Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth A. Ugon et al. (Eds.) © 2018 European Federation for Medical Informatics (EFMI) and IOS Press. This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0). doi:10.3233/978-1-61499-852-5-421

## Signal Detection for Baclofen in Web Forums: A Preliminary Study

# Pierre KARAPETIANTZ<sup>a,1</sup>, Bissan AUDEH<sup>a</sup>, Agnès LILLO-LE LOUËT<sup>b</sup> and Cedric BOUSQUET<sup>a</sup>

<sup>a</sup>Sorbonne Université, Inserm, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, France <sup>b</sup>Centre régional de pharmacovigilance HEGP, AP-HP Paris, France

Abstract. Web forums are proposed as a new complementary source of knowledge to spontaneous reports by patients and healthcare professionals due to underreporting of adverse drug reactions (ADRs). Some authors suggest that signal detection could be a convenient method for gathering mentions of ADRs in patients' posts. Signal detection methods were proposed to mine pharmacovigilance databases, but little is known about their applicability to web forums. We describe a method implementing several traditional decision rules on signal detection with baclofen applied to a set of more than 6 million posts. We then cross-validated four unexpected signals applying a logistic regression method. Most adverse effects (AEs) described in the summary of product characteristics of baclofen were detected by signal detection methods. Some unexpected AEs were too. Therefore, web forums are confirmed as a complementary resource for improving current knowledge in pharmacovigilance by detecting unexpected adverse drug reactions.

Keywords. Baclofen, Adverse drug reaction, Social media, Forums, Signal detection, Logistic regression

## Introduction

The Internet and social media have become part of people's daily lives. In 2016, 86% of the French population was reported as active internet users, and 58% of these users were active on social media [1]. On average, users spend 3h 37m daily on the internet, and 1h 16m on social media [1]. In France, 46% to 71% of the population uses the Internet to seek medical or health related information [2].

Web forums are online platforms used to share information and personal experiences. Health data (diseases, feelings, medication usage and many other aspects) in these forums, represents a rich source of information on diagnoses, treatments and medical researches [3]. Several studies have already demonstrated the added value of mining adverse drug reaction (ADR) signals from social media posts [4], using machine-learning methods (such as support vector machine or associations rules) but little is known about the utility of applying traditional signal detection methods (used to mine pharmacovigilance databases) as well as logistic regression signal detection method in web forums analysis. To our knowledge, only two studies applied one of the traditional signal detection method to web-based pharmacovigilance data [5, 6].

<sup>&</sup>lt;sup>1</sup> Corresponding Author; E-mail: pierre.karapetiantz@gmail.com.

Pharmacovigilance - defined by the World Health Organization (WHO) as "the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problems" [7] - begins during clinical trials and continues after the drug is released onto the market. ADRs may be spontaneous reports (from healthcare professionals or patients), or adverse events (AEs) observed during Phase IV clinical trials driven by pharmaceutical companies and national authorities. All data are stored in spontaneous reporting systems [8].

In Europe, national authorities accept reports from patients since 2011. In France, patients as well as any healthcare professional can send ADRs reports to the corresponding Regional Pharmacovigilance Centre which collects, analyzes and reviews these reports before adding them to the national agency. However, underreporting of ADRs is one of the biggest limitations of the current pharmacovigilance systems. Seven studies have revealed that the underreporting rate in France is between 78% and 99% [9]. To overcome underreporting, web forums were proposed as a potential source of information complementary to case reports.

Our objective was to evaluate the feasibility of ADRs' signal detection using web forums. In this preliminary work, we studied baclofen which has been used since the mid-70s in the management of spasticity in several neurological diseases. Indeed, the off-label high-dosed use of this drug by alcohol-dependent patients has caught the attention of the public health authorities in France in the last 5 years. Recently, signals emerging from patients and health professionals made this drug an ideal subject for such analysis on social media. Thus, we performed a retrospective, observational and descriptive study, based on posts mentioning AEs caused by baclofen, detected by natural language processing.

#### Methods

Our material consisted of a dataset of over 60 million posts extracted from 22 French medical forums from 01/01/2000 to 07/03/2015 using the open source tool Vigi4Med Scrapper [10]. Natural language processing (NLP) techniques were applied to these posts [11]. As causality between drugs and events was not implemented, adverse events (AEs) rather than ADRs were detected. This allowed detecting the mention of at least one drug and one AE in 6,569,555 posts, which corresponded to 55,350,564 drugs-AE couples.

To improve precision, all the mentions of drugs and AEs detected by NLP were normalized. For drugs, the normalization was achieved with a perfect correspondence to the RacinePharma dictionary (5,164 drug terms) linked by the Vidal company to its own drug dictionary (3,971 drug terms) for 83.5% (4,310 drug terms), and to the Anatomical Therapeutic Chemical (ATC) classification system. Due to the difference between labels of drug terms in medical dictionaries and drug mentions in the patients' posts, we added to the normalization a set of different lexical variations that patients could use in web forums to mention baclofen.

Adverse events normalization was achieved using the UMLS-MedDRA (V19.1) mappings within the UMLS metathesaurus. When a UMLS term was linked to more than one MedDRA term we considered all possible couples. Terms corresponding to the System Organ Class (SOC) "Social circumstances" and to High Level Group Terms (HLGT) containing the character string: "therapeutic procedure" were removed. Only drugs-AE couples that were coded in 5th ATC level and preferred term level were taken into account.

After the normalization process, we compared the AEs' SOC mentioned with baclofen in web forums to the SOC found in the Summary of Product Characteristics (SPC). We then applied several decision rules for traditional signal detection methods used in Pharmacovigilance - frequentist and bayesian methods, with or not application of the False Discovery Rate (FDR) [12]. The applied decision rules are described in table 1.

Method		Decision rules				
Frequentist Methods						
Proportional Reporting Ratio	-	PRR1: $n \ge 3, PRR >$				
(PRR)		2 and $\chi^2 > 4$				
		PRR2: $CI_{95\%}(ln(PRR)) >$				
		0, with FDR < 5%				
Reporting Odds Ratio (ROR)	-	$ROR1: CI_{95\%}(ln(ROR)) > 0$				
	-	ROR2: ROR1, with FDR <				
		5%				
Reporting Fisher's Exact Test	-	Fisher's Exact Test p-value				
(RFET)		calculated with 2 methods				
		(RFET and midRFET), with				
		FDR < 5%				
Bayesian Methods						
Bayesian Confidence	-	BCPNN1: Beta distribution a				
Propagation Neural Network		priori and				
(BCPNN)		$Information \ Component >$				
		0				
	-	BCPNN2:				
		BCPNN1, with FDR < 5%				
Gamma Poisson	-	$GPS1: CI_{95\%}(ln(EBGM)) \ge$				
Shrinkage (GPS)		2				
	-	GPS2: GPS1, with FDR <				
		5%				

Table 1. Decision rules applied for each signal detection algorithm

As final step, to cross-check the previous signal detection algorithm, we applied a logistic regression based method: the Class-imbalanced subsampling lasso algorithm (CISL) with a 5% threshold [11], considering each post as an Individual Case Safety Report about four AEs of interest selected by a pharmacist (CB). All statistical analyses were performed with the R language and environment for statistical computing and graphics [14] using the PhVID package.

#### Results

From the 55,350,564 detected drug-AE couples, 5,997 had an exact match with "BACLOFEN". When considering other lexical variations for baclofen like "BAKLO", "BALCO", "BACOLFEN", 40,158 additional couples were identified as relevant.

After the normalization process, the final database consisted of 2,622,445 drug-AE couples from 747,500 different posts from 07/02/2000 to 07/02/2015. 999 distinct AEs were found with baclofen. Among the 10 most frequent AEs detected with baclofen by the NLP tool, two were indications for baclofen's use and represented around 15% of the

AEs found with baclofen. The 10 most frequent AEs found with baclofen represented 54% of baclofen's AEs in web forums. Furthermore, 36 of the 37 AEs present in baclofen' SPC were also detected in association with this drug in the forums. In baclofen' SPC, 30% of the effects are related to the "Nervous system disorders" SOC and 16% to the "Psychiatric disorders" SOC whereas, in the forums, 10% of the AEs were related to "Nervous system disorders". The repartition of adverse events between different organs in the forum is thus different than what is described in the SPC. Table 2 describes the proportion of AEs in baclofen' SPC considered as signal when applying the different algorithms.

Table 2. Proportion of AEs in the SPC considered as a signal for each decision rule

Method	PRR1	PRR2	ROR1	ROR2	RFET	midRFET	BCPNN1	BCPNN2	GPS1	GPS2
%SPC signal	50%	64%	64%	64%	64%	64%	61%	64%	36%	64%

In order to confirm some unexpected AEs detected by these algorithms we used an implementation of logistic regression that takes into account all drugs present in the database and their possible interactions, the CISL algorithm. As this process necessitates large computational resources and long delays, we only considered four unexpected AEs: "Glaucoma", "Bulimia nervosa", "Sleep apnea syndrome" (SAS) and "Nightmare". The signals detected by the traditional methods were confirmed by the CISL algorithm for all of them.

### Discussion

We demonstrated that applying traditional signal detection methods to web forums allows the detection of AEs described in the SPC of: on average, 60% of the SPC's AEs were considered as signals (vs 35% on average for 4 drugs in [15] using co-occurrences statistics). Only 4% of the AEs found with baclofen in forums were in its SPC (vs 99% for statins in [16] where 72 posts were manually reviewed).

The main application of our results is to check if a manual evaluation of the French pharmacovigilance database and of the medical literature confirms the unexpected detected signals. If confirmed, the national agency should be informed of these potential signals. This review is mandatory to prove that signal detection on social media can improve current knowledge on unexpected AEs.

Our study has some limitations. On the one hand, we probably missed information using a perfect match to normalize drug terms. However, we manually identified several lexical variations, which allowed us to lower this bias. On the other hand, several masking effects exist when applying traditional signal detection methods to social media data, such as under or over represented drugs with specific AEs, or UMLS terms corresponding to several AEs. Moreover, the NLP technique could not detect causality between drugs and AEs, which explains the presence of indications captured as potential AEs.

Nevertheless, this study showed that web forums represent a very important data source for a pharmacovigilance purpose as we could detect expected ADRs as well as potential unknown ADRs. Future work should focus on bias limitation and signal detection algorithm adaptation. To date, the last step, i.e. a manual evaluation of potential signals is inevitable.

#### Acknowledgments

This work was funded by two grants from the ANSM, the French agency for drug safety (*Agence nationale de sécurité du médicament et des produits de santé*) AAP-2013-052 through the Vigi4MED project and Convention n°2016S076 through the PHARES project. We acknowledge Vidal for providing structure and information on drugs based on SPCs and other relevant sources, CHU of Rouen, CISMEF team for providing RacinePharma, and CNRS-LIMSI for detecting mentions of drugs and AEs in posts.

#### References

- [1] We are social, "Digital in 2016," We Are Social Singapore, 2016.
- [2] H. Romeyer, "La santé en ligne," Commun. Inf. Médias Théories Prat., no. Vol. 30/1, Feb. 2012.
- [3] S. Katsahian, E. Simond Moreau, D. Leprovost, J. Lardon, C. Bousquet, G. Kerdelhué, R. Abdellaoui, N. Texier, A. Burgun, A. Boussadi, and C. Faviez, "Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in Adverse Drug Reaction Mining," Stud. Health Technol. Inform., vol. 210, pp. 526–530, 2015.
- [4] B. W. Chee, R. Berlin, and B. Schatz, "Predicting Adverse Drug Events from Personal Health Messages," AMIA. Annu. Symp. Proc., vol. 2011, pp. 217–226, 2011.
- [5] G. E. Powell, H. A. Seifert, T. Reblin, P. J. Burstein, J. Blowers, J. A. Menius, J. L. Painter, M. Thomas, C. E. Pierce, H. W. Rodriguez, J. S. Brownstein, C. C. Freifeld, H. G. Bell and N. Dasgupta, "Social Media Listening for Routine Post-Marketing Safety Surveillance", Drug Saf, 39(5):443-54, May 2016.
- [6] S. Yeleswarapu, A. Rao, T. Joseph, V. G. Saipradeep and R. Srinivasan, "A pipeline to extract drug-adverse event pairs from multiple data sources", BMC Med Inform Decis Mak, 14:13, 2014.
- [7] S. Katsahian, E. Simond Moreau, D. Leprovost, J. Lardon, C. Bousquet, G. Kerdelhué, R. Abdellaoui, N. Texier, A. Burgun, A. Boussadi, and C. Faviez, "Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in Adverse Drug Reaction Mining," Stud. Health Technol. Inform., vol. 210, pp. 526–530, 2015.
- [8] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez, "Utilizing social media data for pharmacovigilance: A review," J. Biomed. Inform., vol. 54, pp. 202– 212, Apr. 2015.
- [9] L. Hazell, SA. Shakir. "Under-Reporting of Adverse Drug Reactions: A Systematic Review." Drug Safety 29, no. 5, 2006.
- [10] B. Audeh, M. Beigbeder, A. Zimmermann, P. Jaillon, C. Bousquet. Vigi4Med Scraper: A Framework for Web Forum Structured Data Extraction and Semantic Representation. PloS one. 2017.
- [11] C. Grouin, F. Morlane-Hondere, P. Zweigenbaum, "Identification of Drug-Related Medical Conditions in Social Media", LREC, 2016.
- [12] I. Ahmed. "Détection automatique de signaux en pharmacovigilance : Approche statistique fondée sur les comparaisons multiples" 2009. Thesis.
- [13] I. Ahmed, A. Pariente, P. Tubert-Bitter. Class-Imbalanced subsampling lasso algorithm for discovering adverse drug reactions. Stat Methods Med Res 2016.
- [14] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016.
- [15] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard and J. H. Holmes, "Identifying potential AE using the web: a new approach to medical hypothesis generation", J. Biomed. Inform.; 44(6): 989-996, Dec 2011.
- [16] F. Kheloufi, and M. J. Jean-Pastor, "Sharing experience with adverse drug reactions on internet: Which adverse drug reactions involving statins are described by patients on internet-based-forums?" Fundamental and Clinical Pharmacology, 28:56-56, May 2014.