PhUSE 2018

Adverse Drug Reactions detection on social media: bias and limitation

Erwan Le Covec, Keyrus Biopharma, Lasne, Belgium
Lise Radoszycki, Else Care, Paris, France
Stéphane Chollet, Keyrus Biopharma, Lyon, France

ABSTRACT

Social media are computer-mediated technologies that facilitate the creation & sharing of information and are an important source for Adverse Drug Reaction (ADR) collection. This can help to reduce under-reporting in post-marketing phase products or monitor drug-specific trends.

In the paper "*Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting*" we concluded that the application of social media is subject to challenges because the data are inconsistent, unstructured and region-specific.

In an attempt to solve these challenges, the current paper will focus on the following topics:

- Increasing the number of sources and languages;
- Diversifying and comparing data sources (general social media vs specialized media such as the online patient community Carenity®).

This analysis will provide a better understanding of the collected data, allowing us to estimate biases when working with social media, and will lead to a new field of application: the pharmaco-epidemiology.

INTRODUCTION

With the increasing role of social media, people tend to have a stronger presence on the web and do not hesitate to share information which would have been kept private 10 years ago. It is now possible to find information on a person's health issues, as well as their prescriptions. Adverse Drug Reactions (ADRs) have traditionally involved the use of data mining techniques on spontaneous reports submitted to national surveillance systems, also called Spontaneous Reporting System (SRS). However the interest towards the use of new data sources, especially the ones coming from the Internet, is growing. Several publications have been issued regarding the subject: they deal with the extraction of ADRs from Twitter®, health-related forums and even search queries[1,2,3.]

It appears to be possible to extract medical concepts from these data which can complement the traditional ways of ADRs detection. However, some points need to be taken into account like the fact language has an important impact on the ADRs detection. The choice of the data sources is also crucial. Classic social media does not carry the same information as a health-oriented forum and thus, some maybe be more valuable than others.

The present paper aims to provide a more in-depth understanding of the data which can be collected for ADR detection, to estimate bias and limitations across sources and languages.

CONTEXT

PHARMACOVIGILANCE AND PHARMACOEPIDEMIOLOGY

The Pharmacovigilance science is related to the collection, detection, assessment, monitoring, and prevention of Adverse Events (AEs) with pharmaceutical products. A safety signal is information on a new or known AE (whose frequency increases or decreases) that may be caused by a medicine and requires further investigation. A wide range of sources can be involved to detect these signals. According to the guidelines on Good Pharmacovigilance Practices (GVP), social media can be a source of potential valid Individual Case Safety Reports (ICSRs) because they allow patients and healthcare professionals to communicate ADRs and thus can be used as sources for new signal identification[4].

However finding real signals on social media is difficult and very time consuming. A signal must be related to an identified person which is often complex on the Internet where people are most of the time pseudonymised.

“Pharmacoepidemiology is defined as the study in real conditions and on large populations, of use, effectiveness and risk of drugs. The methods and fields of application of Pharmacoepidemiology are described. They allow to characterize conditions of use, misuse, clinical effectiveness, adverse drug reactions and risks of drugs. The development of Pharmacoepidemiology should allow optimization of ‘‘rational use’’ of drugs”[5].

For this reason, extracting ADRs on social media could be better related to Pharmacoepidemiology than to Pharmacovigilance.

PREVIOUS PAPER

In “*Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting*”, we proposed an automated method to extract, detect and store ADRs from a multi-source unstructured data collection system in English. We focused our analysis on the website Twitter® and Reddit® in English showing that the collected data could not be qualified for signal detection but still carried valuable information for more general reporting. However with this method, it required a lot of manual revision to isolate real ADRs from symptoms or other medical events[6]. Reducing this manual work could be done using more advanced natural language processing techniques but requires a good understanding of the extracted data.

ADVANCED NATURAL LANGUAGE PROCESSING PREPARATION

To enhance ADR detection and to reduce manual revision of the extracted records, machine learning can be used with the help of word embedding. Using such techniques requires a large amount of annotated data. But to have an efficient algorithm, data must be representative of reality. Bias must also be known and understood because such algorithms are known to reproduce them6. That is why using several languages and various kinds of sources can help us in this task.

Social media can induce multiple biases. Users of online forums may not be representative of the world population and they may not share medical data in the same way they do with their physician. But the current lexicon-based method can also induce biases by being less effective on some sources or languages. By comparing the data, it is possible to measure potential errors and be able to correct them to prepare machine learning training.

METHOD

USED METHOD

The used method aims to detect, normalize and store ADRs in text extracted from social media. Detection is done by the lexicon-based methods which consist in searching if a word of a lexicon is present in a text. Data exclusion is performed after the extraction, during the cleaning and processing of the data. A normalization step to code the drugs and the ADR detected completes the process using the WHO Drug Dictionary Enhanced (WHO-DDE) and the Medical Dictionary for Regulatory Activities (MedDRA).

DATA SOURCES

Reddit® and Twitter® were the two sources used in “Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting”. Even if ADR extraction results were not the same, these two sources are both generalist social media and thus may carry biased or false information7. In addition to the previous sources, a new one has been selected: Carenity®. Adding this source can help us to confirm our existing hypothesis but also in finding new evidence showing social media as interesting sources for ADR detection. As it is a health-oriented forum, it carries other information than more generalist social media like Twitter®.

Carenity® was chosen because it is an Internet social health network for patients suffering from chronic diseases, their families and the caregivers. It provides discussion groups and health monitoring space, as well as information on diseases and treatments. As of 2018, the site has more than 300,000 active members, and more than 1,200 pathologies (chronic and rare diseases) are listed8. As a specialized website, it also contains valuable metadata regarding patient information which is sometimes difficult to get from other websites. These metadata (age, gender, disease…) can provide new insights and help define the population that uses social media for ADR reporting.

LANGUAGES

The choice of the language was made with the aim of having the best worldwide coverage. Languages which do not use the Latin alphabet were excluded because due to the complexity posed to the systems[9].

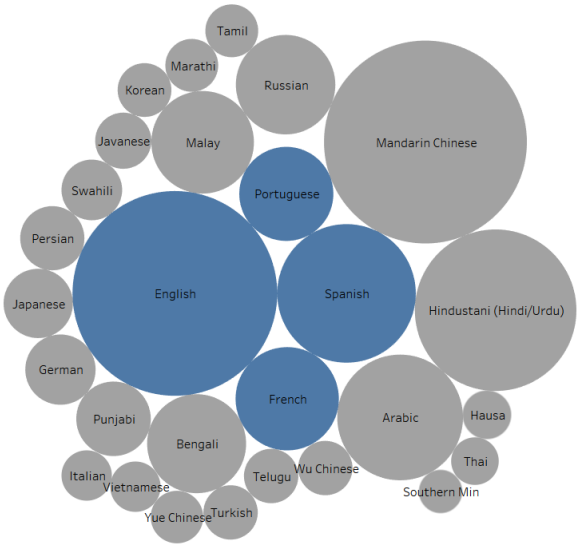

Figure 1: Most used languages in the world.
In blue, the languages selected for the study, representing 32% of the world's population

We selected the following languages, covering potentially 32% of the world's population (figure 1):

- English
- French
- Spanish
- Portuguese

Using these four languages could help to detect and cultural biases in the data.

For Reddit®, only data in English was extracted because it is the main language of these website and insufficient data in other languages. Since Carenity platforms are only available in English, Spanish, French, German and Italian, Portuguese content derives exclusively from Twitter®. However, the majority of the records are in French as it was the primary language of the website.

ANALYSIS METHOD

The analysis tries to answer four main questions on ADR detection:

- Does the number of sources have an influence?
- Are health-related websites more accurate?
- Does the language have an influence?
- How can we use metadata for ADRs detection?

As health-oriented websites were targeted for the analysis, drugs were chosen accordingly to the number of records it was possible to find on Carenity®. As the drugs with the most records were for specific treatments, two additional drugs (Aspirin and Ibuprofen) were added to allow for a better comparison. Extraction for Carenity® and Reddit® was done on all the historic data. The oldest record for Carenity® was from 2011 for French, 2014 for English and 2015 for Spanish (dates that the sites were created). Regarding Reddit® the oldest record is from 2007. For Twitter®, one year of data was extracted because of Twitter's® limitations which means the oldest available record was from 2017.

The following drugs were targeted for data extraction:

- Ibuprofen: anti-inflammatory
- Acetylsalicylic acid: anti-inflammatory
- Baclofen: for spasticity
- Gabapentin: for partial seizures, neuropathic pain, hot flashes, and restless legs syndrome
- Clonazepam: tranquilizer to treat seizures, panic disorder, and movement disorder
- Pregabalin: for epilepsy, neuropathic pain, fibromyalgia, and generalized anxiety
- Levothyroxine: for thyroid hormone deficiency
- Duloxetine: for major depressive disorder, generalized anxiety disorder, fibromyalgia and neuropathic pain
- Methotrexate: for cancer, autoimmune diseases, ectopic pregnancy, and for medical abortions
- Etanercept: for autoimmune diseases
- Dimethylfumarate: for multiple sclerosis
- Amitriptyline: antidepressant
- Prednisone (with Prednisolone or Hydrocortisone as synonym): for inflammatory disease, autoimmune disease and cancer
- Hydroxychloroquine: for malaria, rheumatoid arthritis, and lupus
- Venlafaxine: antidepressant
- Adalimumab: for rheumatoid arthritis, psoriatic arthritis, ankylosing spondylitis, Crohn's disease, ulcerative colitis, chronic psoriasis, hidradenitis suppurativa, and juvenile idiopathic arthritis

The drug molecule and the brand name were used for all extractions, except for Aspirin were only ‘Aspirin’ was used. As an example, ‘Levothyrox’ was used for the extraction in French as it is most known with this term but also ‘Lyrica’, a commercial name for Pregabalin.

RESULTS

Drug	Twitter	
Ibuprofen	152.459	*43,28%*
Acetylsalicylic acid	107.133	*30,41%*
Gabapentin	25.511	*7,24%*
Clonazepam	25.440	*7,22%*
Levothyroxine	12.492	*3,55%*
Pregabalin	6.796	*1,93%*
Adalimumab	5.265	*1,49%*
Baclofen	4.672	*1,33%*
Venlafaxine	4.500	*1,28%*
Methotrexate	3.514	*1,00%*
Duloxetine	2.852	*0,81%*
Etanercept	1.056	*0,30%*
Dimethylfumarate	445	*0,13%*
Amitriptyline	133	*0,04%*
Prednisone	14	*0,00%*
Hydroxychloroquine	3	*0,00%*
Total		**352.285**
Median		*4.586*

Drug	Reddit	
Ibuprofen	29.756	*18,21%*
Acetylsalicylic acid	27.070	*16,56%*
Gabapentin	17.095	*10,46%*
Clonazepam	16.306	*9,98%*
Levothyroxine	15.570	*9,53%*
Baclofen	14.272	*8,73%*
Pregabalin	13.219	*8,09%*
Duloxetine	9.418	*5,76%*
Methotrexate	5.349	*3,27%*
Venlafaxine	5.117	*3,13%*
Etanercept	4.962	*3,04%*
Dimethylfumarate	4.577	*2,80%*
Adalimumab	595	*0,36%*
Amitriptyline	120	*0,07%*
Prednisone	0	*0,00%*
Hydroxychloroquine	0	*0,00%*
Total		**163.426**
Median		*7.384*

Drug	Carenity	
Adalimumab	1.959	*15,50%*
Pregabalin	1.684	*13,33%*
Methotrexate	1.159	*9,17%*
Duloxetine	1.135	*8,98%*
Etanercept	1.102	*8,72%*
Clonazepam	1.057	*8,36%*
Hydroxychloroquine	863	*6,83%*
Dimethylfumarate	858	*6,79%*
Amitriptyline	767	*6,07%*
Levothyroxine	584	*4,62%*
Ibuprofen	570	*4,51%*
Gabapentin	358	*2,83%*
Prednisone	283	*2,24%*
Acetylsalicylic acid	136	*1,08%*
Baclofen	119	*0,94%*
Venlafaxine	3	*0,02%*
Total		**12.637**
Median		*813*

Table 1: Records extracted

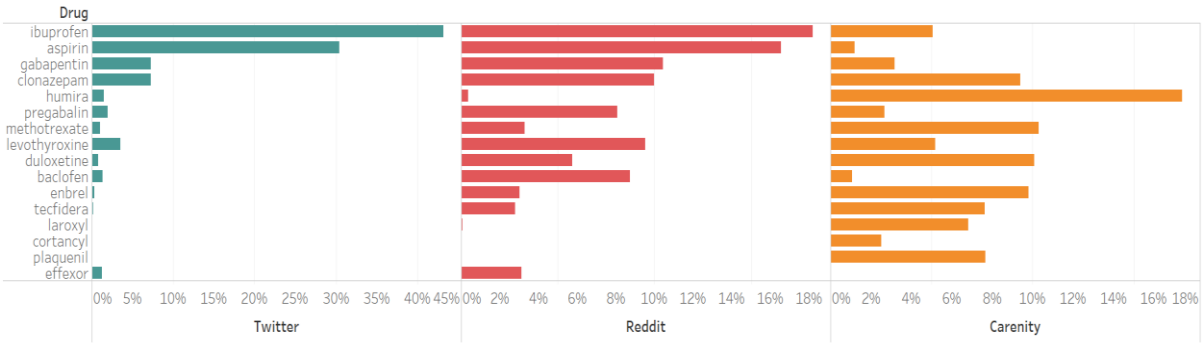

Figure 2: Drug representation in Twitter®, Reddit® and Carenity®

Table 1 and Figure 2 describe the number of records extracted from Twitter®, Reddit® and Carenity® depending on the targeted drug. There were overwhelmingly more records from Twitter® than the other networks (2x Reddit® and 28x Carenity®), even though they were extracted within a shorter time period. However, more than 70% of the Twitter’s® data concerned ‘Ibuprofen’ and ‘Aspirin’. For Reddit®, these two drugs represented 34% of the data and for Carenity®, it was not even 10% of the data collected.

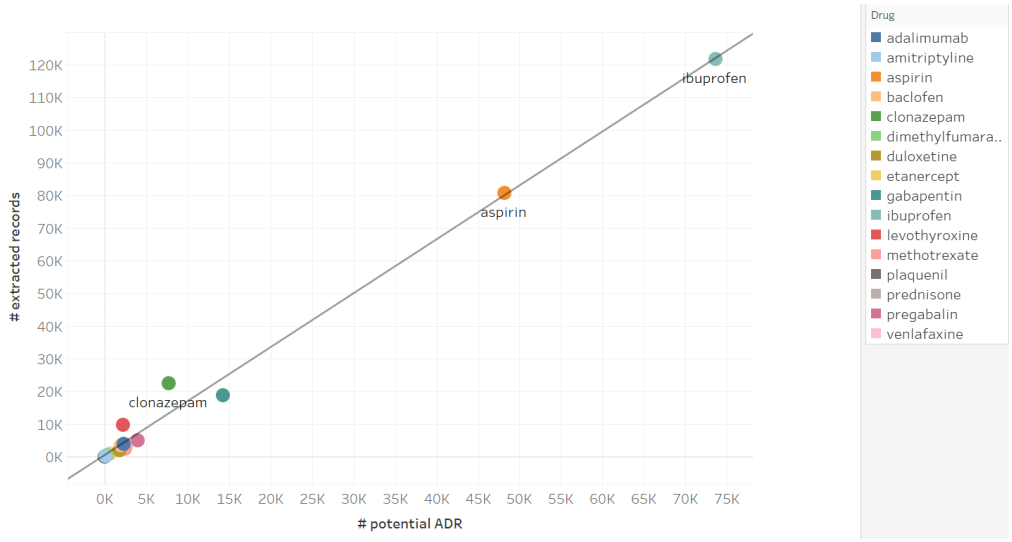

Figure 3: Potential ADR ratio on Twitter® (detected ratio = 35%)

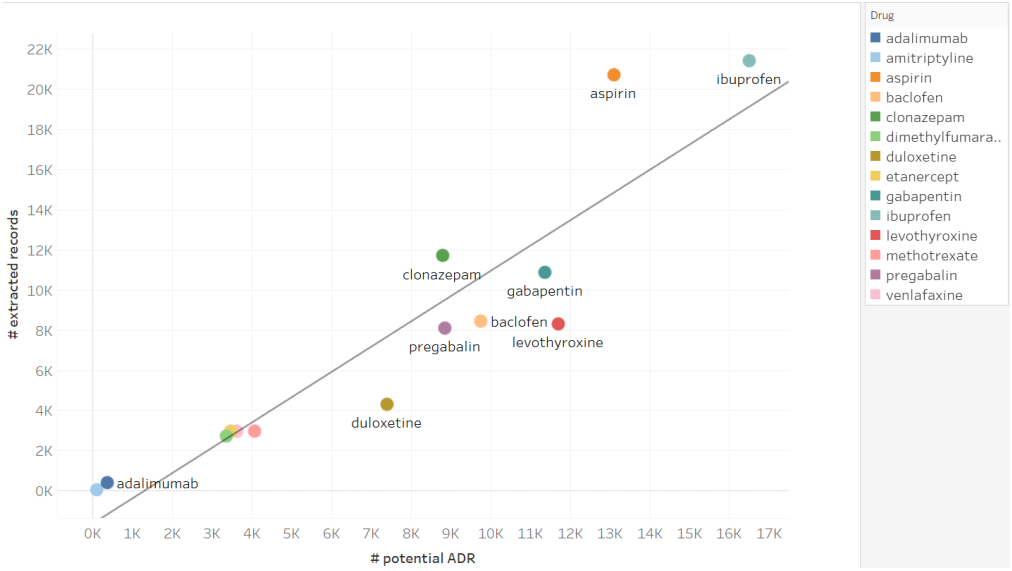

Figure 4: Potential ADR ratio on Reddit® (detection ratio = 74%)

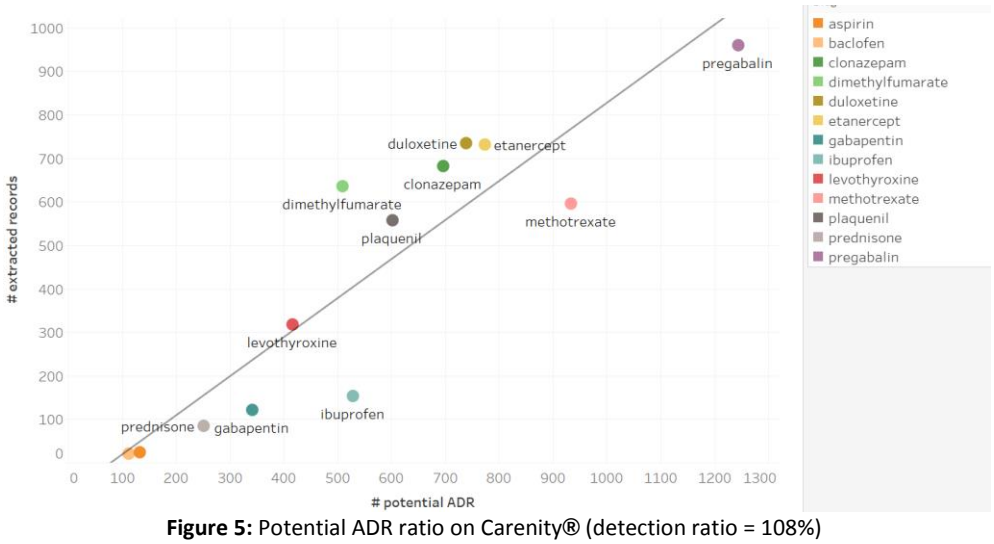

Figure 5: Potential ADR ratio on Carenity® (detection ratio = 108%)

Figures 3, 4 and 5 show the number of potential ADRs detected on Twitter®, Reddit® and Carenity®. The ratio represents the number of potential ADRs divided by the number of records. The detection ratio gives an idea of how many records contained potential ADR. For Reddit®, the ratio was high with a potential ADR ratio of 74%. For Carenity®, the ratio was even stronger indicator of ADR with 108%. A ratio over 100% signifies that several ADRs could potentially be found within a single record. Regarding Twitter®, the detection ratio was much lower with a potential ADR of 35%. The trend curve can be used to identify which drugs could have more potential

ADRs than others. When a drug falls below the trend line, there is a stronger likelihood that more ADRs could be identified with a smaller sample size. Inversely, when a drug is above the trend line, less ADRs would likely to be detected.

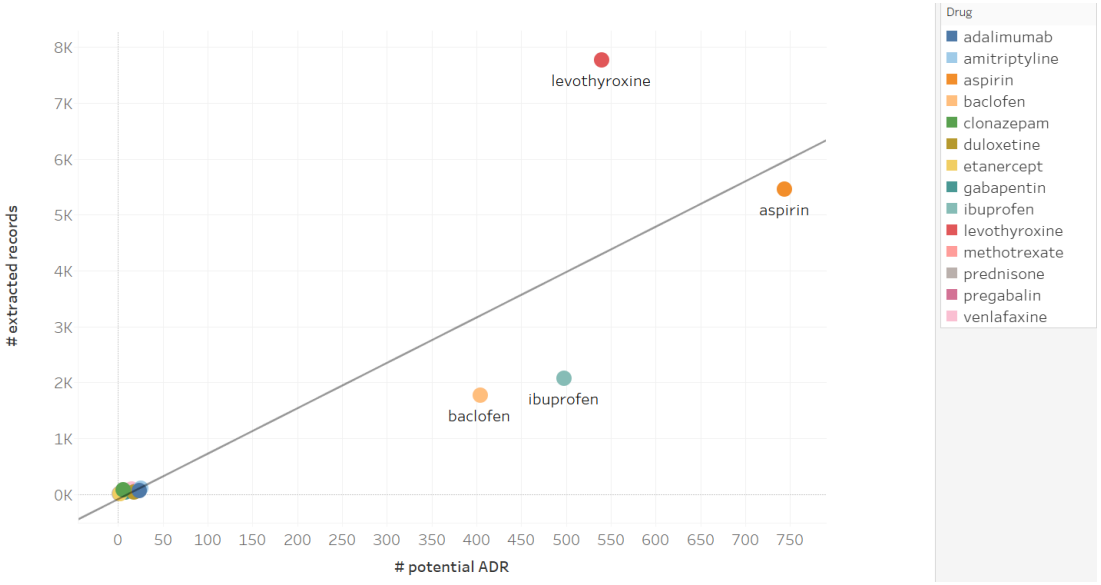

Figure 6: Potential ADR ratio for French language on Twitter® (detection ratio = 15%)

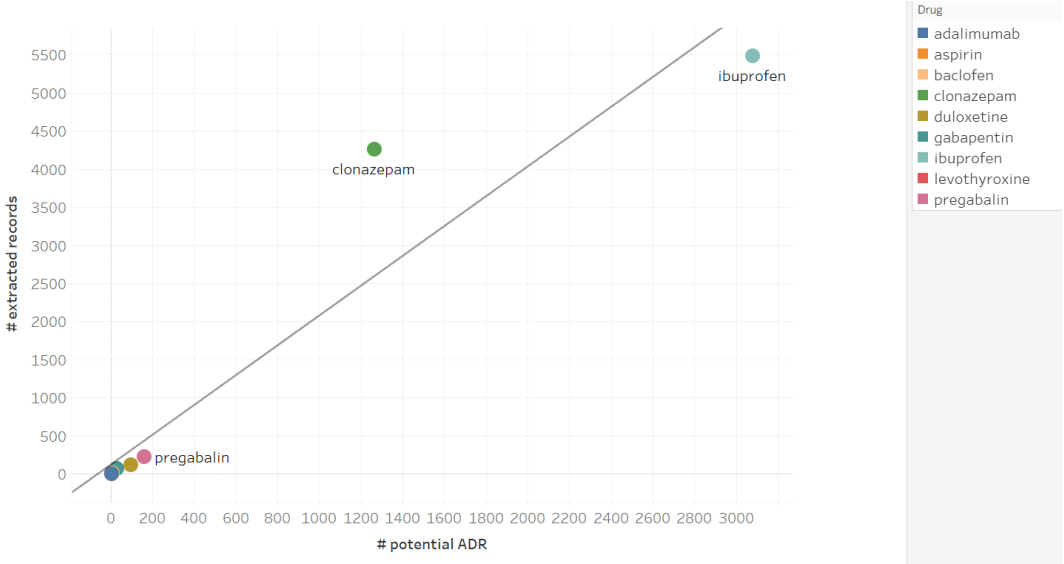

Figure 7: Potential ADR ratio for Portuguese language on Twitter® (30%)

In Figure 6 data are filtered to show only the data in French and in Figure 7 only data in Portuguese are shown. These are data derived from Twitter®. The total number of potentially identified ADRs varies based on the language, it being higher for Portuguese records (30%) than French records (15%). Pertaining to the number of records per drug, we detected variations in the volume of specific-drug mentioned between languages. For example, for French Aspirin and Levothyroxine were the top drugs meanwhile for Portuguese, it was Ibuprofen and Clonazepam.

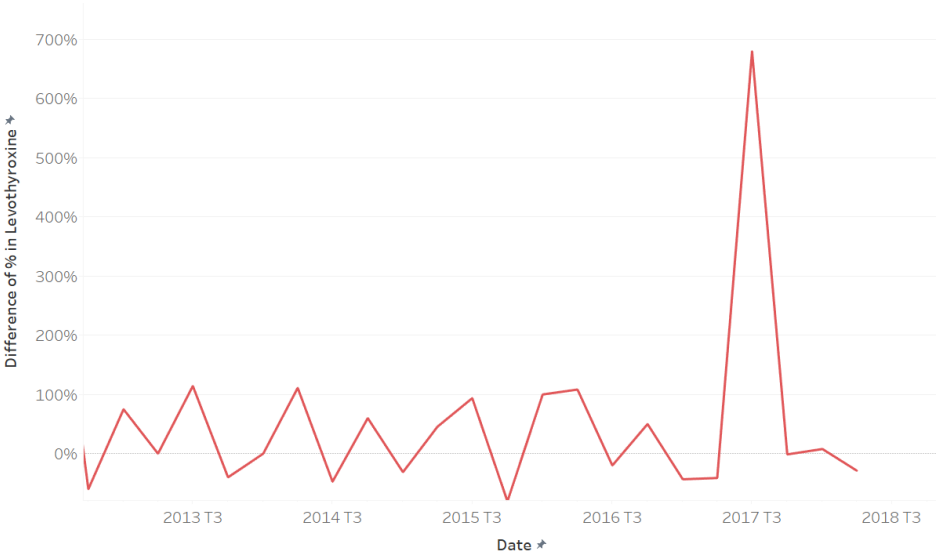

Figure 8: Mention of levothyroxine in Carenity® over the time

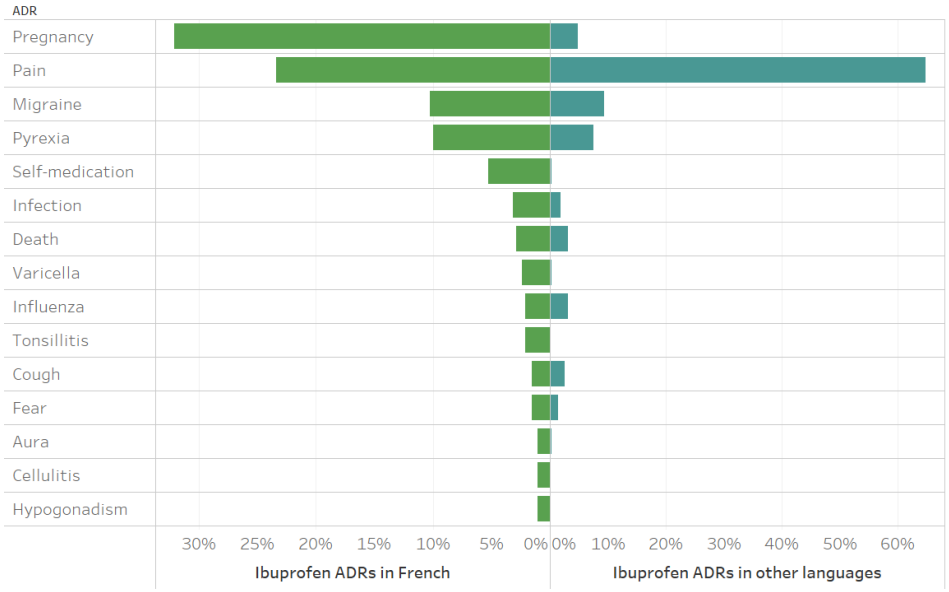

Figure 9: Comparison of ADRs mention for Ibuprofen between tweets in French vs tweets in other languages (Twitter®)

As shown in figure 6, many French tweets contained information about levothyroxine. Figure 8 shows that on Carenity®, the number of French posts about levothyroxine had dramatically increased in 2017. This language particularity is common in both sources.

Regarding potential ADRs, figure 9 shows the difference of ADRs detected for Ibuprofen between French tweets and the other languages. It appears the main potential ADRs were not similar. In French, pregnancy represented more than 30% of the records while its equivalent in the other languages would be pain (30% of records). This over representation of pregnancy may be due to a peak beginning in early 2018.

Figure 10: Age and gender influence

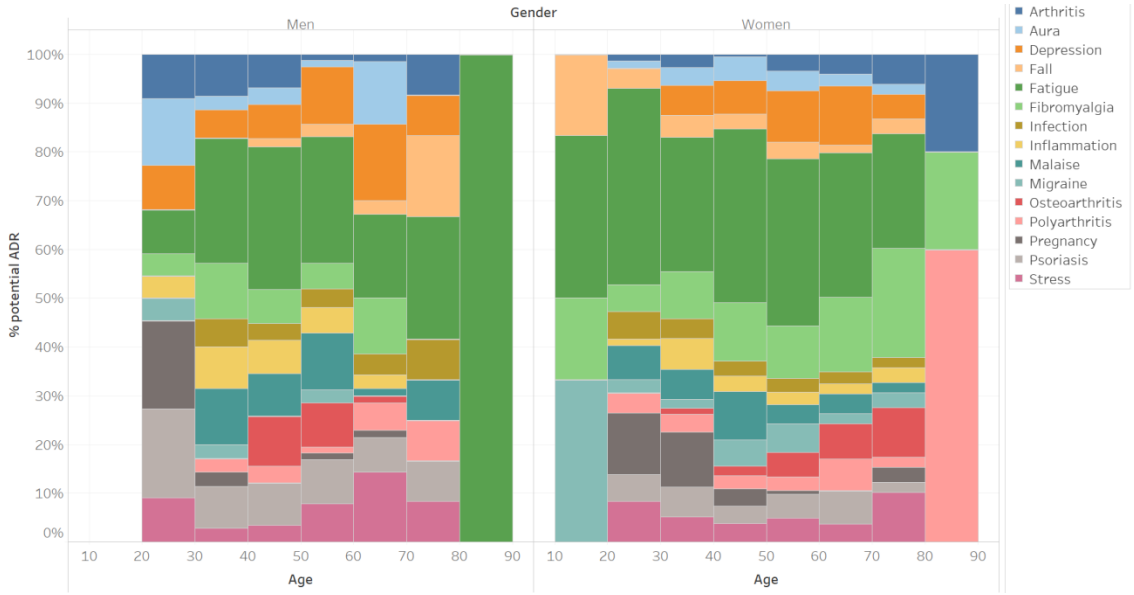

Figure 11: ADR per gender and age

Figures 10 and 11 focus on the metadata of Carenity® because no metadata was available for the other sources. It appears that most of the records come from women whatever the age group. They represented 80% of the records when men represented 20%. The distribution between the total number of records and the potential ADRs are similar except for women between the ages of 30 and 40 who report more ADRs per record. As for the ADR per gender, we can see they are not identical. Men tend to discuss more from stress and depression than women who mostly talked about fatigue. For the last age class, only a few records are available which explains why the ADRs are not diverse.

	Reddit® En	Twitter® En	Carenity® En	Total En	Carenity® Fr	Final Total
# detected ADR	318	318	318	*954*	318	*1,272*
# source record	128	194	78	*400*	157	*557*
# real ADR *% of the total*	31 *10%*	17 *5%*	54 *17%*	*102* *11%*	23 *7%*	*125* *10%*
# false ADR	287	301	264	*852*	287	*1,139*
# ADR missed *% of missed*	6 *16%*	4 *19%*	4 *7%*	*14* *12%*	32 *58%*	*46* *27%*

Table 2: True/false positive and false negative in a subset of Methotrexate ADR detection

Table 2 is the result of a manual analysis performed to identify, for each source, if the potential ADRs are real, missed or false. For this table, 318 potential ADRs were chosen randomly for each source. These 318 ADRs were analyzed to understand if any of the potential ADRs were real. Additionally, by examining the records, we were able to identify other real ADRs that were not detected by our methodology. Methotrexate was chosen because it was the only drug having enough data allowing this comparison. Data for Twitter® in French were too few but gives similar results to Carenity® in French. The 318 potential ADRs represent only 128 unique records for Reddit®, 194 for Twitter® in English, and 78 for Carenity® in English and 157 for Carenity® in French.

Amongst these 1,272 potential ADR, few are real ADRs (only 125), the others are medical conditions, symptoms or other problems which are not linked to the drug and can be considered as false positive. This table also shows that on average few real ADRs are missed, (e.g. false negative detection) except for Carenity® in French where most of the ADRs are not well detected whereas for Carenity® in English, 54 real ADRs were found for 78 distinct records - this could be a method limitation in relation to the language.

DISCUSSION

SOURCES

Using various sources show the limitations of some sources and their data such as misrepresentation of drugs, low ADRs presence or difficulty to analyze the text. It also showed that quality of data is not equal between all sources and the most useful sources with our methodology appear to be the health-oriented website in English, Carenity®. Regarding Twitter®, even if this source is widely used in the research of ADRs[1,2], it does not carry a lot of usable information and is very complex to correctly detect ADRs as shown in the table 2, Twitter® has the lower rate of real ADR detection.

It is shown that the number of various drugs mentioned in the comments is much more congruent in Reddit® and Carenity® than Twitter®. The over representation of Aspirin and Ibuprofen on Twitter® was not noted in the two other sources, even if they were also the most common drugs on Reddit®. This contrasts with Carenity® which had less record for common drugs and did have more content about disease-specific drugs. However, this was consistent with the fact that Carenity® focuses on less generic health conditions. These results showed that websites like forums where people can interact with others seem to bring more value when various or disease-specific drugs need to be analyzed.

In general, there was a larger dataset from Twitter® than the other sources, even with Twitter's® data being extracted for one year as opposed to a longer period for the other sources. They are several explanations for this. Firstly, Twitter® limits the number of characters. Thus, a user needs to create additional comments in order to express an idea. This is a non-existing issue for Reddit® or Carenity®. Secondly, the purpose of Twitter® is to share instant life statuses where users can post tweets several times a day and are more inclined to talk about everyday random life events. For forums, users tend to post only when they face a particular issue or when they have need of advice.

Carenity® was the source with the least number of records. It can be explained by the fact that it is a health-oriented website devoted to patients and their caregivers. Because of this nuance, users are less likely to be quality of data tends to be more valuable (see table 2 for English). For Carenity®, the potential ADR ratio was the highest at 108%. Amongst the potential ADRs, a majority were real ADRs and very few real ADR ones were not identified among the potential ADRs. The false-positive terms detected by our methodology seems to be the symptoms caused by the diseases themselves. For Twitter®, few real ADRs were detected nor were missed. It potentially signifies that this source does not carry much information for ADR detections. The ADR detected tended to be mostly symptoms or events which were not related to drugs. Reddit® was between Twitter® and Carenity® which means its data was more valuable than Twitter® but not as indicative as Carenity®. However, this observation is true for the English language though not for the content in other languages.

From a machine learning perspective, as each source is very different, one algorithm per source and language or should be used. Additionally, final results should be weighted according to the relevance of the source.

LANGUAGES

The ADR ratios and records vary significantly from one language to another (according to figures 3, 6 and 7). French content had a much lower ratio than English on Twitter® and data repartition was also not similar. With our current methodology, results were less relevant when we used French data than English data. It is confirmed by table 2 which shows for Carenity® in French, a significant amount of real ADRs are missed. It can be explained by the complexity of the French grammar and syntax structures when applied to a computer algorithm as the method, based on dictionaries, cannot grasp certain subtleties of languages. These expressions can be difficult to handle with our current lexicon-based method but could be detected with a new method based on syntactic comprehension of the sentences. New deep learning techniques on natural language processing such as GloVe or Word2Vec could be in use to analyze our French dataset.

Language had also an impact on the representation of drugs. In French, Levothyroxine was frequently debated while in Portuguese it was Clonazepam. Clonazepam appears to be widely used in Brazil for recreational purposes and would probably explain why it was one of the most commonly mentioned drugs in Portuguese[10]. Regarding Levothyroxine, it could be explained by the fact it was mentioned throughout France media outlets because of a chemical change that may have caused complications to many patients. The timeframe it was reported clearly represented on figure 8 where a peak of levothyroxine is observed at the end of 2017. A similar behavior can also be noted for Ibuprofen in French. In this language, the first detected ADR was pregnancy while it being the 12th position for other languages. Pregnancy is not an ADR of Ibuprofen, though it is not a suggested drug during pregnancy, there was a lot of mentions detected with the two. The peculiarity for French was that at the beginning of 2018, national drug safety agency said the drug is contraindicated from the beginning of the 6th month of pregnancy, regardless of the duration of treatment and the route of administration. Screening of the news in each of the countries involved in the analysis is needed to avoid misconception and misunderstanding. But it can also be used to measure the impact of the media on a population.

As for the sources, to go further in the analysis, one algorithm per language should be used to capture the subtlety of each language and better capture ADRs. Regarding the news effect, data to be used for the training of machine learning algorithm could be limited to a time period that would not be affected.

METADATA

Carenity® was the only source which provided reliable metadata such as age or gender as unlike Twitter®, the website is medically oriented. It allowed for a better understanding of the extracted data and the profile of the users. Twitter® also provides some metadata but it is highly inconsistent as users have a lower relevancy to specify their socio-medical profile.

In figure 10, it appears that most of the patients talking about drugs are women. Men represented less than 20% of the contributors. It is possible that women seek more information about their conditions and are more willing to share their experiences. In France, at least, it appears women do more medical research on the Internet which can explain the gender proportion of Carenity[11]. This over representation of women needs to be taken into account because of the potential different behavior in gender (as demonstrated by J. Rowley et al. [12]). Even if gender information is not specified on Reddit®, studies have showed that most of the users are men (67%) 13. Studies regarding male pathologies may want to focus more on Reddit® and studies regarding female pathologies may want to focus on Carenity®.

Gender had also an influence on the type of ADRs detected. Figure 11 at first glance the distribution appears similar; however it is possible to identify gender-specific variations in the dataset. There were more mentions of

fatigue and malaise amongst women while men seem more open to discuss about stress or depression. It can be explained by the fact that a majority of men who had talked about stress or depression had psoriasis, which is known to have this kind of side effects. For women, fibromyalgia is the most present among those who mentioned fatigue.

On overall, the data repartition between the age groups was consistent except for women between 30 and 40. It appeared that at this age, users tend to report significantly more ADRs by record. It is likely that this group has a higher familiarity and comfort with social media and thus is more open to discuss publicly[14,15]. It means no specific correction on the age class should be made for future machine learning but in the other hand, records from Carenity® should be pondered according to the gender frequency.

CONCLUSION

Detected ADRs on social media is a growing interest of concern and attention with the pharmaceutical and scientific communities. Existing research often focus only on the use of English in Twitter®. This paper tried to examine several different sources and languages to evaluate and eventually find ways to eliminate inherent biases and limitations of readily available data sources. With the end goal to obtain a higher quality of data to utilize in conjunction with machine learning algorithms. This research also allowed for a better understanding of the reliability of our lexicon-based method. Due to the lack of user information for several sources, using social media as a detector of safety signals is challenging but could be interesting in the context of Pharmacoepidemiology.

Our research highlighted the main differences between social media platforms and languages. With Forums, users can explain more in depth their issues and be in active discussions with others tend to be more valuable whereas sources such as Twitter® carry copious amounts of noise, have large amount of unusable data. A solution of using deep syntactic analysis in order to enhance the detection of other languages (not English) and to avoid the consideration of medical symptoms as ADRs might be used. Thanks to the bias we identified in this study, we will be able to correct our data set using Machine Learning in order to better detect ADRs.

DISCLAIMER

The data contained in this paper was obtained for demonstration purposes only using the techniques presented. Anyone analyzing data from social media sites, either public or membership based, should investigate the source terms and conditions relevant to data reuse and obtain any necessary permissions if required.

REFERENCES

1. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, L Smith K. Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions [Internet]. 2017 [cited 13 July 2017]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419871/

2. Nikfarjam A, Gonzalez G. Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments [Internet]. 2017 [cited 13 July 2017]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243273/

3. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features [Internet]. 2017 [cited 13 July 2017]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4457113/

4. Guideline on good Pharmacovigilance practices (GVP) - Module VI [Internet]. 1st ed. 2017 [cited 25 April 2017]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf

5. Academie-medecine.fr [Internet]. 2018 [cited 22 September 2018]. Available from: http://www.academie-medecine.fr/wp-content/uploads/2016/06/PAGE-263-274.pdf

6. Biased bots: Artificial-intelligence systems echo human prejudices [Internet]. Princeton University. 2018 [cited 11 September 2018]. Available from: https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices

7. Le Covec E, Ghanem E, Chollet S. Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting [Internet]. 2018 [cited 10 September 2018]. Available from: https://www.phusewiki.org/docs/Conference%202017%20DH%20Papers/DH01.pdf

8. Who are we? - Carenity [Internet]. Carenity.co.uk. 2018 [cited 23 September 2018]. Available from: https://www.carenity.co.uk/who-we-are

9. Ethnologue: Languages of the World [Internet]. Ethnologue. 2018 [cited 10 September 2018]. Available from: https://www.ethnologue.com/

10. Rivotril, a droga da paz química [Internet]. Carte Capital. 2015 [cited 23 September 2018]. Available from: https://www.cartacapital.com.br/saude/rivotril-a-droga-da-paz-quimica-3659.html

11. Lartizien G. Le patient internaute qui est-il ? : que recherche-t-il ? : comment lui adapter nos pratiques ?. [Internet]. 2012 [cited 23 September 2018];. Available from: http://gedscd.univ-lille2.fr/nuxeo/site/esupversions/6ae3f8f7-871e-43e5-b73c-7abd586c5528

12. Rowley, J. , Johnson, F. and Sbaffi, L. Gender as an influencer of online health information-seeking and evaluation behavior. J Assn Inf Sci Tec. 2017 [cited 23 September 2018]. Available from: https://doi.org/10.1002/asi.23597

13. Barthel M, Stocking G, Holcomb J, Mitchell A. 1. Reddit news users more likely to be male, young and digital in their news preferences [Internet]. Pew Research Center's Journalism Project. 2018 [cited 11 September 2018]. Available from: http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

14. Share of internet users among the population in France in 2016 a. France: internet usage penetration by age 2016 | Statistic [Internet]. Statista. 2018 [cited 23 September 2018]. Available from: https://www.statista.com/statistics/410850/france-internet-usage-penetration-by-age/

15. Share of adults in the United States who use the internet in 2018 a. U.S. internet reach by age group 2018 | Statistic [Internet]. Statista. 2018 [cited 23 September 2018]. Available from: https://www.statista.com/statistics/266587/percentage-of-internet-users-by-age-groups-in-the-us/

ACKNOWLEDGMENTS

This paper was made possible thanks to Carenity and its data science team who provided us with their data and gave us some helpful insights, Keyrus Biopharma's Pharmacovigilance team who helped us at every step of the project, people who helped in reviewing, and everyone who encouraged us and helped us in this work with their kind words and good mood.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Erwan Le Covec
Keyrus Biopharma
Drève Richelle 161, bte 18, Bât. L
Waterloo / 1410
Belgium
Email: erwan.lecovec@keyrus.com

Lise Radoszycki
Else Care
1 rue de Stockholm
Paris / 75008
France
Email: lise@carenity.com

Stéphane Chollet
Keyrus Biopharma
1 place Giovanni da Verrazano
Lyon / 69009
France
Email: stephane.chollet@keyrus.com

Web: http://www.keyrusbiopharma.com

Web: https://www.carenity.com/