a safe drugs usage. *J Biomed Inform.* 2012;**45**(2):231–9.

[10] European Commission (EC). Guideline on the readability of the labelling and package leaflet of medicinal products for human use; 2009.

[11] Structured Product Labeling. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=96