Pharmacology and social media: Potentials and biases of web forums for drug mention analysis-case study of France

Bissan Audeh, François-Élie Calvier, Florelle Bellet, Marie-Noëlle Beyens, Antoine Pariente, Agnès Lillo-Le Louet, Cedric Bousquet

▶ **To cite this version:**

HAL Id: hal-02317748

https://hal.sorbonne-universite.fr/hal-02317748

Submitted on 16 Oct 2019

Original Article

Pharmacology and social media: Potentials and biases of web forums for drug mention analysis—case study of France

Bissan Audeh (iD)
Sorbonne Université and Université Paris 13, France

François-Elie Calvier
CHU University Hospital of Saint-Etienne, France

Florelle Bellet and Marie-Noëlle Beyens (iD)
Centre Hospitalier Universitaire, France

Antoine Pariente
University of Bordeaux, France; CHU de Bordeaux, France

Agnès Lillo-Le Louet
Assistance publique—Hôpitaux de Paris (AP-HP), France

Cedric Bousquet
Sorbonne Université and Université Paris 13, France; CHU University Hospital of Saint-Etienne, France

Abstract

The aim of this study is to analyze drug mentions in web forums to evaluate the utility of this data source for drug post-marketing studies. We automatically annotated over 60 million posts extracted from 21 French web forums. Drug mentions detected in this corpus were matched to drug names in a French drug database (Theriaque®). Our analysis showed that a high proportion of the most frequent drug mentions in the selected web forums correspond to drugs that are usually prescribed to young women, such as

Corresponding author:
Bissan Audeh, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS), Campus des Cordeliers, INSERM U 1142, 15 rue de l'école de médecine, 75006 Paris.
Email: bissan.audeh@gmail.com

combined oral contraceptives. The most mentioned drugs in our corpus correlated weakly to the most prescribed drugs in France but seemed to be influenced by events widely reported in traditional media. In this article, we conclude that web forums have high potential for post-marketing drug-related studies, such as pharmacovigilance, and observation of drug utilization. However, the bias related to forum selection and the corresponding population representativeness should always be taken into account.

Introduction

Web forums have become a major platform for information sharing. Online discussions reflect the interest of the population for a topic at a given time. As health is a major preoccupation for many people, medical-related topics regarding diseases and treatments are often present in web forums. Such data constitute a valuable resource for research in public health.

In the pharmacovigilance domain, studies examining the interest of social media for drug safety have gained much attention in the past few years.[1,2] Several neologisms were even proposed to introduce this new approach, such as "Pharmacovigilance 2.0,"[3] "Cyber pharmacovigilance"[4] and "Digital pharmacovigilance."[5] In addition to research on adverse drug reactions (ADRs) performed from preclinical or clinical data, social media are also an appealing information resource for other drug-related domains, such as drug utilization studies,[6–9] drug misuse,[10,11] attitudes of patients concerning safety issues,[12] or drug development process.[13] Although most of the studies agree on the high potential of using social media in drug research, many challenges have been identified in this perspective, including, for instance, spams elimination and patient language processing and interpretation.[14,15] As the cost of retrieving and managing big data from social media is high, the evaluation of the added value of such a resource is necessary.

In this context, the main aim of our work was to analyze drug mentions in web forums in order to assess the utility of this data resource for drug-related studies. This global objective was characterized through the following study questions:

1. What are the most recurrent drugs in users' posts?
2. Does the evolution over time of a drug mentioned in web forums correspond to the events reported for this drug in the traditional media (Press, TV, and Radio)?
3. Do the most prescribed drugs in France correspond to the most mentioned drugs in web forums?

In the literature, few studies have addressed these issues. To our knowledge, only Wiley et al.[16] and Carbonell et al.[17] investigated drug mentions in social media without focusing on a specific type of drug or disease. Wiley et al.[16] studied drug categories involved in users posts within several types of social media, including Twitter, Google+ and web forums. Their analysis was based on drugs categories, and did not point out the most popular active ingredients or trade names. Carbonell et al.[17] analyzed mentions of trade names in twitter over a limited period of 3 weeks. Recently, Mahroum et al.[18] exploited statistics about web search queries related to the vaccination against the Human Immunodeficiency Virus (HIV). This last approach required predefining search keywords and evaluating the proportion of Google queries that are related to these keywords. Unlike the

study of Mahroum et al.,[18] our study is interested in users' discussions in web forums rather than online queries, and it does not focus on a specific disease.

None of the previous studies achieved a general analysis of drug mentions in social media. They also did not analyze the evolution of drug mentions in web forums over time and the relation between this evolution and related events in traditional media. Our work in this article aims to fill these gaps.

Methods

To achieve our study, the following three steps were realized: (1) data extraction and preparation, (2) automatic detection of drug mentions, and (3) analysis of drug mentions over users' posts. This article is an ancillary work of the French national project Vigi4Med (Vigilance for Medication in web forums) that aimed to detect ADRs from user posts in Web forums. Thus, the first two steps have been exhaustively described in previous publications related to the Vigi4Med project.[19,20] The result of the first step[19] was the extraction of more than 63 million posts between the years 2000 and 2015 from 21 web forums selected by pharmacovigilance experts. These forums were chosen using Google keyword-based search and a list of certified health websites as described in Karapetiantz et al.[21] The names of these forums and the number of posts extracted from each forum can be found in Audeh et al.[19] The second step allowed the detection of drugs and pathologies mentions from users' posts using a machine learning approach.[20] This approach used two successive classifiers based on well-known methods in machine learning: a Conditional Random Fields classifier[22] to identify medical-related named entities and a Support Vector Machine classifier[23] to identify relations between the entities.

The data extracted for the Vigi4Med project were filtered out to keep only the posts containing automatically detected mentions of at least one drug and one pathology without considering causality assessment. As we can see from Figure 1, these data consisted of 55,350,564 couples (drug, pathology) corresponding to 6,572,528 unique posts and 829,363 unique drug mention. The matching process that we will describe in this article will lead to 4165 drug mentions found in 1,627,793 posts that represent the final corpus of our analysis. We imported all the data into a Mysql database which we used as our analysis environment. In the following subsections, we describe the processes that we applied within the third step of our study regarding the analysis of drug mentions over users' posts.

Ethical approval

In this work, confidentiality was considered by storing the data on a server with restricted and secured access. Although this work was achieved before the application of the European General Data Protection Regulation (GDPR), a special attention was paid to respect users' privacy by anonymizing users' identities and submitting an official declaration about data usage to the National Committee on Computers and Liberties (CNIL).[21]

Normalization by matching to a drug database

In order to explore our data, we matched drug mentions detected in step 2 to global trade names (e.g. Actifed®) or active ingredients (e.g. ibuprofen) in an official drug database. An active ingredient is a substance that alone or in combination with other ingredients is considered to fulfill the intended activity of a medicine. We chose not to match drug mentions to complete trade names that include dosage and forms information (e.g. Actifed 10 mg tablet). In fact, complete trade names are
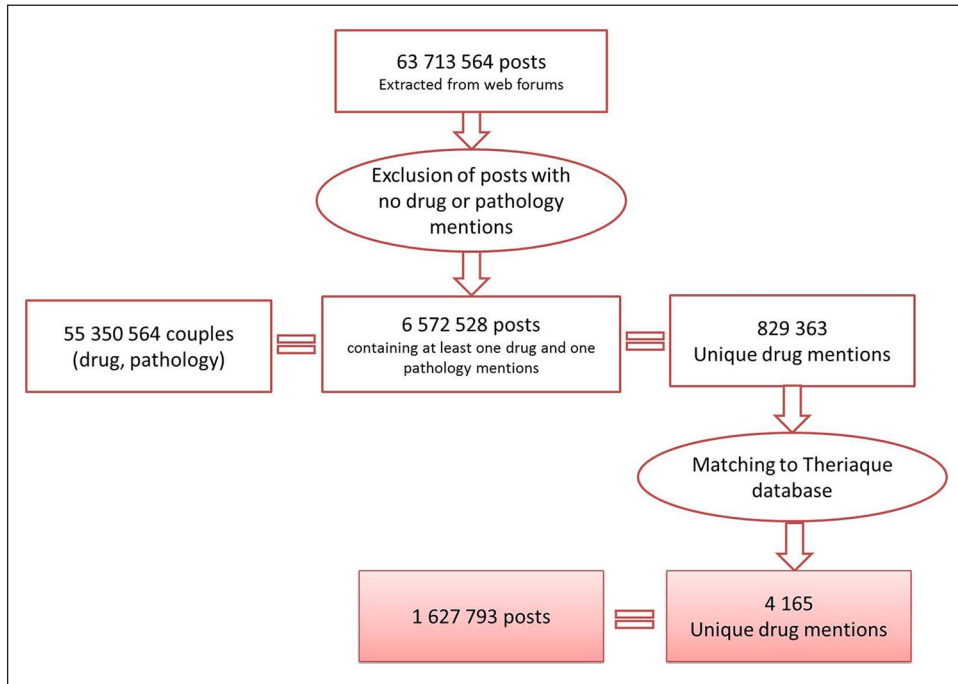
Figure 1. Illustration about the selection process that led to the posts and mentions included in our study.

often composed of several terms and numbers. Our observation of multiple users' posts showed that these complex names are not often employed in users' posts. In addition, the probability of precisely matching them to complete trade names in a reference drug database is low.

The resource that we used to match drug names was Theriaque®, a French drug database developed by the National Hospital Centre on Medicines Information (CNHIM).[24] Table 1 illustrates some associations extracted from Theriaque® for the trade name ACTIFED® and some of its complete trade names.

The preprocessing of drug mentions consisted in removing points, extra spaces and commas. All matching operations were performed using case and accent insensitive comparisons. For each drug mention, the matching process starts by checking if there is a match to an active ingredient in Theriaque®. If a match is found, the drug mention is associated with the matched active ingredient; otherwise, the same process is repeated to match a trade name. If a match is found, the mention is associated with both the corresponding trade name and all its active ingredients. Table 2 demonstrates examples of matching a drug mention to an active ingredient or a trade name.

Expert intervention for unmatched mentions

The mentions that did not match any active ingredient or trade name were listed with the number of posts associated with them and presented in descending order to a pharmacovigilance specialist. This list contained more than 800,000 mentions. Thus, the expert evaluated only unmatched mentions that were present in more than a significant number of posts (arbitrary fixed to 350) and decided if they should be considered in the study. Table 3 shows examples of such unmatched mentions and the corresponding actions suggested by the expert.

Table 1. Associations for the trade name ACTIFED® in Theriaque database.

Global trade name	Complete trade name	Active ingredient
ACTIFED®	Actifed cetirizine 10 mg CPR	Cetirizine
ACTIFED®	Actifed jour et nuit CPR	Diphenhydramine
		Paracetamol
		Pseudoephedrine
ACTIFED®	Actifed rhume CPR	Paracetamol
		Pseudoephedrine
		Triprolidine

Table 2. Examples of the table containing posts that match a trade name or an active ingredient.

Post ID	Drug mention as detected in user's post	Matched trade name in Theriaque	Matched (or corresponding) active ingredient(s) in Theriaque
04de3e064d5986283a	Actifed	ACTIFED®	Cetrazine
			Diphenhydramine
			Paracetamol
			Pseudoephedrine
			Triprolidine
04de3e064d5986283a	Paracetamol	–	Paracetamol
00002625588ff3f2bf55d	Clomid	CLOMID®	Clomiphene

Table 3. Examples of mentions unmatched to Theriaque® database that were presented to an expert.

Unmatched mention	Number of posts	Suggested action
Hormones	142,463	Ignore
Antidepresseur	68,318	Ignore
Antibio	62,501	Ambiguous: ignore
Vitamines	53,733	Ambiguous: ignore
Pim-pam	52,412	Ignore
Diane	43,266	Ambiguous: Ignore (frequently used as a first name)
Baclo	22,203	Replace by Baclofene
Morfine	1213	Replace by Morphine
Lévotyrox	925	Replace by Levothyrox

Statistical analysis

For statistics about number of posts per drug mentions, a post was counted only once for each trade name or active ingredient that it contained. For active ingredients, a post was counted if it contained an explicit mention of the active ingredient, or a product name that contains this active ingredient. Thus, a post that had a mention of a drug that contained several active ingredients was counted as one occurrence for each of these active ingredients. For example, if a post mentioned the trade name "Actifed®," this post was counted once for the trade name "Actifed®" and once for

each of its active ingredients "cetirizine," "diphenhydramine," "paracetamol," "pseudoephedrine" and "triprolidine" (cf. Table 3).

In order to evaluate the correlation between the most prescribed drugs and the most mentioned ones, we used Open Medic. This open dataset is provided and certified by the French Health Insurance System. It details the list of reimbursements performed for all deliveries of active ingredients in community pharmacies in France over a selected period. We considered the list of the most mentioned active ingredients in web forums during the year 2015 in order to compare them to the number of deliveries related to the active ingredients prescribed in France in the same year. To enable statistical correlation tests between these lists, only the 510 active ingredients in common in both lists were taken into account.

Another aspect of our study was to verify if the temporal trends of drug mentions in social media were influenced by medical events in traditional media. For this part of the study, we chose a set of drugs involved in mediatized "crises" in France:

- *Combined oral contraceptives (COCs).* In December 2012, a case of stroke in a young woman related to a third-generation COC was reported in a French newspaper. This event alarmed health professionals and national health authorities and opened a debate concerning the use and prescription of COCs.[25–27] To prepare data for the analysis relative to COCs, an expert identified the active ingredients that characterize the first, second, third or fourth generation of COCs.
- *Champix®.* In September 2006, Champix® (varenicline) was approved in Europe as an aid to smoking cessation treatment and marketed in France in February 2007. In December 2007, The European Medicines Agency (EMA) warned doctors and patients that suicide attempt cases were reported with this drug. The French Health Insurance stopped reimbursing Champix® in 2011 on the basis of an unfavorable reassessment of its benefit–risk balance. In 2017, Champix® was finally admitted again for reimbursement in France.
- *Baclofene.* In January 2012, the off-label use of baclofene at high dosages in alcohol-dependent patients was debated and received high media coverage.
- *Mirena®.* In 2017, women raised attention on the potential ADRs related to the use of Mirena®, a levonorgestrel-releasing intrauterine device, which was largely echoed in the media. Although the data used in our study were collected in 2015, we aimed to investigate if Mirena® was already discussed in web forums before the crisis.

Results

What are the most recurrent drugs in users' posts?

Table 4 presents the 50 most mentioned ingredients, and Table 5 presents the most mentioned trade names in the studied web forums. The high and correlated numbers of occurrences of sodium-based active ingredients draw our attention. We found out that this situation was the result of the high frequency of a trade name associated with the antacid drug "Gaviscon®," which contains a combination of sodium-based ingredients. As we counted all active ingredients associated with the detected trade names, this naturally led to high occurrences of sodium-based active ingredients in our calculations. In order to synthesize the results, we grouped these ingredients (in addition to calcium carbonate) in one line in Table 4 (line 23) with the maximum number of occurrences found with its ingredients.

Tables 4 and 5 show that the most mentioned drugs in web forums are related to pregnancy, contraception or ovulation stimulation. This finding was confirmed by the results in Table 6, which presents the proportions of posts associated with each Anatomical Therapeutic

Table 4. Top 50 active ingredients in web forums. To facilitate the analysis and improve the presentation of drug counts, raw results of active ingredients counts were reviewed by a pharmacist who grouped frequently associated ingredients and ignored synonyms of the same ingredient.

Order	Active ingredient	Number of posts	Percentage of the total number of posts
1	Clomiphene	101,033	6.21
2	Paracetamol	96,276	5.91
3	Ethinylestradiol	87,780	5.39
4	Insulin	74,002	4.55
5	Levonorgestrel	71,210	4.37
6	Phloroglucinol	62,013	3.81
7	Progesterone	57,648	3.54
8	Dydrogesterone	53,453	3.28
9	Cortisone	42,505	2.61
10	Magnesium	40,097	2.46
11	Baclofene	33,561	2.06
12	Morphine	33,388	2.05
13	Estradiol	33,356	2.05
14	Follitropin	32,266	1.98
15	Choriogonadotropin	31,538	1.94
16	Nicotine	30,519	1.87
17	Desogestrel	30,181	1.85
18	Copper	28,877	1.77
19	Isotretinoin	27,668	1.70
20	Alprazolam	24,362	1.50
21	Cyproterone	24,299	1.49
22	Folic acid	20,999	1.29
23	Set of gaviscon active ingredients: calcium carbonate, sodium carbonate, sodium bicarbonate, sodium alginate, sodium carbonate acid	20,954	1.29
24	Venlafaxine	20,739	1.27
25	Paroxetine	20,434	1.26
26	Ibuprofene	20,066	1.23
27	Misoprostol	19,191	1.18
28	Varenicline	19,168	1.18
29	Triptorelin	18,293	1.12
30	Drospirenone	18,205	1.12
31	Salbutamol	16,861	1.04
32	Acetylsalicylic acid	16,521	1.01
33	Escitalopram oxalate	16,108	0.99
34	Bromazepam	15,877	0.98
35	Fluoxetine	15,857	0.97
36	Glucose	15,512	0.95
37	Valproate	15,135	0.93
38	Finasteride	15,015	0.92
39	Gestodene	14,517	0.89
40	Metformin	14,081	0.87
41	Glycerol	13,382	0.82
42	Domperidone	12,619	0.78

(Continued)

Table 4. (continued)

Order	Active ingredient	Number of posts	Percentage of the total number of posts
43	Codeine	11,931	0.73
44	Caffeine	11,733	0.72
45	Interferon	11,128	0.68
46	Lysine acetylsalicylate	10,992	0.68
47	Clonazepam	10,189	0.63
48	Gonadotrophine human menopos(e)(ale)	9714	0.60
49	Tramadol	9584	0.59
50	Sertraline	9049	0.56
Total		1,459,886	89.68

Table 5. Top 50 trade names in web forums.

| Order	Product name	No. of posts	Simplified class proposal	Medical specialty (SPC therapeutic indications)
1	Clomid®	98,174	Ovulation induction	Gynecology
2	Spasfon®	61,595	Antispasmodic	Gastroenterology Gynecology
3	Doliprane®	53,645	Analgesic and antipyretic	Neurology
4	Duphaston®	53,141	Progestin	Gynecology
5	Puregon®	31,563	Ovulation induction	Gynecology
6	Ovitrelle®	31,538	Ovulation induction	Gynecology
7	Mirena®	23,713	Local contraceptive progestin	Gynecology
8	Xanax®	22,362	Anxiolytic	Psychiatry
9	Roaccutane®	21,856	Anti-acne treatment	Dermatology
10	Cerazette®	21,610	Oral contraceptive progestin	Gynecology
11	Effexor®	20,382	Antidepressant	Psychiatry
12	Utrogestan®	19,495	Progestin	Gynecology
13	Gaviscon®	19,434	Antacid	Gastroenterology
14	Champix®	19,132	Smoking cessation	Neurology
15	Cytotec®	19,018	Antiulcer agent prostaglandin	Gastroenterology
16	Decapeptyl®	18,207	Analog of luteinizing hormone releasing hormone (LHRH), GnRH agonist, hormonal therapy in cancer	Gynecology Cancer
17	Deroxat®	16,581	Antidepressant	Psychiatry
18	Aspirine®	16,474	Analgesic and antipyretic	Neurology
19	Acide Folique®	16,306	Vitamin	Hematology
20	Seroplex®	15,988	Antidepressant	Psychiatry
21	Lexomil®	15,177	Anxiolytic	Psychiatry
22	Prozac®	14,961	Antidepressant	Psychiatry
23	Jasmine®	12,466	Combined oral contraceptive pill	Gynecology
24	Diane 35®	12,295	Anti-acne treatment	Dermatology

(Continued)

Table 5. (continued)

| Order	Product name	No. of posts	Simplified class proposal	Medical specialty (SPC therapeutic indications)
25	Propecia®	11,885	Treatment of alopecia	Dermatology
26	Ventoline®	11,845	Anti-asthmatic	Pneumonology
27	Androcur®	11,749	Antiandrogen hormonal therapy in cancer	Gynecology Cancer
28	Trinordiol®	11,065	Combined oral contraceptive pill	Gynecology
29	Motilium®	10,795	Drug increasing gastrointestinal motility	Gastroenterology
30	Rivotril®	10,023	Anticonvulsant	Neurology
31	Lantus®	9977	Insulin	Endocrinology
32	Aspegic®	9892	Analgesic and antipyretic	Neurology
33	Menopur®	9714	Ovulation induction	Gynecology
34	Provames®	9644	Estrogen	Gynecology
35	Leeloo®	9376	Combined oral contraceptive pill	Gynecology
36	Dafalgan®	9149	Analgesic and antipyretic	Neurology
37	Advil®	9039	Analgesic non-steroidal anti-inflammatory drug	Neurology Rheumatology
38	Valium®	8667	Anxiolytic	Psychiatry
39	Aotal®	8617	Alcohol withdrawal	Neurology
40	Zoloft®	8494	Antidepressant	Psychiatry
41	Tarceva®	7958	Cytotoxic drug	Cancer
42	Atarax®	7955	Anxiolytic H1 antagonist	Psychiatry Allergology
43	Lutenyl®	7655	Progestin	Gynecology
44	Levothyrox®	7591	Thyroid hormone Replacement therapy	Endocrinology
45	Lovenox®	7532	Antithrombotic	Hematology
46	Stilnox®	7481	Sedative	Psychiatry
47	Microval®	7473	Contraceptif Oral Progestatif Seul	Gynecology
48	Minidril®	7093	Combined oral contraceptive pill	Gynecology
49	Depakine®	7047	Anticonvulsant	Neurology
50	Lamictal®	6967	Anticonvulsant	Neurology Psychiatry

Chemical (ATC) class. It is important to note that a post can be associated with several ATC classes depending on the types of mentions that it contains.

In addition to knowing the most used drug mentions in web forums, we wanted to verify if patients use mostly trade names or active ingredients when talking about drugs in web forums. Figure 2 shows the proportion of trade name mentions for the most frequent active ingredients. These results demonstrate that patients usually use trade names and not active ingredients when discussing about drugs on web forums. An exception to this observation was the frequent use of the mentions "Insulin" and "Cortisone" by forum users. We expect that users use these mentions to refer to drugs containing insulin-based preparations or variations of corticosteroids. In order to match trade names that correspond to "Insuline" and "Cortisone," we considered these two mentions as umbrella terms to any active ingredients whose ATC code, respectively, starts with "A10A" (Insulins and analogues) or "H02AB" (corticosteroids for systemic use).

Table 6. Posts' counts in web forums per ATC class.

ATC class	Number of mentions (unique per post)	% of total
G: Genito-urinary system and sex hormones	796,570	23.04%
A: Alimentary tract and metabolism	678,833	19.64%
N: Nervous system	659,954	19.09%
L: Antineoplastic and immunomodulating agents	220,056	6.37%
D: Dermatologicals	181,375	5.25%
R: Respiratory system	170,175	4.92%
S: Sensory organs	155,877	4.51%
B: Blood and blood forming organs	137,823	3.99%
M: Musculo-skeletal system	117,908	3.41%
J: Antiinfectives for systemic use	94,049	2.72%
C: Cardiovascular system	90,728	2.62%
H: Systemic hormonal preparations, excluding sex hormones and insulins	77,244	2.23%
V: Various	63,830	1.85%
P: Antiparasitic products, insecticides and repellents	12,275	0.36%
Total number of unique mentions	3,456,697	100%
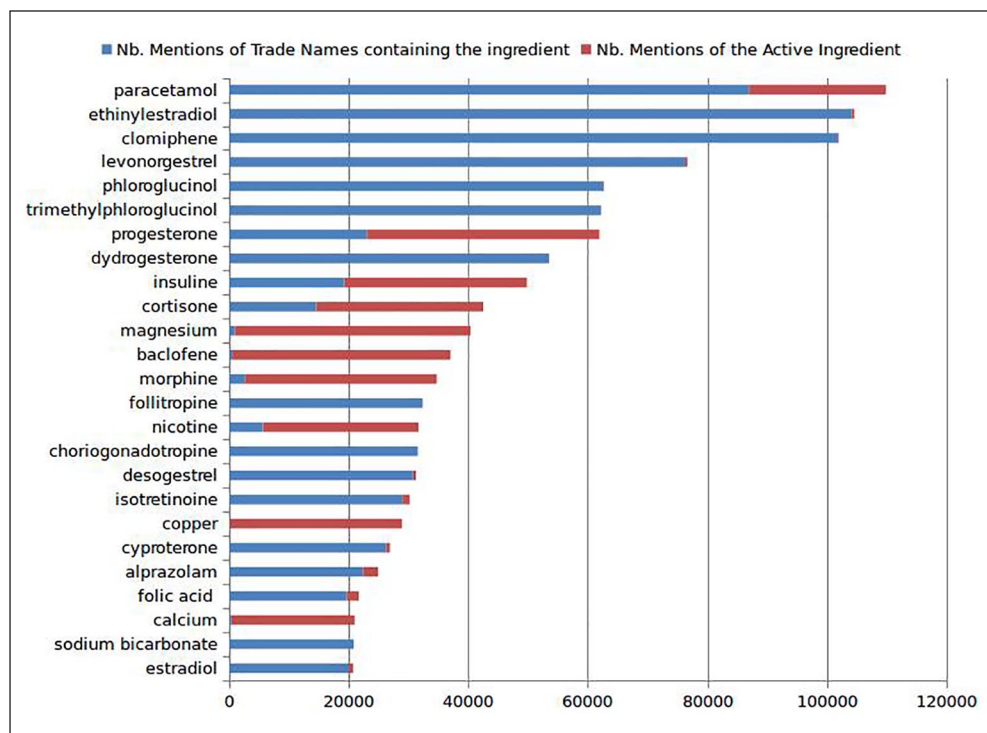
Figure 2. The proportion of using trade names versus active ingredients names in web forums.
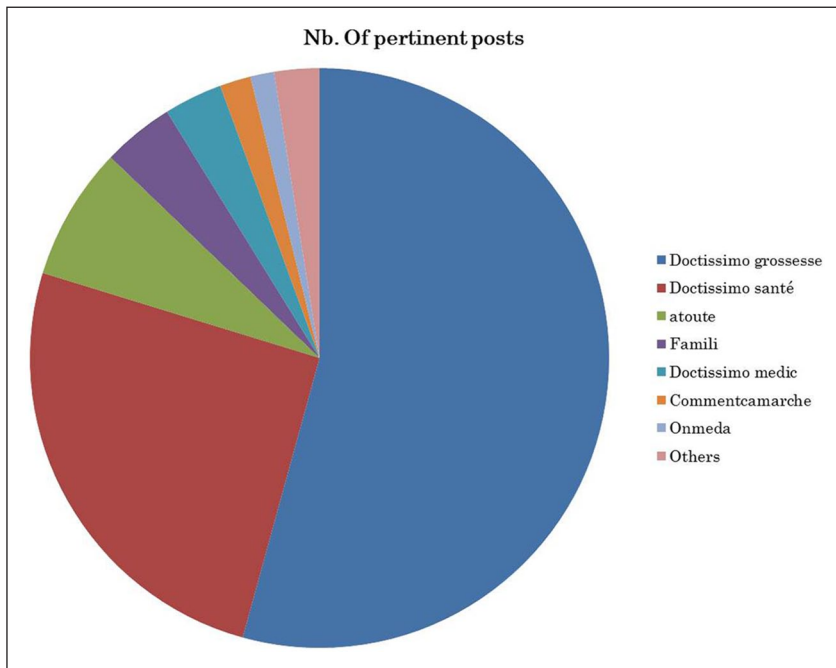
**Figure 3.** The distribution of posts containing active ingredients over web forums.

In this study, we noticed that Doctissimo forums were the most dominant in our study (Figure 3). Figure 4 depicts the distribution of the top 10 mentioned active ingredients in web forums (excluding cortisone and magnesium) over the dominant forums.

Figures 5 and 6 show the proportions of baclofene and ethinylestradiol, respectively, within the total number of users' posts for each of the studied forums. As expected, Figure 5 confirms the high contribution of two French forums dedicated to discussions on baclofene for drug mentions of baclofene, while several generic forums contained relatively important counts of the ethinylestradiol mention.

Does the evolution over time of a drug mention in web forums correspond to the events reported for this drug in the traditional media?

Figure 7 shows the evolution of "old" generation versus "new" generation of COCs. The third and fourth generations of COCs appeared more popular until March 2013. After this date, the tendency is inversed and posts mentioning old-generation pills, especially second generation, are more frequent in web forums.

The evolution of the number of posts mentioning Baclofene, Champix® and Mirena® from July 2007 to May 2015 in French forums is represented in Figure 8. The analysis of these mentions revealed that the increase of interest in Baclofene started in 2010. Conversely, mentions regarding Champix® that frequently appeared in web forums till April 2007 continuously declined through the following years. Overall, the number of mentions of Mirena® in web forums was stable, at the exception of a peak observed at the start of 2013.
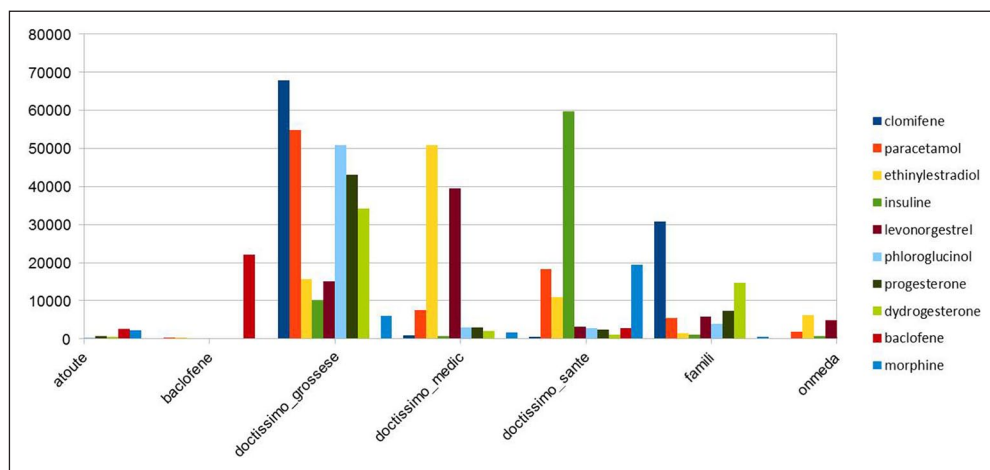
Figure 4. The distribution of the top active ingredients on the dominant forums in our study.

Do the most prescribed drugs in France correspond to the most mentioned drugs in web forums?

Table 7 shows a comparison of the top 10 active ingredients in 2015 in web forums with the top 10 active ingredients in Open Medic over the same year. Kendall-tau correlation coefficient value between the ranking of active ingredients found in the forums and in Open Medic was estimated at 0.31, which signifies a weak correlation between both rankings.[28] The scatterplot of the rankings shown in Figure 9 confirms this interpretation. In this figure, each point is an active ingredient whose rank in the Open Medic List is in the X axis, and rank in the Forums List is in the Y axis.

Discussion

Web forums represent an interesting source of information for drug-related studies. In this article, we presented a generic method for analyzing drug mentions that are detected in automatically annotated posts extracted from web forums. This method was applied to the case study of France. The use of French drugs and active ingredients names and a French database for drug matching (Theriaque®) does not decrease the generality of our method which could be applied to any web forum of any language as long as a convenient drug database is used for matching drug names.

In this study, most of the posts were dedicated to topics on contraception or pregnancy, including fertility and pregnancy development. From the top 50 mentioned drugs, a single active ingredient concerned men (finasteride), which was ranked #38 position among the most mentioned active ingredients in the forum. These results could be explained by the fact that women aged 18–45 years correspond to the most important population of web forum users.[29,30] Moreover, although our study detected the presence of drugs related to chronic diseases like diabetes mellitus (insulin/Lantus®, metformine), asthma (salbutamol/Ventoline®), hypothyroidism (Levothyrox®), epilepsy (clonazepam/Rivotril®, Depakine®, Lamictal®), depression (venlafaxine/Effexor®, paroxetine/Deroxat®, escitalopram/Seroplex®, fluoxetine/Prozac®, sertraline/Zoloft®) or cancer (Tarceva®, triptoreline/Decapeptyl®), most of the highly mentioned drugs concerned mainly young patients

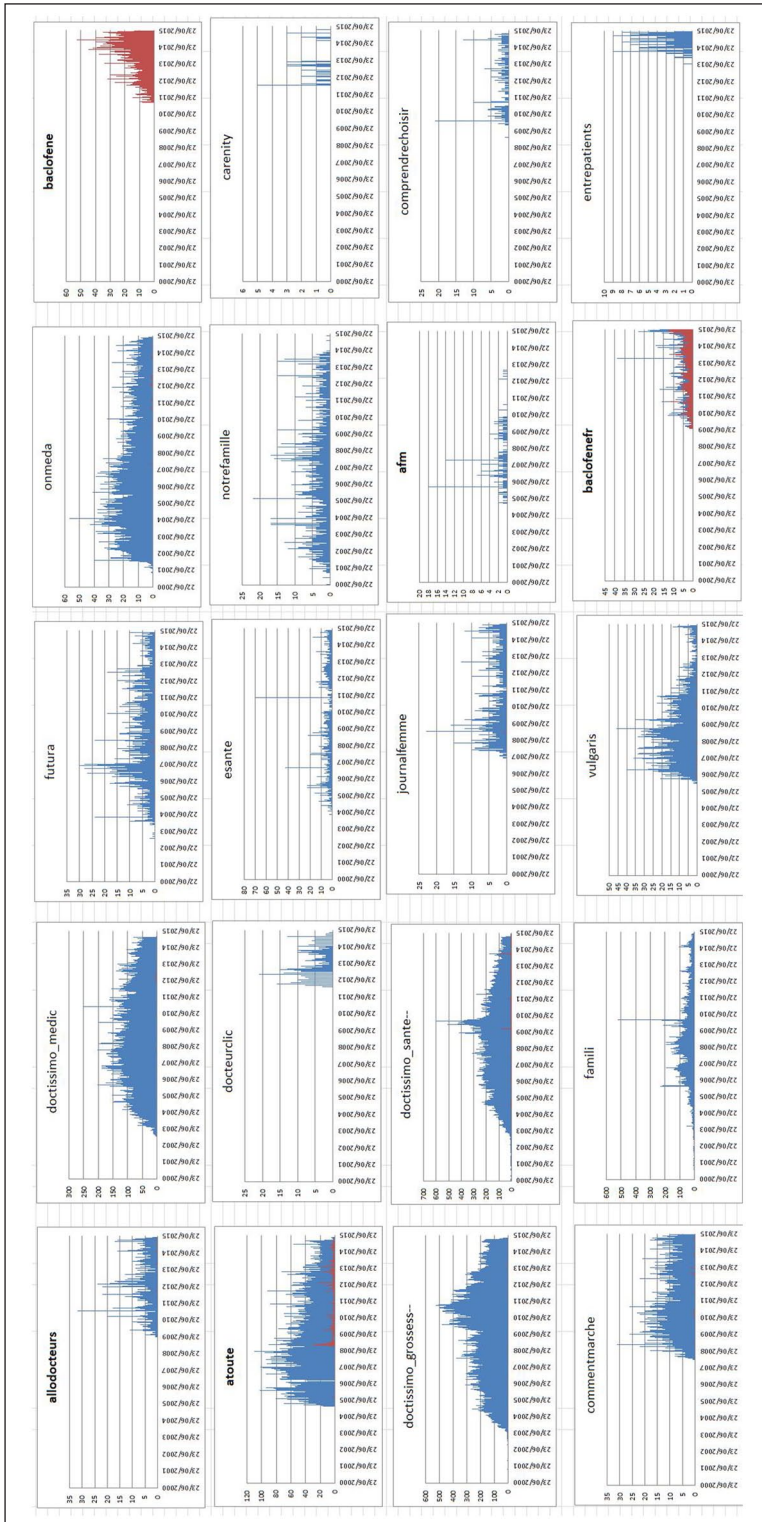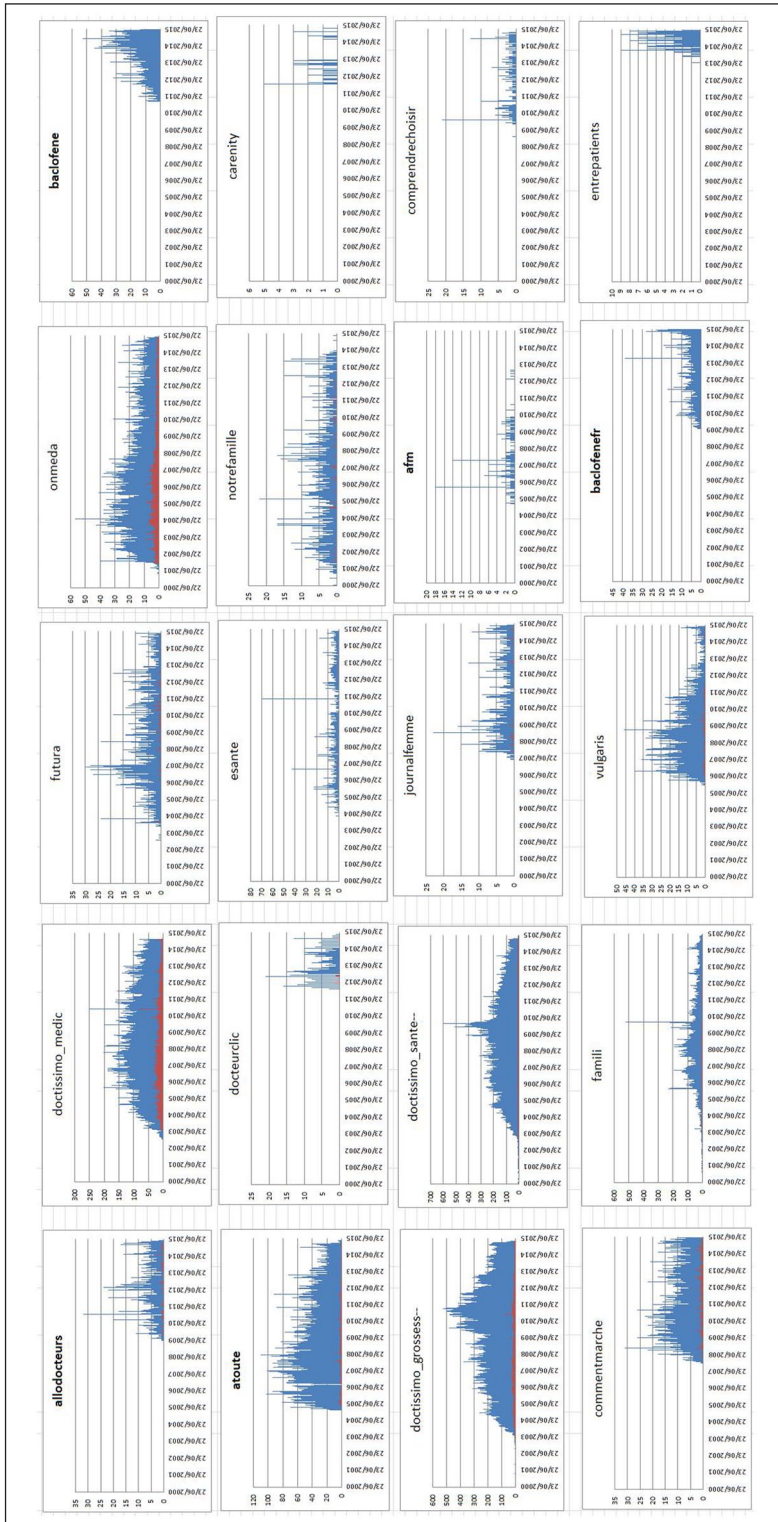**Figure 5.** Proportion of baclofene (red) to the total number of posts (blue) per forum.

Figure 6. Proportion of ethinylestradiol (red) to the total number of posts (blue) per forum.
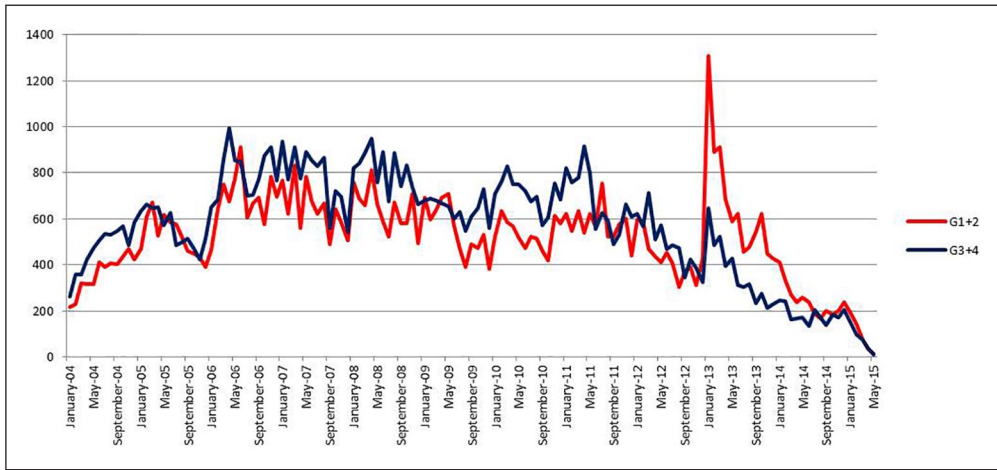
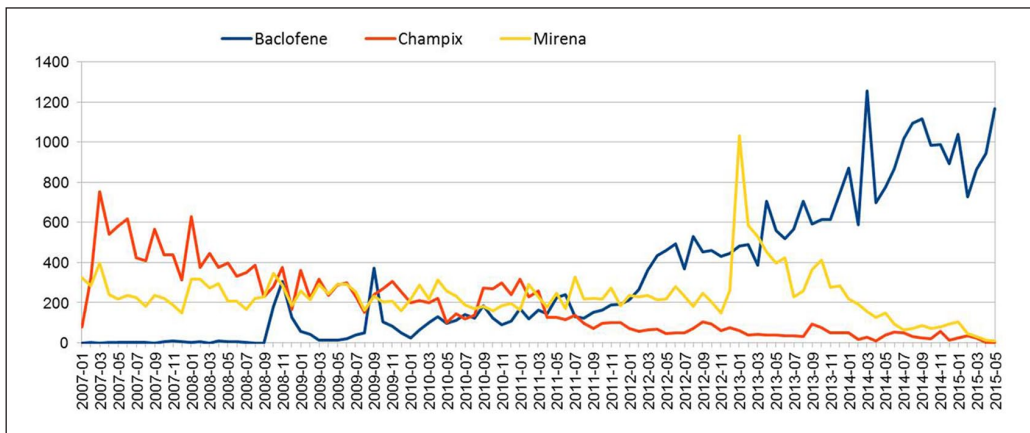**Figure 7.** Outcome of old versus new COC mentions in web forums over 11 years.

Figure 8. The use of Baclofene, Champix®, and Mirena® in web forums.

Table 7. Top 10 active ingredients mentions versus top prescribed active ingredients in 2015 in France.

Rank	OPEN MEDIC	Web Forum
1	Paracetamol	Baclofene
2	Ibuprofen	Clomiphene
3	Amoxicillin	Paracetamol
4	Cholecalciferol	Progesterone
5	Diclofenac	Phloroglucinol
6	Prednisolone	Dydrogesterone
7	Tixocortol	Magnesium
8	Phloroglucinol	Estradiol
9	Betamethasone	Levonorgestrel
10	Omeprazole	Morphine
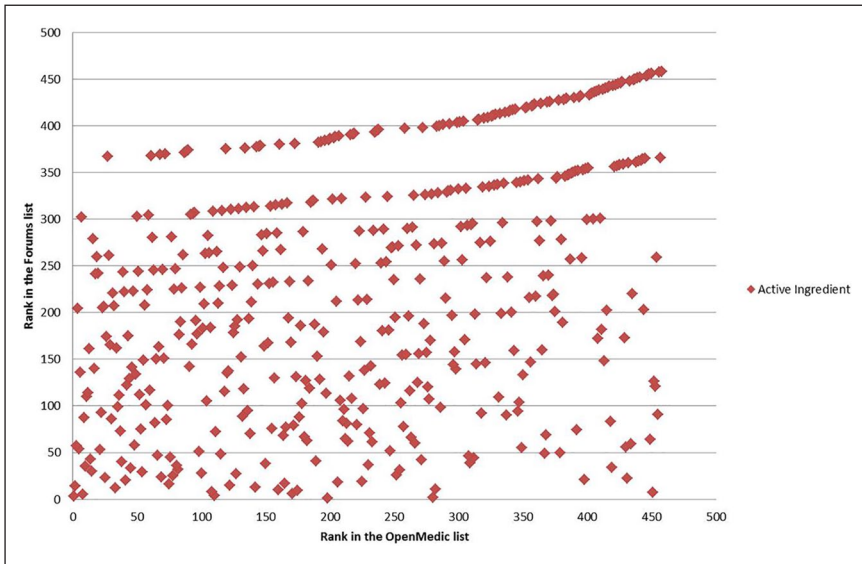
Figure 9. Scatterplot of the active ingredient ranks in the Forum's list and Open Medic List.

without serious pathologies. In the second position, we found painkiller medications (paracetamol, ibuprofen, morphine, etc.). From the top 50 active ingredients, psychotropic-related active ingredients (alprazolam, venlafaxine, paroxetine, escitalopram, bromazepam, fluoxetine, valproate, clonazepam and sertraline) were represented in 147,750 posts, while weaning-related active ingredients for tobacco and alcohol (baclofene, nicotine, varenicline) were represented by 83,248 posts.

These findings were confirmed by the high number of posts associated with the ATC classes G (Genito-urinary system and sex hormones), A (Alimentary tract and metabolism) and N (Nervous system). Our results are consistent with the findings of Wiley et al.[16] about the high frequency of discussions concerning nervous system, hormones and respiratory agents in social media.

There is a clear trend to use trade names in web forums (cf. Figure 2). However, some active ingredients such as morphine, copper and glucose were frequently used in our analyzed web forums. To understand the context in which these ingredients were used, we checked out the pathological conditions detected by machine learning associated with these three ingredients (Table 8). High frequencies of morphine and copper mentions (33,388 and 28,877 mentions, respectively) were particularly unexpected. Findings in Table 8 suggest that copper is probably used by women describing non-hormonal contraceptive devices, morphine is mentioned as a strong painkiller in serious diseases like cancer, while glucose was probably used by diabetic patients to describe their glycemic status.

For the analysis of the temporal trends of the case study on COCs, Figure 7 showed that at the beginning of 2013, discussions about the old COCs became more important than discussions about the new generations. Actually, the use of COCs in general had been significantly reduced after 2013, as many young women stopped using this contraception method after the mediatized information about the risks. However, this finding is difficult to confirm based on the reimbursement data we used in our study, as numerous COCs among those available in France are not reimbursed by the Health Insurance system.

For the case study of Champix®, a correspondence was observed between Champix® mentions in web forums and the dates of events related to its marketing and reimbursement (Figure 8).

Table 8. Pathologies associated with unexpected active ingredients often mentioned by users.

Morphine		Copper		Glucose	
Pathology	Frequency	Pathology	Frequency	Pathology	Frequency
Douleurs *pain*	13,324	Mal *ache*	3641	Diabète *diabetes*	1974
mal *ache*	7124	Douleurs *pain*	3531	mal *ache*	1464
fatigué *tired*	2723	saignements *bleedings*	2076	intolérance *intolerance*	901
souffre *suffering*	2046	acne *acne*	1327	diabète gestationnel *Gestational Diabetes*	756
pleuré *cried*	1776	prise de poids *weightgain*	1196	fatigue *Tiredness*	671
cancer *cancer*	1233	abondantes *abundant*	1112	diabétique *diabetic*	556
maladie *illness*	821	retard *late*	907	hypo *probably for hypoglycemia*	441
nausées *sickness*	776	Fatigue *Tiredness*	907	hyperglycémie *hyperglycemia*	437
dépendance *addiction*	671	infection *infection*	883	hypoglycémie *hypoglycemia*	424
Stress *Stress*	647	Nausées *Sickness*	799	stress *Stress*	423
angoisse *anxiety*	614	règles abondante *abundant period*	743	vomir *throwing up*	421

Unfortunately, we could not study the impact of the recent media interest on Champix® mentions as our study did not cover the year 2017. Nevertheless, we studied the case of Baclofene, which also constituted a particular case study in our analysis. In fact, Baclofene has been used for years as a muscle relaxant under the trade name of Lioresal®. After the media coverage concerning its new use and potential interest for alcohol abstinence, it became more common on web forums. Indeed, a clear increase in mentions was observable since January 2012 for this drug, which meets the date at which French authorities authorized its prescribing for alcoholic dependence treatment. Furthermore, we noticed an increasing interest in baclofene in 2008, which could be related to the release of a book about the efficiency of baclofene in alcohol withdrawal. Finally, for the Mirena® temporal case study, Figure 8 shows the high number of posts mentioning Mirena® in January 2013. We found out that at this date, a report comparing the uterine perforation rate between Mirena® and Copper-based IUD was published, which could have influenced the discussions in web forums at that period. From these case studies, we conclude that temporal trends of drug mentions seemed to be influenced by events widely reported in traditional media. One positive consequence of this influence is that we can collect in near real-time descriptions of ADRs by patients. Such a process is much expensive and time consuming when using traditional reporting procedures.

Another important finding of our study is that the most mentioned drugs in web forums are not necessarily the most prescribed. For example, although Levothyrox® is highly prescribed in France (Levothyroxine Sodique is at position 33 in 2015 in Open Medic), this drug does not appear in the list of top 50 mentioned drugs in web forums over the same studied period. It is important to keep

in mind that the posts included in our study were all published before the change of the composition of Levothyrox® in France, which happened in March 2017[31] and was widely covered by media. Indeed several adverse reactions were reported by patients after lactulose was replaced by mannitol as an excipient. Our study showed out that the most mentioned drug in the selected web forums is clomiphene, which is an ovulation stimulant for women with infertility problems. This finding confirms the results of a previous study.[32]

Limitations

The methods used in this article are generic, but the application proposed in this work was limited to 21 selected forums in French language for a period that ends in 2015. The domination of some big forums biased our results by representing mostly young women and pregnancy-related topics. In our case study, matching mentions in web forums to the Theriaque® drug database led to counting ingredients that are not relevant to the study, such as glucose. Another limitation was the disambiguation of misspelled drug mentions in users' discussions, which required an expert intervention. Finally, using social media to analyze patients' reactions excludes patients who do not have access to the Internet or who are not familiar with the use of online discussions.

Perspectives

Extending the period of our study to include recent posts and English forums will be the next follow-up for our work. This extension will allow us to analyze the echo of recent events in social media and the influence of cultural and linguistic particularities on the results. Considering the bias regarding the age, sex and health context of the studied population, any future work should be careful about the crucial choice of forums to consider. Furthermore, in order to minimize expert intervention, it is important for future work to consider a list of drug ingredients that should not be taken into account when counting active ingredients. Another perspective to this work is to consider a system that analyzes drug mentions in web forums "On the fly." This procedure will be part of the current PHARES project that aims to establish a pipeline for detecting ADRs and off-label uses in web forums.

Finally, this article focused only on how patients mention drugs in social media. A future work will focus on extracting causality relation between drugs and adverse events mentioned on social media posts. The development and the evaluation of a robust machine learning approach are necessary to correctly detect ADRs.

ORCID iDs

Bissan Audeh   https://orcid.org/0000-0001-8550-8724
Marie-Noëlle Beyens   https://orcid.org/0000-0002-5831-8957

References

1. Ghosh R and Lewis D. Aims and approaches of Web-RADR: a consortium ensuring reliable ADR reporting via mobile devices and new insights from social media. *Expert Opin Drug Saf* 2015; 14(12): 1845–1853.

2. Golder S, Norman G and Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *Br J Clin Pharmacol* 2015; 80(4): 878–888.

3. Micoulaud-Franchi JA. One step more toward pharmacovigilance 2.0. *Presse Med* 2011; 40(9 pt 1): 790–792.

4. Bagheri H, Lacroix I, Guitton E, et al. Cyberpharmacovigilance: what is the usefulness of the social networks in pharmacovigilance? *Therapie* 2016; 71(2): 235–239.

5. Salathe M. Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. *J Infect Dis* 2016; 214(suppl. 4): S399–S403.

6. Brown JC, Tuuri RE, Akhter S, et al. Lacerations and embedded needles caused by epinephrine autoinjector use in children. *Ann Emerg Med* 2016; 67(3): 307.e8–315.e8.

7. Vaughan Sarrazin MS, Cram P, Mazur A, et al. Patient perspectives of dabigatran: analysis of online discussion forums. *Patient* 2014; 7(1): 47–54.

8. Risson V, Saini D, Bonzani I, et al. Patterns of treatment switching in multiple sclerosis therapies in US patients active on social media: application of social media content analysis to health outcomes research. *J Med Internet Res* 2016; 18(3): e62.

9. Palosse-Cantaloube L, Lacroix I, Rousseau V, et al. Analysis of chats on French internet forums about drugs and pregnancy. *Pharmacoepidemiol Drug Saf* 2014; 23(12): 1330–1333.

10. Kazemi DM, Borsari B, Levine MJ, et al. Systematic review of surveillance by social media platforms for illicit drug use. *J Public Health* 2017; 39(4): 763–776.

11. Anderson LS, Bell HG, Gilbert M, et al. Using social listening data to monitor misuse and nonmedical use of bupropion: a content analysis. *JMIR Public Health Surveill* 2017; 3: e6.

12. Abou Taam M, Rossard C, Cantaloube L, et al. Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator®) withdrawal in France. *J Clin Pharm Ther* 2014; 39(1): 53–55.

13. Rothman M, Gnanaskathy A, Wicks P, et al. Can we use social media to support content validity of patient-reported outcome instruments in medical product development? *Value Health* 2015; 18: 1–4.

14. Lardon J, Abdellaoui R, Bellet F, et al. Adverse drug reaction identification and extraction in social media: a scoping review. *J Med Internet Res* 2015; 17(7): e171.

15. Ravoire S, Lang M and Perrin E. Advantages and limitations of online communities of patients for research on health products. *Therapie* 2017; 72(1): 135–143.

16. Wiley MT, Jin C, Hristidis V, et al. Pharmaceutical drugs chatter on Online Social Networks. *J Biomed Inform* 2014; 49: 245–254.

17. Carbonell P, Mayer MA and Bravo A. Exploring brand-name drug mentions on Twitter for pharmacovigilance. *Stud Health Technol Inform* 2015; 210: 55–59.

18. Mahroum N, Bragazzi NL, Brigo F, et al. Capturing public interest toward new tools for controlling human immunodeficiency virus (HIV) infection exploiting data from Google Trends. *Health Informatics J*. Epub ahead of print 11 April 2018. DOI: 10.1177/1460458218766573.

19. Audeh B, Beigbeder M, Zimmermann A, et al. Vigi4Med scraper: a framework for web forum structured data extraction and semantic representation. *PLoS One* 2017; 12(1): e0169658.

20. Morlane-Hondère F, Grouin C and Zweigenbaum P. Identification of drug-related medical conditions in social media. In: Proceedings of the international conference on language resources and evaluation (LREC'2016), Portorož, 23–28 May 2016.

21. Karapetiantz P, Bellet F, Audeh B, et al. Descriptions of adverse drug reactions are less informative in forums than in the French pharmacovigilance database but provide more unexpected reactions. *Front Pharmacol* 2018; 9: 439.

22. Lafferty JD, McCallum A and Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning, Williams College, Williamstown, MA, 28 June–1 July 2001.

23. Chang CC and Lin CJ. LIBSVM: a library for support vector machines. *ACM T Intel Syst Tec* 2011; 2(3): 27.

24. Husson MC. Theriaque: independent-drug database for good use of drugs by health practitioners. *Ann Pharm Fr* 2008; 66(5–6): 268–277.

25. Arie S. French doctors are told to restrict use of third and fourth generation oral contraceptives. *BMJ* 2013; 346: f121.
26. Hugon-Rodin J, Gompel A and Plu-Bureau G. Epidemiology of hormonal contraceptives-related venous thromboembolism. *Eur J Endocrinol* 2014; 171: R221–R230.
27. Emmerich J, Thomassin C and Zureik M. Contraceptive pills and thrombosis: effects of the French crisis on prescriptions and consequences for medicine agencies. *J Thromb Haemost* 2014; 12(9): 1388–1390.
28. Cohen J. A power primer. *Psychol Bull* 1992; 112: 155–159.
29. Sadah SA, Shahbazi M, Wiley MT, et al. A study of the demographics of web-based health-related social media users. *J Med Internet Res* 2015; 17(8): e194.
30. Duggan M and Brenner J. *The demographics of social media users – 2012*. Washington, DC: Pew Research Center, 2013.
31. ANSM. Levothyrox (lévothyroxine): changement de formule et de couleur des boîtes—Point d'Information, http://ansm.sante.fr/S-informer/Points-d-information-Points-d-information/Levothyrox -levothyroxine-changement-de-formule-et-de-couleur-des-boites-Point-d-Information
32. Sadah SA, Shahbazi M, Wiley MT, et al. Demographic-based content analysis of web-based health-related social media. *J Med Internet Res* 2016; 18(6): e148.