Publications scientifiques

Kap Code

FROM DATA TO HEALTH

L'équipe Kap Code

Stéphane Schück
PRÉSIDENT

Nathalie Texier
VICE-PRÉSIDENT

Adel Mebarki
DIRECTEUR GÉNÉRAL

Paméla Voillot
RESPONSABLE ÉTUDES & OPÉRATIONS

Pierre Foulquié
RESPONSABLE DATA-SCIENCE

Simon Renner
RESPONSABLE MÉDICAL

EFMI EUROPEAN FEDERATION for MEDICAL INFORMATICS

2015

Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in adverse drug reaction mining

Agnes Lillo-le-louet, Sandrine Katsahian, Erica Simond Moreau, Damien Leprovost, Jeremy Lardon, Cedric Bousquet, Gaétan Kerdelhué, Redhouane Abdellaoui, Nathalie Texier, Anita Burgun, Abdelali Boussadi, and Carole Faviez

Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in Adverse Drug Reaction Mining

Agnes LILLO-LE-LOUET, Sandrine KATSAHIAN, Erica SIMOND MOREAU, Damien LEPROVOST, Jeremy LARDON, Cedric BOUSQUET, Gaetan KERDELHUE, Redhouane ABDELLAOUI, Nathalie TEXIER, Anita BURGUN, Abdelali BOUSSADI, Carole FAVIEZ

CONTEXT

Suspected adverse drug reactions (ADR) reported by patients through social media can be a complementary tool to already existing ADRs signal detection processes. However, several studies have shown that the quality of medical information published online varies drastically whatever the health topic addressed.

Objective: The aim of this study is to use an existing rating tool on a set of social network web sites in order to assess the capabilities of these tools to guide experts for selecting the most adapted social network web site to mine ADRs.

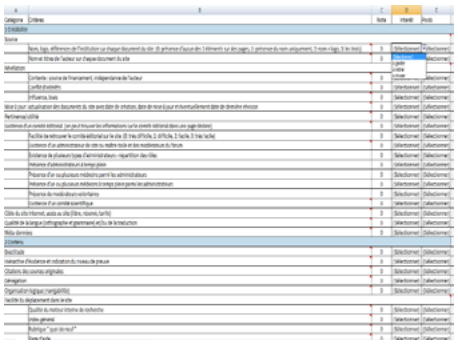
METHOD

Step 1: Forums and Internet social networks selection according to three criteria

- **The number of visits**: estimated through the Cismef web site (www.cismef.org), a catalog and index of French language health resources on the Internet
- **The notoriety of the forum**: estimated through Google, Alexa (http://www.alexa.com/) and Yoovi (http://www.yoovi.com/)
- **The number of messages posted in relation with health and drug therapy**: evaluated using the number of messages per day through the 1001 forums website (one of the larger French forum's indexer on the Internet)

Step 2: Data collection

6 Experts rated all websites using the following grid



Step 3: Data analysis

Scores are calculated for each websites aand each raters. "Not found" values are treated in three different ways:

- Omitted
- Correspond to the worst score of 0.
- Correspond to a "penalty" of "-2"

Agreement between experts assessed using weighted kappa pooled using mean in order to take into consideration the distance between the scores.

RESULTS

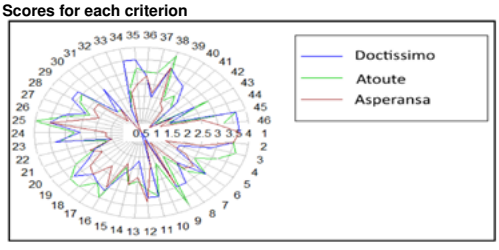
Websites selected

- **Asperansa**: (http://www.asperansa.org) Internet forum dedicated to adolescents, young adults with autism and high level of Asperger syndrome and their parents.
- **Atoute**: (http://www.atoute.org/) generalist web sites and forums dedicated to health topics
- **Doctissimo**: (http://www.doctissimo.fr/) : generalist web sites and forums dedicated to health topics

Three scores computed for each Internet social network

		Doctissimo	Atoute	Asperansa
score A	mean	80%	78%	73%
	median	80%	79%	75%
score B	mean	72%	71%	66%
	median	73%	70%	62%
score C	mean	71%	69%	62%
	median	73%	69%	58%

Scores for each criterion



Kappa's coefficients

	Doctissimo	Atoute	Asperansa
Where « Not found » values are omitted	0.03870397	0.06779932	0.05367282
Where « Not found » values correspond to 0	0.08845810	0.09893061	0.07135463

DISCUSSION

The results show that if some criteria are quite consensual, others are randomizing the results. The personal opinion of the expert seems to have a major impact, undermining the relevance of the criterion. In order to maximize the agreement between experts, it appears we need to make a rigorous selection of criteria. Our future work is to collect results given by this evaluation grid and proposes a new scoring tool adapted to Internet social networks assessment.

* The ADR prism project is financed by the DGE and territorial collectivities under the 16th FUI's call for project.

EFMI EUROPEAN FEDERATION for MEDICAL INFORMATICS
2015

Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in adverse drug reaction mining

Sandrine Katsahian, Erica Simond Moreau, Damien Leprovost, Jeremy Lardon, Cedric Bousquet, Gaétan Kerdelhué, Redhouane Abdellaoui, Nathalie Texier, Anita Burgun, Abdelali Boussadi, and Carole Faviez

Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in Adverse Drug Reaction Mining

Sandrine Katsahian[a,c,d,h], Erica Simond Moreau[h], Damien Leprovost[b,c,f], Jeremy Lardon[b,e], Cedric Bousquet[b,e], Gaétan Kerdelhué[g] , Redhouane Abdellaoui[i], Nathalie Texier[i], Anita Burgun[a,c,d,h], Abdelali Boussadi[a,c,d,h] and Carole Faviez[i]

[a] INSERM, UMR_S 1138, équipe 22, Centre de Recherche des Cordeliers, F-75006, Paris, France

[b] INSERM, U1142, LIMICS, F-75006, Paris, France

[c] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1138, Centre de Recherche des Cordeliers, F-75006, Paris, France

[d] Université Paris Descartes, Sorbonne Paris Cité, UMR_S 1138, Centre de Recherche des Cordeliers, F-75006, Paris, France

[e] University of Saint Etienne, Department of Public Health and Medical Informatics, Saint-Etienne, France.

[f] Université Paris 13, Sorbonne Paris Cité, F-93430, Villetaneuse, France.

[g] CISMeF, Rouen University Hospital, Cour Leschevin, Rouen, Cedex, France.

[h] AP-HP, HEGP, Département d'Informatique Hospitalière, Paris, France.

[i] Kappa Santé, Paris, France.

Abstract. Background and objectives: Suspected adverse drug reactions (ADR) reported by patients through social media can be a complementary tool to already existing ADRs signal detection processes. However, several studies have shown that the quality of medical information published online varies drastically whatever the health topic addressed. The aim of this study is to use an existing rating tool on a set of social network web sites in order to assess the capabilities of these tools to guide experts for selecting the most adapted social network web site to mine ADRs. **Methods:** First, we reviewed and rated 132 Internet forums and social networks according to three major criteria: the number of visits, the notoriety of the forum and the number of messages posted in relation with health and drug therapy. Second, the pharmacist reviewed the topic-oriented message boards with a small number of drug names to ensure that they were not off topic. Six experts have been chosen to assess the selected internet forums using a French scoring tool: Net scoring. Three different scores and the agreement between experts according to each set of scores using weighted kappa pooled using mean have been computed. **Results:** Three internet forums were chosen at the end of the selection step. Some criteria get high score (scores 3-4) no matter the website evaluated like accessibility (45-46) or design (34-36), at the opposite some criteria always have bad scores like quantitative (40-42) and ethical aspect (43-44), hyperlinks actualization (30-33). Kappa were positives but very small which corresponds to a weak agreement between experts. **Conclusion:** The personal opinion of the expert seems to have a major impact, undermining the relevance of the criterion. Our future work is to collect results given by this evaluation grid and proposes a new scoring tool for Internet social networks assessment.

Keywords. Adverse drug reaction, Internet, quality measurement, evaluation

Kap Code - Book de publications scientifiques

JMIR Publications
The Leading eHealth Publisher

2015

Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review

Jérémy Lardon, PhD; Redhouane Abdellaoui, MS; Florelle Bellet, PharmD; Hadyl Asfari, PharmD; Julien Souvignet, MS; Nathalie Texier, PharmD; Marie-Christine Jaulent, PhD; Marie-Noëlle Beyens, MD; Anita Burgun, MD, PhD; Cédric Bousquet, PharmD, PhD

Review

Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review

Jérémy Lardon[1,2*], PhD; Redhouane Abdellaoui[3,4*], MS; Florelle Bellet[5], PharmD; Hadyl Asfari[2,6], PharmD; Julien Souvignet[2,6], MS; Nathalie Texier[4], PharmD; Marie-Christine Jaulent[6], PhD; Marie-Noëlle Beyens[7], MD; Anita Burgun[8,9], MD, PhD; Cédric Bousquet[2,6], PharmD, PhD

[1]Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), (Unité Mixte de Recherche en Santé, UMR_S 1142), F-93430, Villetaneuse, France, Sorbonne Universités, University of Pierre and Marie Curie (UPMC) Université Paris 06, Unité Mixte de Recherche en Santé (UMR_S) 1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Institut National de la Santé et de la Recherche Médicale (INSERM), U1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Paris, France

[2]Service de Santé Publique et de l'Information Médicale (SSPIM), Department of Public Health and Medical Informatics, Centre Hospitalier Universitaire (CHU) University Hospital of Saint Etienne, Saint-Etienne, France

[3]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1138, équipe 22, Centre de Recherche des Cordeliers, Université Paris Descartes, Sorbonne Paris Cité, F-75006, Paris, France

[4]Kappa Santé, Paris, France

[5]Centre de Pharmacovigilance, Centre Hospitalier Universitaire (CHU) University Hospital of Saint Etienne, Saint-Etienne, France

[6]Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), (Unité Mixte de Recherche en Santé, UMR_S 1142), F-93430, Villetaneuse, France, Sorbonne Universités, University of Pierre and Marie Curie (UPMC) Université Paris 06, Unité Mixte de Recherche en Santé (UMR_S) 1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Paris,, Institut National de la Santé et de la Recherche Médicale (INSERM), U1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Paris, France

[7]Centre de Pharmacovigilance, Centre Hospitalier Universitaire (CHU) University Hospital of Saint Etienne, Saint-Etienne, France

[8]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1138, équipe 22, Centre de Recherche des Cordeliers, Université Paris Descartes, Sorbonne Paris Cité, F-75006, Paris, France

[9]Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Européen Georges-Pompidou (HEGP), Department of Medical Informatics, Paris, France

[*]these authors contributed equally

Corresponding Author:
Jérémy Lardon, PhD
Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), (Unité Mixte de Recherche en Santé, UMR_S 1142), F-93430, Villetaneuse, France
Sorbonne Universités, University of Pierre and Marie Curie (UPMC) Université Paris 06, Unité Mixte de Recherche en Santé (UMR_S) 1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006
Institut National de la Santé et de la Recherche Médicale (INSERM), U1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006
Campus des Cordeliers, Institut National de la Santé et de la Recherche Médicale (INSERM) U 1142 - Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS) Esc D - 2ème étage
15 rue de l'École de Médecine
Paris, 75006
France
Phone: 33 144279190
Fax: 33 144279192
Email: jeremy.lardon@chu-st-etienne.fr

Abstract

Background: The underreporting of adverse drug reactions (ADRs) through traditional reporting channels is a limitation in the efficiency of the current pharmacovigilance system. Patients' experiences with drugs that they report on social media represent a new source of data that may have some value in postmarketing safety surveillance.

Objective: A scoping review was undertaken to explore the breadth of evidence about the use of social media as a new source of knowledge for pharmacovigilance.

XSL•FO

2016

Web-based signal using medical forums in france from 2005-2015

Marie-Laure Kürzinger MSc, Nathalie Texier PharmD, Stéphane Schuck MD, MSc, Carole Faviez MSc, Thierry Deliens MSc, Ling Zhang MSc, Stéphanie Tcherny-Lessenot MD, MSc, Juhaeri Juhaeri PhD, Susan Welsh MD.

Web-based signal detection using medical forums in France from 2005-2015

Abstract #750716

Marie-Laure Kürzinger MSc(1), Nathalie Texier PharmD(2), Stéphane Schuck MD, MSc(2), Carole Faviez MSc(2), Thierry Deliens MSc(3), Ling Zhang MSc(4), Stéphanie Tcherny-Lessenot MD, MSc(1), Juhaeri Juhaeri PhD(4), Susan Welsh MD(4).

1 Sanofi, Global Pharmacovigilance and Epidemiology, Chilly-Mazarin, France
2 Kappa Santé, France
3 Sanofi, Information Technology & Solutions, Chilly-Mazarin, France
4 Sanofi, Global Pharmacovigilance and Epidemiology, Bridgewater, New Jersey, United States

Background

- ✓ Traditional signal detection methods in pharmacovigilance are based on individual case safety reports
- ✓ The use of web-based data (such as social media) is emerging among regulators, industry and academia.
- ✓ The strength of web-based data relies on their real time availability allowing early signal detection.

Objectives

- ✓ The study aims at assessing the ability of identifying early signals from web-based patient's medical forums in France on 3 products over the last 11 years:
  - To compare the detected signals from the patient's medical forums in France to signals detected in VigiBase® (cumulative)
  - To evaluate time difference in the detection of signals from the patient forums and from VigiBase®

Methods

- ✓ **Data sources**
  - Data were extracted from the Detec't database, composed of messages posted on patients' forums between 2005 and 2015
  - 13 French patients forums: Atoute, Doctissimo, E-sante, SanteMedecine, Onmeda, Futura sciences, Psoriasis, AlarmAsso, Morphee, Albi, Renaloo, Allodocteurs, Forum Sclérose en plaques
  - WHO adverse events reporting system (VigiBase®)
- ✓ **8 disproportionality methods/definitions**:
  - EB05 ≥2; EBGM ≥2; EBGM ≥4; PRR ≥2, N ≥3, Chi-square ≥4; PRR025 ≥1; ROR025≥1; IC025 > 0; RFET p-value ≤ 0.05
- ✓ **Three single agent drugs**
  - teriflunomide, insulin glargine, zolpidem
- ✓ Comparison of signals detected from the forums to signals detected in VigiBase® was done by describing the **sensitivity**, **specificity**, **positive predictive value (PPV)**, **negative predictive value (NPV)**, **accuracy and ROC curves**.
- ✓ For expected signals, time difference in months between the detection date of signals from the patients' forums and date of signals from VigiBase® was provided.

Results

Figure 1. Number of intake messages for the three drugs

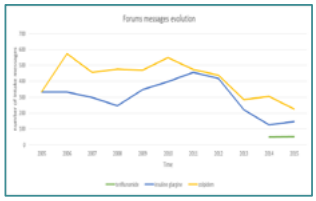


Figure 3. Number of drug-event pairs in VigiBase® and in the Forums

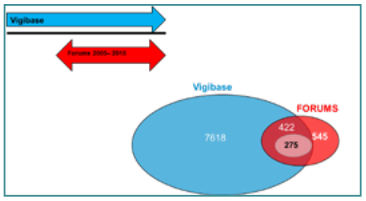


Figure 2. Extraction & methodology

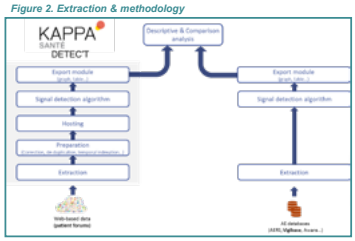
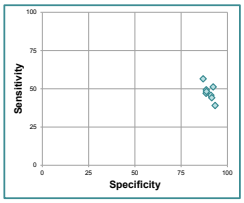


Table 1. Comparison between patient's forums and VigiBase® signals (all drugs together

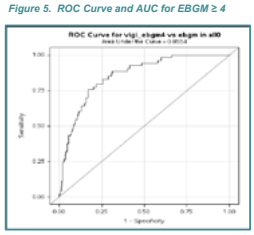
Definition	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
EB05 ≥2	29.06	95.53	62.5	84.03	81.99
EBGM ≥2	48.24	89.28	62.5	82.33	78.19
EBGM ≥4	39.62	94.57	51.21	91.6	87.67
PRR ≥2, N ≥3, Chi-square ≥4	31.93	94.05	67.85	77.86	76.54
Lower 95% CI of PRR ≥1	37.34	87.5	64.13	70	68.72
Lower 95% CI of ROR ≥1	36.96	87.93	66.3	68.48	68
IC025 > 0	33.33	94.18	75.38	72.54	72.98
RFET: p-value ≤ 0.05	50.6	86.09	68.1	74.83	72.98

Figure 4. Specificity and Sensitivity whatever the methods



- ✓ The number of posts from the patients' forums containing teriflunomide, insulin glargine, zolpidem and with the mention of intake were respectively over the 2005-2015 period: 102, 3326, 4584 (Fig. 1)
- ✓ Comparison analysis shows that according to metrics used, the sensitivity ranges from 29.1 to 50.6%, the specificity from 86.1 to 95.5%, the PPV from 51.2 to 75.4%, the NPV from 68.5 to 91.6% and the accuracy from 68 to 87.7% (Table 1).
- ✓ The AUC reaches 0.85 when considering EBGM ≥ 4 (Fig. 5)
- ✓ The time analysis shows that 30% of the signals are detected earlier (up to 128 months earlier) in the forums than in VigiBase® and 20% are detected at the same time but are available earlier.

Figure 5. ROC Curve and AUC for EBGM ≥ 4



Conclusions

- ✓ Web-based signal detection using patients' medical forums in France appears to have good performance compared to signals detected in traditional sources. Half of the web-based signals from the French forums are detected or available earlier than in the traditional data sources. Those signals relate to serious medical events as well as patients related symptoms (stress, hunger …).
- ✓ These results indicate that using patients' medical forums should be considered as a complementary source of data to traditional sources allowing signals to be detected earlier and thus ensuring increased safety of the patients.
- ✓ Further enhancements are needed to investigate the reliability and validation of patient's medial forums worldwide.

Conflict of Interest Disclosure

This study was entirely funded by Sanofi. Authors are employees of Sanofi and the other authors are employees at Kappa Santé which was the company doing the data processing and analysis.

KAPPA SANTÉ

SANOFI Research & Development

2016

Comparison of Web-based signal detection using medical forums data in France from 2005-2015 with signals from Vigibase®

Schück S, Kürzinger ML, Abdellaoui R, Texier N, Pouget J, Faviez C, Zhang L, Tcherny-Lessenot S, Juhaeri J, Welsh S

Comparison of Web-based signal detection using medical forums data in France from 2005-2015 with signals from Vigibase®

Schück S[1], Kürzinger ML[2], Abdellaoui R[1], Texier N[1], Pouget J[3], Faviez C[1], Zhang L[4], Tcherny-Lessenot S[2], Juhaeri J[4], Welsh S[4]

[1]Kappa Santé, Paris, France
[2]Global Pharmacovigilance and Epidemiology, Sanofi, Chilly-Mazarin, France
[3]Information Technology Solutions, Sanofi, Lyon, France
[4]Global Pharmacovigilance and Epidemiology, Sanofi, Bridgewater, NJ, USA

Background

Post-marketing drug surveillance is largely based on signals found from spontaneous reports from patients and health care providers. Rare adverse drug reactions and adverse events (AEs) which may develop after long-term exposure to a drug or from drug interactions may be missed. It has been proposed by the FDA and others [1, 2, 3, 4, 5] that web-based data could be mined as a resource to detect latent signals associated with adverse drug reactions. While traditional signal detection methods in PV are based on individual case safety reports, the use of Web-based data (such as, query logs and social media) is emerging among regulators (FDA and EMA), industry and academia. The strength of Web-based data relies on their real time availability allowing early signal detection - 6 months to 1 year earlier than traditional data sources used for signal detection (spontaneous reporting system, electronic medical records, and claims databases).

Objectives

While traditional signal detection methods in pharmacovigilance are based on individual case safety reports, the use of social media data is emerging among regulators and industry. This study aims at assessing the reliability of signals from web-based patient's forums in France on 3 products (*teriflunomide*, *insuline glargine*, *zolpidem*). Signals detected from this source were compared with those from adverse events reporting system.

Methods

Figure: Extraction & methodology

Data sources

Data were extracted from *the Detec't database*, composed of messages posted on patients' forums between 2005 and 2015. Thirteen French patients forums: *Atoute*, *Doctissimo*, *E-sante*, *SanteMedecine*, *Onmeda*, *Futura sciences*, *Psoriasis*, *AlarmAsso*, *Morphee*, *Albi*, *Renaloo*, *Allodocteurs*, *Forum Sclérose en plaques* were screened. Another data source for this study was the WHO adverse events reporting system (VigiBase®) for drug-events comparison. Three drugs *teriflunomide*, *insulin glargine*, *zolpidem* were targeted.

Disproportionally approaches

Methods	Statistics	Criteria
EBGM	EB05	≥ 2
EBGM	EBGM	≥ 2
EBGM	EBGM	≥ 4
		PRR ≥ 2
PRR Composite	PRR, N, χ^2	N ≥ 3
		$\chi^2 \geq 4$
PRR	LB95(log(PRR))	≥ 1
ROR	LB95(log(ROR))	≥ 1
BCPNN	IC025	≥ 1
RFET	p.value	$p \leq 0.05$

Figure: Signals detection strategy

Evaluation

Signals detected from forums were compared to signals detected in VigiBase® by describing *sensitivity*, *specificity*, *positive predictive value (PPV)*, *negative predictive value (NPV)*, *accuracy* and *ROC curves*.

Results

The number of posts from the patients' forums containing *teriflunomide*, *insulin glargine*, *zolpidem* and with the mention of intake were respectively over the 2005-2015 period: 102, 3326, 4584.

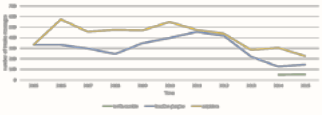


Figure: Messages' intake number for the three drugs

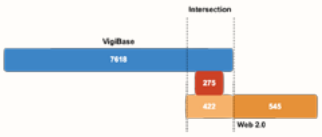


Figure: Drug-event pairs' frequencies

According to metrics used, the sensitivity ranges from 29.1 to 50.6%, the specificity from 86.1 to 95.5%, the PPV from 51.2 to 75.4%, the NPV from 68.5 to 91.6% and the accuracy from 68 to 87.7%

Method	Sen (%)	Spe (%)	PPV (%)	NPV (%)	Acc (%)
EB05	29.06	95.53	62.5	84.03	81.99
EBGM	48.24	89.28	62.5	82.33	78.19
EBGM	39.62	94.57	51.21	91.6	87.67
Composite	31.93	94.05	67.85	77.86	76.54
PRR	37.34	87.5	64.13	70	68.72
ROR	36.96	87.93	66.3	68.48	68
BCPNN	33.33	94.18	75.38	72.54	72.98
RFET	50.6	86.09	68.1	74.83	72.98

Figure: Comparison between patient's forums and VigiBase® signals (all drugs together)

The AUC reaches 0.85 when considering EBGM ≥ 4

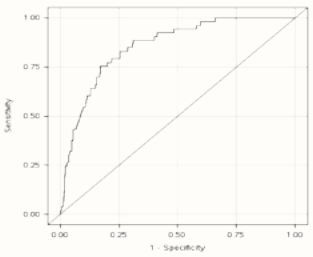


Figure: ROC Curve and AUC for EBGM ≥ 4

The time analysis shows that 30% of the signals are detected earlier (up to 128 months earlier) in the forums than in VigiBase® and 20% are detected at the same time but are available earlier.

Conclusion

Web-based signal detection using forums data in France appears to have good performance features compared to adverse events reporting system. Further analysis will explore time difference between the date of signals from the forums and date of signals from VigiBase®

Conflict of Interest Disclosure

This study was entirely funded by Sanofi. Authors are employees of Sanofi and the other authors are employees at Kappa Santé which was the company doing the data processing and analysis.

References

[1] Sloane R, Osanlou O, Lewis D, Bollegala D, Maskell S, Pirmohamed M.
Social media and pharmacovigilance: A review of the opportunities and challenges.
Br J Clin Pharmacol. 2015 Oct;80(4):910–20.

[2] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al.
Utilizing social media data for pharmacovigilance: A review.
J Biomed Inform. 2015 Apr;54:202–12.

[3] Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al.
Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review.
J Med Internet Res. 2015 Jul;17(7):e171.

[4] White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E.
Toward enhanced pharmacovigilance using patient-generated data on the internet.
Clin Pharmacol Ther. 2014 Aug;96(2):239–46.

[5] Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al.
Text mining for adverse drug events: the promise, challenges, and state of the art.
Drug Saf. 2014 Oct;37(10):777–90.

MedInfo2017
21-25 AUGUST, 2017 • HANGZHOU CHINA
2017

Mining adverse drug reactions in social media with named entity recognition and semantic methods

Xiaoyi Chen, Myrtille Deldossi, Rim Aboukhamis, Carole Faviez, Badisse Dahamna, Pierre Karapetiantz, Armelle Guenegou-Arnoux, Yannick Girardeau, Sylvie Guillemin-Lanne, Agnès Lillo-Le-Louët, Nathalie Texier, Anita Burgun, Sandrine Katsahian

Mining adverse drug reactions in social media with named entity recognition and semantic methods

Xiaoyi Chen[a], Myrtille Deldossi[b], Rim Aboukhamis[c], Carole Faviez[d], Badisse Dahamna[e,f,g], Pierre Karapetiantz[a], Armelle Guenegou-Arnoux[a], Yannick Girardeau[h,i], Sylvie Guillemin-Lanne[b], Agnès Lillo-Le-Louët[c], Nathalie Texier[d], Anita Burgun[a,h,i], Sandrine Katsahian[a,h,i]

[a] INSERM, UMRS1138, équipe 22, Centre de Recherche des Cordeliers, Paris, France
[b] Expert System, 75012 Paris, France
[c] Centre Régional de Pharmacovigilance, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France
[d] Kappa Santé, 75002 Paris, France
[e] Service d'Informatique Biomédicale, CHU de Rouen, France
[f] LITIS-TIBS EA 4108, 76031 Rouen Cedex, France
[g] INSERM, U1142, LIMICS, 75006 Paris, France
[h] Université Paris Descartes, Sorbonne Paris Cité, UMRS1138, Centre de Recherche des Cordeliers, Paris, France
[i] Département d'Informatique Hospitalière, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France

Abstract

Suspected adverse drug reactions (ADR) reported by patients through social media can be a complementary source to current pharmacovigilance systems. However, the performance of text mining tools applied to social media to discover ADRs needs to be evaluated. In this paper, we introduce the approach developed to mine ADR from French social media. A protocol of evaluation is highlighted, which includes a detailed sample size determination and corpus constitution. Our text mining approach provided very encouraging preliminary results with F-measures of 0.94, 0.81 and 0.70 for recognition of drugs, symptoms and ADRs respectively, thus this approach is promising for downstream pharmacovigilance analysis.

Keywords:

Social Media, Pharmacovigilance, Data Mining

Introduction

The rapid expansion of the Internet and social media is changing the way people gather information about disease and treatment, as well as how they share personal health experiences with others [1]. The *Digit in 2016* [2] reported that, in France, 86% of the population are active internet users. This proportion is higher than Western Europe's average of 83% and slightly lower than North America's average of 88%. Various questionnaire statistics [3]–[5] showed that a large proportion of French people (46% to 71%) use the Internet to seek medical or health related information. Many people also use social media, such as forums, to communicate with others with the same health concerns and share information related to their illnesses, feelings, medication use and many other aspects [6], which offers promising opportunities for public health surveillance with a rich internet-based, patient-generated source.

The World Health Organization (WHO) defines Pharmacovigilance as "the science relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problems". It begins during clinical trials and continues after the drug is released onto the market. However a study [7] showed that 60% of potentially fatal ADRs were not described in initial drug labels and 39% were not included in any report of randomized controlled trials. The main pharmacovigilance tools are spontaneous reporting systems, driven by drug agencies, like the U.S. FDA' (Food and Drug Administrations) Adverse Event Reporting System (FAERS) which gathers voluntary reports by healthcare professionals and consumers (59% by professionals Vs. 41% by consumers in 2006, and 46% Vs. 54% in Q1 2015 [8]). It can also include Phase IV clinical trials driven by pharmaceutical companies and governmental agencies [9]. Despite such systems, the underreporting of ADRs by the patients as well as by the health professionals remains a significant limitation [10][11].

Several studies have already demonstrated the value of mining ADR from social media posts [12]–[14]. However, in contrast to the numerous studies in social media, the potential of utilising this data for pharmacovigilance has not yet been fully exploited. It represents only 0.5% of publications with "social media" (SM) keyword query in the PubMed database (Figure 1A). Figure 1B shows that the number of publications with "SM + pharmacovigilance" as keywords has increased exponentially in the last five years.

A recent scoping review [11] outlined five complete steps that should be taken for processing ADR extraction in social media: (1) data collection, (2) preprocessing, (3) entity recognition (for drugs and symptoms), (4) identifying the relationship between drug and symptom, and (5) results evaluation. Since content and language of medical social media differ from those of general social media and of clinical documents, specific text mining methods or techniques based on Natural Language Processing (NLP) are necessary for step (3) and step (4) in order to identify medical concepts (such as drugs, symptoms, etc.) and relations among them [15]. It is evident that the performance of the text mining methods plays a decisive role in ADR signal detection.

From a text mining perspective, the key challenge is that internet users' expressions are usually informal and colloquial, especially when they describe their feelings and symptoms. However, researches have progressed using (i) various data sources, such as forum messages [16][17], Twitter micro blogs [18][19] and Yahoo Wellness Groups [12], in (ii) different languages, (iii) diverse approaches, such as Support Vector

Kap Code - Book de publications scientifiques

JMIR Publications
The Leading eHealth Publisher

2017

Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help?

Redhouane Abdellaoui, MSc; Stéphane Schück, MSc, MD; Nathalie Texier, PharmD; Anita Burgun, MD, PhD

Original Paper

Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help?

Redhouane Abdellaoui[1,2], MSc; Stéphane Schück[2], MSc, MD; Nathalie Texier[2], PharmD; Anita Burgun[1,3], MD, PhD

[1]INSERM, UMRS 1138 Team 22, Université Pierre et Marie Curie, Paris, France

[2]Kappa Santé, Innovation, Paris, France

[3]Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Européen Georges-Pompidou (HEGP), Medical Informatics, Paris, France

Corresponding Author:
Redhouane Abdellaoui, MSc
INSERM
UMRS 1138 Team 22
Université Pierre et Marie Curie
4 rue de Cléry
Paris, 75002
France
Phone: 33 1 44 82 74 74
Fax: 33 1 44 82 74 75
Email: redhouane.abdellaoui@kappasante.com

Abstract

Background: With the increasing popularity of Web 2.0 applications, social media has made it possible for individuals to post messages on adverse drug reactions. In such online conversations, patients discuss their symptoms, medical history, and diseases. These disorders may correspond to adverse drug reactions (ADRs) or any other medical condition. Therefore, methods must be developed to distinguish between false positives and true ADR declarations.

Objective: The aim of this study was to investigate a method for filtering out disorder terms that did not correspond to adverse events by using the distance (as number of words) between the drug term and the disorder or symptom term in the post. We hypothesized that the shorter the distance between the disorder name and the drug, the higher the probability to be an ADR.

Methods: We analyzed a corpus of 648 messages corresponding to a total of 1654 (drug and disorder) pairs from 5 French forums using Gaussian mixture models and an expectation-maximization (EM) algorithm .

Results: The distribution of the distances between the drug term and the disorder term enabled the filtering of 50.03% (733/1465) of the disorders that were not ADRs. Our filtering strategy achieved a precision of 95.8% and a recall of 50.0%.

Conclusions: This study suggests that such distance between terms can be used for identifying false positives, thereby improving ADR detection in social media.

(JMIR Public Health Surveill 2017;3(2):e36) doi:10.2196/publichealth.6577

KEYWORDS

pharmacovigilance; social media; text mining; Gaussian mixture model; EM algorithm; clustering; density estimation

Introduction

Background

Adverse drug reactions (ADRs) cause millions of injuries worldwide each year and require billions of Euros in associated costs [1,2]. Drug safety surveillance targets the detection, assessment, and prevention of ADRs in the postapproval period. A promise of augmenting drug safety with patient-generated data drawn from the Internet was called for by several scientific committees related to pharmacovigilance in the United States and in Europe [3,4].

There are now sites for consumers that enable patients to report ADRs. Patients who experience ADRs want to contribute drug safety content, share their experience, and obtain information and support from other Internet users [5-8].

XSL•FO
RenderX

JMIR Publications
The Leading eHealth Publisher

2017

The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process

Cedric Bousquet, PharmD, PhD; Badisse Dahamna, MSc; Sylvie Guillemin-Lanne, MSc; Stefan J Darmoni, MD, PhD; Carole Faviez, MSc; Charles Huot, PhD; Sandrine Katsahian, MD, PhD; Vincent Leroux, MD; Suzanne Pereira, PhD; Christophe Richard, MD; Stéphane Schück, MPH, MD; Julien Souvignet, MSc; Agnès Lillo-Le Louët, MD; Nathalie Texier, PharmD

Original Paper

The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process

Cedric Bousquet[1,2], PharmD, PhD; Badisse Dahamna[3], MSc; Sylvie Guillemin-Lanne[4], MSc; Stefan J Darmoni[1,3], MD, PhD; Carole Faviez[5], MSc; Charles Huot[4], PhD; Sandrine Katsahian[6], MD, PhD; Vincent Leroux[7], MD; Suzanne Pereira[8], PhD; Christophe Richard[9], MD; Stéphane Schück[5], MPH, MD; Julien Souvignet[1], MSc; Agnès Lillo-Le Louët[10], MD; Nathalie Texier[5], PharmD

[1]Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé, U1142, Institut National de la Santé et de la Recherche Médicale, Paris, France
[2]Service de Santé Publique et de l'Information Médicale, Centre Hospitalier Universitaire de Saint Etienne, Saint-Etienne, France
[3]Department of Biomedical Informatics, Rouen University Hospital, Rouen, France
[4]Expert System, Paris, France
[5]Kappa Santé, Paris, France
[6]Unité mixte de recherche 1138, équipe 22, Institut National de la Santé et de la Recherche Médicale, Centre de Recherche des Cordeliers, Paris, France
[7]Institut de Santé Urbaine, Saint Maurice, France
[8]Vidal, Issy Les Moulineaux, France
[9]Santeos, Paris, France
[10]Assistance Publique-Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Centre Régional de Pharmacovigilance, Paris, France

Corresponding Author:
Cedric Bousquet, PharmD, PhD
Service de Santé Publique et de l'Information Médicale
Centre Hospitalier Universitaire de Saint Etienne
Chemin de la Marandière
Bâtiment CIM42 - Hôpital Nord
Saint-Etienne,
France
Phone: 33 477 27974
Email: cedric.bousquet@chu-st-etienne.fr

Abstract

Background: Adverse drug reactions (ADRs) are an important cause of morbidity and mortality. Classical Pharmacovigilance process is limited by underreporting which justifies the current interest in new knowledge sources such as social media. The Adverse Drug Reactions from Patient Reports in Social Media (ADR-PRISM) project aims to extract ADRs reported by patients in these media. We identified 5 major challenges to overcome to operationalize the analysis of patient posts: (1) variable quality of information on social media, (2) guarantee of data privacy, (3) response to pharmacovigilance expert expectations, (4) identification of relevant information within Web pages, and (5) robust and evolutive architecture.

Objective: This article aims to describe the current state of advancement of the ADR-PRISM project by focusing on the solutions we have chosen to address these 5 major challenges.

Methods: In this article, we propose methods and describe the advancement of this project on several aspects: (1) a quality driven approach for selecting relevant social media for the extraction of knowledge on potential ADRs, (2) an assessment of ethical issues and French regulation for the analysis of data on social media, (3) an analysis of pharmacovigilance expert requirements when reviewing patient posts on the Internet, (4) an extraction method based on natural language processing, pattern based matching, and selection of relevant medical concepts in reference terminologies, and (5) specifications of a component-based architecture for the monitoring system.

Results: Considering the 5 major challenges, we (1) selected a set of 21 validated criteria for selecting social media to support the extraction of potential ADRs, (2) proposed solutions to guarantee data privacy of patients posting on Internet, (3) took into

Kap Code - Book de publications scientifiques

Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach

Redhouane Abdellaoui, MSc; Pierre Foulquié, MSc; Nathalie Texier, PharmD; Carole Faviez, MSc; Anita Burgun, MD, PhD; Stéphane Schück2, MSc, MD

Original Paper

Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach

Redhouane Abdellaoui[1], MSc; Pierre Foulquié[2], MSc; Nathalie Texier[2], PharmD; Carole Faviez[2], MSc; Anita Burgun[1,3], MD, PhD; Stéphane Schück[2], MSc, MD

[1]Unité de Mixte de Recherche 1138 Team 22, Institut National de la Santé et de la Recherche Médicale / Université Pierre et Marie Curie, Paris, France
[2]Kappa Santé, Innovation (Kap Code), Paris, France
[3]Medical Informatics, Hôpital Européen Georges-Pompidou, Assistance Publique-Hôpitaux de Paris, Paris, France

Corresponding Author:
Redhouane Abdellaoui, MSc
Unité de Mixte de Recherche 1138 Team 22
Institut National de la Santé et de la Recherche Médicale / Université Pierre et Marie Curie
15 Rue de l'École de Médecine
Paris, 75006
France
Phone: 33 648094269
Email: redhouane.a@gmail.com

Abstract

Background: Medication nonadherence is a major impediment to the management of many health conditions. A better understanding of the factors underlying noncompliance to treatment may help health professionals to address it. Patients use peer-to-peer virtual communities and social media to share their experiences regarding their treatments and diseases. Using topic models makes it possible to model themes present in a collection of posts, thus to identify cases of noncompliance.

Objective: The aim of this study was to detect messages describing patients' noncompliant behaviors associated with a drug of interest. Thus, the objective was the clustering of posts featuring a homogeneous vocabulary related to nonadherent attitudes.

Methods: We focused on escitalopram and aripiprazole used to treat depression and psychotic conditions, respectively. We implemented a probabilistic topic model to identify the topics that occurred in a corpus of messages mentioning these drugs, posted from 2004 to 2013 on three of the most popular French forums. Data were collected using a Web crawler designed by Kappa Santé as part of the Detec't project to analyze social media for drug safety. Several topics were related to noncompliance to treatment.

Results: Starting from a corpus of 3650 posts related to an antidepressant drug (escitalopram) and 2164 posts related to an antipsychotic drug (aripiprazole), the use of latent Dirichlet allocation allowed us to model several themes, including interruptions of treatment and changes in dosage. The topic model approach detected cases of noncompliance behaviors with a recall of 98.5% (272/276) and a precision of 32.6% (272/844).

Conclusions: Topic models enabled us to explore patients' discussions on community websites and to identify posts related with noncompliant behaviors. After a manual review of the messages in the noncompliance topics, we found that noncompliance to treatment was present in 6.17% (276/4469) of the posts.

(J Med Internet Res 2018;20(3):e85) doi:10.2196/jmir.9222

KEYWORDS

medication adherence; compliance; infodemiology; social media; text mining; depression; psychosis; peer-to-peer support; virtual community

frontiers in Pharmacology

2018

Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate

Xiaoyi Chen, Carole Faviez, Stéphane Schuck, Agnès Lillo-Le-Louët, Nathalie Texier, Badisse Dahamna, Charles Huot, Pierre Foulquié, Suzanne Pereira, Vincent Leroux, Pierre Karapetiantz, Armelle Guenegou-Arnoux, Sandrine Katsahian, Cédric Bousquet and Anita Burgun

Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate

OPEN ACCESS

Edited by:
Iñaki Gutiérrez-Ibarluzea, Basque Office for Health Technology Assessment (OSTEBA), Spain

Reviewed by:
Bryan Martin Bennett, Adelphi (United Kingdom), United Kingdom
Ana Paula Martins, Universidade de Lisboa, Portugal

***Correspondence:**
Xiaoyi Chen
xiaoyi.chen@inserm.fr
Carole Faviez
carole.faviez@kapcode.fr

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Pharmaceutical Medicine and Outcomes Research, a section of the journal Frontiers in Pharmacology

Received: *15 December 2017*
Accepted: *04 May 2018*
Published: *24 May 2018*

Citation:
Chen X, Faviez C, Schuck S, Lillo-Le-Louët A, Texier N, Dahamna B, Huot C, Foulquié P, Pereira S, Leroux V, Karapetiantz P, Guenegou-Arnoux A, Katsahian S, Bousquet C and Burgun A (2018) Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate. Front. Pharmacol. 9:541. doi: 10.3389/fphar.2018.00541

***Xiaoyi Chen*[1*†], *Carole Faviez*[2*†], *Stéphane Schuck*[2], *Agnès Lillo-Le-Louët*[3], *Nathalie Texier*[2], *Badisse Dahamna*[4,5], *Charles Huot*[6], *Pierre Foulquié*[2], *Suzanne Pereira*[7], *Vincent Leroux*[8], *Pierre Karapetiantz*[1], *Armelle Guenegou-Arnoux*[1], *Sandrine Katsahian*[1,9], *Cédric Bousquet*[10] *and Anita Burgun*[1,9]**

[1] UMRS 1138, équipe 22, Institut National de la Santé et de la Recherche Médicale, Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France, [2] Kappa Santé, Paris, France, [3] Centre Régional de Pharmacovigilance, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France, [4] Service d'Informatique Biomédicale, Centre Hospitalier Universitaire de Rouen, Rouen, France, [5] Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes-TIBS EA 4108, Rouen, France, [6] Expert System, Paris, France, [7] Vidal, Issy Les Moulineaux, France, [8] Institut de Santé Urbaine, Saint-Maurice, France, [9] Département d'Informatique Médicale, Hôpital Européen Georges Pompidou, Paris, France, [10] Sorbonne Université, Inserm, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, Paris, France

Background: The Food and Drug Administration (FDA) in the United States and the European Medicines Agency (EMA) have recognized social media as a new data source to strengthen their activities regarding drug safety.

Objective: Our objective in the ADR-PRISM project was to provide text mining and visualization tools to explore a corpus of posts extracted from social media. We evaluated this approach on a corpus of 21 million posts from five patient forums, and conducted a qualitative analysis of the data available on methylphenidate in this corpus.

Methods: We applied text mining methods based on named entity recognition and relation extraction in the corpus, followed by signal detection using proportional reporting ratio (PRR). We also used topic modeling based on the Correlated Topic Model to obtain the list of the matics in the corpus and classify the messages based on their topics.

Results: We automatically identified 3443 posts about methylphenidate published between 2007 and 2016, among which 61 adverse drug reactions (ADR) were automatically detected. Two pharmacovigilance experts evaluated manually the quality of automatic identification, and a f-measure of 0.57 was reached. Patient's reports were mainly neuro-psychiatric effects. Applying PRR, 67% of the ADRs were signals, including most of the neuro-psychiatric symptoms but also palpitations. Topic modeling showed that the most represented topics were related to *Childhood and Treatment initiation*, but also *Side effects*. Cases of misuse were also identified in this corpus, including recreational use and abuse.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2018

Détection automatique du mésusage des neuroleptiques dans le trouble anxieux et la démence à partir des réseaux sociaux

Stéphane Schück, Pierre Foulquié, Adel Mebarki, Paméla Voillot, Nathalie Texier, Carole Faviez

DÉTECTION AUTOMATIQUE DU MÉSUSAGE DES NEUROLEPTIQUES DANS LE TROUBLE ANXIEUX ET LA DÉMENCE À PARTIR DES RÉSEAUX SOCIAUX

Stéphane Schück[1], Pierre Foulquié[1], Adel Mebarki[1], Paméla Voillot[1], Nathalie Texier[1], Carole Faviez[1]
[1]Kap Code, Paris, France

KAP CODE

INTRODUCTION

Les neuroleptiques sont des psychotropes utilisés dans le traitement des troubles psychotiques, tels que la schizophrénie et les troubles bipolaires. Dès 2011, l'étude réalisée par Alexander et al. met en évidence un usage des neuroleptiques en dehors de leurs autorisations de mise sur le marché (AMM) grandissant aux Etats-Unis [1].

En parallèle, on estime que **50% des français sont actifs** sur les réseaux sociaux, un grand nombre d'entre eux échangeant à propos de leur santé.

Dans ce contexte, l'objectif de cette étude consistait à (1) évaluer s'il était possible d'identifier des usages hors AMM des neuroleptiques sur les réseaux sociaux français dans le cadre de la démence et des troubles anxieux et (2) de mettre en place une méthodologie de détection automatique de ces cas.

MATÉRIEL & MÉTHODES

Les messages associés à deux **neuroleptiques, aripiprazole** et **rispéridone**, ont été extraits à partir de forums médicaux généralistes français via l'outil Detec't [2] développé par la société Kap Code.

La méthode présentée ici s'organisait en deux étapes. Dans un premier temps, des échantillons des deux corpus ont été annotés manuellement afin d'identifier les messages associés à une prise médicamenteuse dans le cadre d'un des deux mésusages recherchés.

Dans un deuxième temps, des champs lexicaux correspondant à ces mésusages ont été créés. Ils ont été initialement constitués à partir du dictionnaire d'effets indésirables MedDRA (*Medical Dictionary for Regulatory Activities*). Un modèle de sujet (topic model) a ensuite permis l'identification, au sein du corpus **aripiprazole**, de messages contenant un vocabulaire patient évocateur du mésusage. Ces messages ont été revus et le vocabulaire identifié a été utilisé pour compléter les champs lexicaux. Des synonymes issus du langage courant ont enfin été ajoutés afin de parfaire les champs lexicaux.

La recherche des éléments de ces champs lexicaux au sein des corpus avait pour objectif d'identifier automatiquement les messages associés aux mésusages en question.

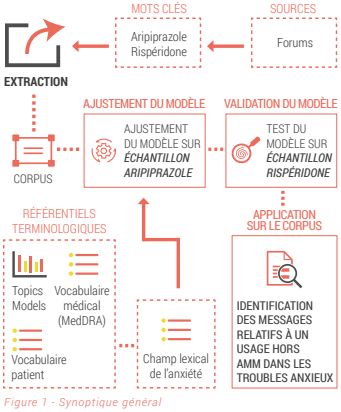


Figure 1 - Synoptique général

Cette méthode de détection a été appliquée aux deux échantillons préalablement annotés, dans un premier temps à l'échantillon **aripiprazole** pour ajustement du champ lexical et de la méthode, puis dans un deuxième temps à l'échantillon **rispéridone** pour test du modèle.

RÉSULTATS

Les corpus associés à **aripiprazole** et **rispéridone** contenaient, respectivement, **9 528** et **3 868 messages**.

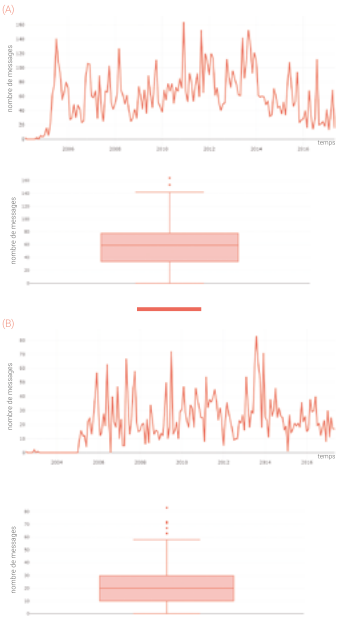


Figure 2 - Distribution des observations mensuelles. (A) - Corpus aripiprazole / (B) - Corpus rispéridone

Annotation des messages

1 000 messages issus du corpus **aripiprazole** et **400 messages issus du corpus rispéridone** ont été annotés manuellement.

Au sein du corpus **aripiprazole** (resp. rispéridone), **35 messages associés au trouble anxieux (resp. 15), soit 3,5% (IC 95% : 2,4%-4,6%) des messages de l'échantillon (resp. 3,8%, IC 95% : 2,0%-5,6%)**, ont été identifiés.

Aucun message associé à la démence n'a pu être identifié dans aucun des deux corpus.

Tableau 1 – Effectifs en nombre de messages au sein des échantillons annotés.

TRAITEMENT	Échantillon	Troubles anxieux	Démences
Aripiprazole	1 000 messages	35 messages	0 message
Rispéridone	400 messages	15 messages	0 message

Constitution des champs lexicaux

La recherche de termes évocateurs du trouble anxieux au sein du dictionnaire MedDRA a permis l'identification de **1 116 termes**. En parallèle, la revue de messages et l'ajout de vocabulaire patient a permis d'identifier **227 termes supplémentaires**.

Ces **1 343 termes** ont été répartis en trois classes en fonction de leur spécificité par rapport au trouble anxieux ; la classe 1 contenait des termes très spécifiques du trouble anxieux (tels que *angoisse, anxiété*, etc.), la classe 2 des termes évocateurs du trouble anxieux mais qui pouvaient également être associés à d'autres pathologies (*agitation, insomnie*, etc.) et la classe 3 des termes encore moins spécifiques du trouble anxieux.

Tableau 2 – Répartition des termes du champ lexical du trouble anxieux

ORIGINE	Classe 1	Classe 2	Classe 3	Total
MedDRA	205	889	22	1 116
Termes supplémentaires	54	172	1	227
Total	**259**	**1 061**	**23**	**1 343**

Test du modèle

Le test de différents modèles utilisant les différentes classes de ces champs lexicaux sur l'échantillon **aripiprazole** a permis d'identifier le modèle possédant les meilleures performances.

Ce modèle était uniquement construit à partir de la classe 1 du champ lexical, c'est à dire les termes les plus spécifiques (soit **205 termes** issus de MedDRA et **54 termes** issus de l'annotation des messages du corpus **aripiprazole**, identifiés par le *topic model*).

Le modèle a alors été réappliqué sur l'échantillon **rispéridone**. Les performances obtenues ont été équivalentes sur les deux échantillons, le rappel et la précision allant respectivement de **97** à **100%** et de **19** à **20%**.

*Tableau 3 – Performances du modèle **trouble anxieux** sur les échantillons*

TRAITEMENT	Sensibilité	Spécifité	VPP	VPN
Aripiprazole	97%	86%	20%	100%
Rispéridone	100%	83%	19%	100%

CONCLUSION

Une prise de neuroleptiques dans le cadre des troubles anxieux a été identifiée sur les réseaux sociaux. Le mésusage dans le cadre de la démence n'a pas été identifié. La méthodologie de détection automatique de ces cas de mésusages a obtenu de bonnes performances pour détecter ces messages d'intérêt.

RÉFÉRENCES

[1] Alexander, G. Caleb, et al. «Increasing off-label use of antipsychotic medications in the United States, 1995–2008.» Pharmacoepidemiology and drug safety 20.2 (2011): 177-184.

[2] Abdellaoui R et al. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help?. JMIR public health and surveillance, 2017, vol. 3, no 2.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2018

Que nous apportent les réseaux sociaux quant à la crise sanitaire du Levothyrox d'Août 2017 ?

Stéphane Schück, Paméla Voillot, Pierre Foulquié, Carole Faviez, Adel Mebarki, Nathalie Texier, Tristan Gauvin

QUE NOUS APPORTENT LES RÉSEAUX SOCIAUX QUANT À LA CRISE SANITAIRE DU LEVOTHYROX® D'AOÛT 2017 ?

Stéphane Schück[1], Paméla Voillot[1], Pierre Foulquié[1], Carole Faviez[1], Adel Mebarki[1], Nathalie Texier[1], Tristan Gauvin[2]
[1]Kap Code, Paris, France - [2]Lagardère Active, Levallois Perret, France

KAP CODE

INTRODUCTION

Le **Levothyrox®**, prescrit en France à plus de 3 millions de patients atteints de pathologies thyroïdiennes, a fait l'objet d'une polémique en août 2017 suite à un changement de formule en mars 2017. Un nombre important d'effets indésirables a été reporté auprès des autorités sanitaires, des incertitudes persistant concernant la nature de ces effets ou leur spécificité par rapport à la nouvelle formule.

En parallèle, **32 millions d'internautes** sont actifs sur les réseaux sociaux, favorisant les échanges à propos de leur santé.

Dans ce contexte, l'objectif de cette étude consistait à identifier via les réseaux sociaux ces échanges et les thématiques abordées par les patients concernant le **Levothyrox®**, ainsi que les effets indésirables mentionnés.

MATÉRIEL & MÉTHODES

Les messages évoquant le **Levothyrox®** publiés entre 2007 et 2017 ont été extraits à partir du forums français Doctissimo via l'outil Detec't [1] développé par la société Kap Code. L'identification des effets indésirables était effectuée via ce dernier.

L'application d'un algorithme de veille basé sur le volume de messages postés permettait la classification des observations d'une période selon leur activité, plus ou moins soutenue.

Les thématiques de discussion étaient modélisées à l'aide d'un modèle de sujet (*topic model*). L'application d'un partitionnement k-moyennes (*k-means clustering*) permettait de catégoriser les auteurs des messages selon leur comportement.

Enfin, la détection de signal consistait en l'application de la méthode du PRR Composite* sur trois périodes : 2007-2015, 2007-2016 et 2007-2017.

RÉSULTATS

Le corpus d'analyse **Levothyrox®** contenait près de **30 000 messages** publiés entre **2007 et 2017** par plus de **8 000 internautes** différents.

L'évolution temporelle (fig.1) a permis de mettre en évidence un pic de messages en août et septembre 2017, concomitant au bruit médiatique lié au changement de formule du **Levothyrox®** en mars 2017.

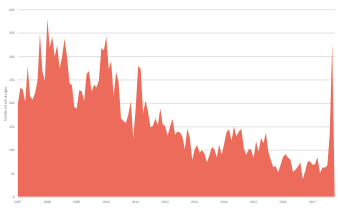


Figure 1 – Évolution temporelle du nombre de messages

Plus de **34 000 effets secondaires** ont été détectés au sein des messages entre 2007 et 2017. L'évolution du nombre d'effets suit l'évolution du nombre de messages. Ces effets sont regroupés en catégories (fig.2).

Quatre de ces catégories de troubles regroupent plus de **75%** des effets détectés.

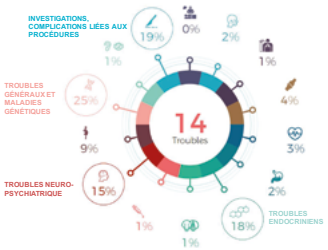


Figure 2 – Catégories de troubles identifiés

L'algorithme de veille a permis de classer les observations de l'été 2017 en activité soutenue (fig.3).

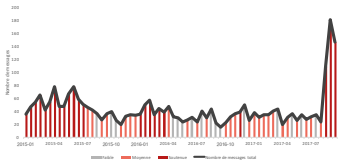
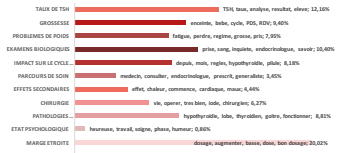


Figure 3 – Évolution temporelle du nombre de messages

Les thématiques de discussions (fig.4) identifiées couvraient les différents aspects des pathologies thyroïdiennes tels que les actes médicaux relatifs à la maladie, les modalités de la prise du traitement et le parcours de soins.



* Un message peut être associé à plusieurs thèmes

Figure 4 – Proportions de messages postés associés à chaque thème de discussions*

La catégorisation des utilisateurs (fig.5) a fait émerger un groupe de **nouveaux internautes** n'ayant quasiment posté qu'en 2017. Ils étaient les auteurs de **77%** des messages postés cette année là.

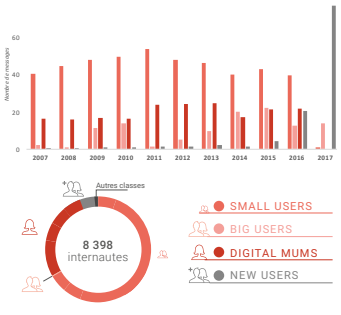


Figure 5 – Répartition annuelles des messages par typologie d'utilisateurs

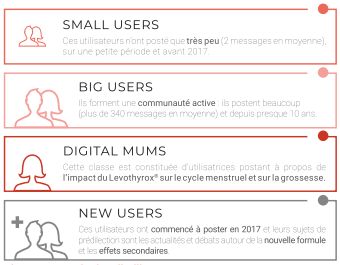
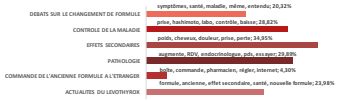


Figure 6 - Typologies d'utilisateurs

Toujours sur l'année 2017, nous avons observé que le changement de formule était au cœur des discussions. Les patients échangeaient (fig.7) également autour de la nature du changement des **effets secondaires** que le **Levothyrox®** est accusé de provoquer ainsi que de la possibilité de se procurer l'ancienne formule à l'étranger.



* Un message peut être associé à plusieurs thèmes

Figure 7 – Proportions de messages postés en 2017 associés à chaque thème de discussions*

La détection de signaux (fig.8) a mis en avant des signaux déjà connus tels que les **chutes de cheveux ou les vertiges**. Les deux nouveaux signaux observés en 2017 étaient la **froideur des extrémités et la fatigue musculaire**.

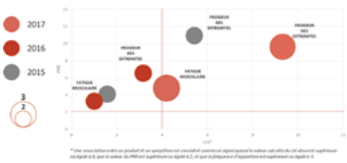


Figure 8 – Détection de signaux sur les périodes 2007-2015; 2007-2016 et 2007-2017*

CONCLUSION

Les messages publiés en 2017 concernaient principalement le changement de formule et ses conséquences. Ils étaient majoritairement publiés par de nouveaux utilisateurs. Les effets secondaires médiatisés étaient déjà connus mais ont augmenté en 2017. La surveillance des réseaux sociaux et des effets indésirables rapportés par de nouveaux utilisateurs apparait comme un outil d'alerte supplémentaire dans la chaine de surveillance du médicament.

RÉFÉRENCES

[1] Abdellaoui R et al. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help?. JMIR public health and surveillance, 2017, vol. 3, no 2.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2018

Étude de l'usage du Méthylphénidate sur les réseaux sociaux

Pierre Foulquié, Paméla Voillot, Carole Faviez,
Adel Mebarki, Xiaoyi Chen, Nathalie Texier, Stéphane Schück

ÉTUDE DE L'USAGE DU MÉTHYLPHÉNIDATE SUR LES RÉSEAUX SOCIAUX

Pierre Foulquié[1], Paméla Voillot[1], Carole Faviez[1],
Adel Mebarki[1], Xiaoyi Chen[2], Nathalie Texier[1], Stéphane Schück[1]
[1]Kap Code, Paris, France - [2]INSERM, UMRS1138, équipe 22, Centre de Recherche des Cordeliers, Paris, France

KAP CODE

INTRODUCTION

Le **méthylphénidate**, prescrit en France à plus de **49 000 patients** atteints de troubles du déficit de l'attention avec hyperactivité (TDAH) âgés de 6 à 17 ans, a fait l'objet de plusieurs évaluations par l'EMA et l'ANSM concernant son usage hors AMM.

Parallèlement, la FDA et l'EMA ont reconnu les réseaux sociaux comme nouvelle source de données pour renforcer la surveillance des médicaments [1][2].

Dans ce contexte, l'objectif de cette étude consistait en **l'exploration du comportement** des patients vis-à-vis du **méthylphénidate** à travers les réseaux sociaux.

MATÉRIEL & MÉTHODES

Les messages ayant trait au **méthylphénidate** ont été extraits à partir de cinq forums français généralistes via l'outil Detec't [3], développé par la société Kap Code.

L'analyse s'est déroulée en deux étapes :

1. Premièrement, un modèle de sujet (*topic model*) était appliqué pour identifier les thèmes de discussions (*topic*) relatifs au **méthylphénidate**. Ces derniers permettaient de classer les messages selon les thématiques qu'ils contenaient.

2. Les messages associés aux thématiques qui se rapportaient aux effets secondaires, aux mésusages et aux abus du **méthylphénidate** faisaient ensuite l'objet d'une classification hiérarchique descendante (CHD), directement appliquée aux mots. Différentes sous-thématiques étaient ainsi obtenues.

RÉSULTATS

Les corpus d'analyse méthylphénidate était constitué de **3 343 messages** publiés **entre 2007 et 2016**.

L'évolution temporelle du volume de messages postés (fig.1) n'a pas permis en général pas de mettre en évidence de pics aux dates clés associées au profil de sécurité du **méthylphénidate**. Ces dates consistaient en des décisions à l'échelle française (2006, 2011, 2012 et 2013) et européenne (2007 et 2009). Seule la décision de la HAS (2012) semble avoir eu une influence sur le volume de messages observé.

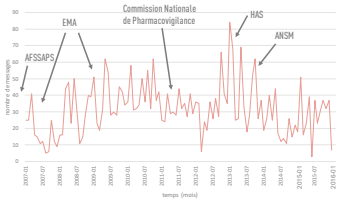


Figure 1 – Évolution temporelle des messages associés au méthylphénidate

La création d'un **nuage de mots** a permis d'observer les termes les plus présents dans le **corpus**.

Plusieurs champs lexicaux tels que ceux de la TDAH, de la prescription et de la prise, de l'enfance ainsi que des inquiétudes concernant les **effets indésirables** ont pu être identifiées.

Le nuage de mots (fig.2) indique une plus grande fréquence d'utilisation du méthylphénidate chez les enfants, en accord avec l'indication du médicament.

Figure 2 – Nuage de mots des messages méthylphénidate

Par la suite, l'application d'un topic model a permis de mettre en avant 14 thématiques de discussions principales (fig. 3).

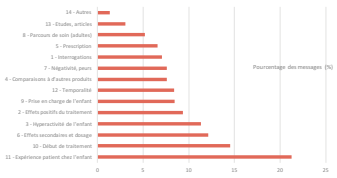


Figure 3 –Thématiques du topic model

Parmi les topics principaux, un topic était centré sur les « **effets secondaires** » (**12,15%** des messages), deux sur le mésusage (intitulés « **Comparaison aux autres produits** » et « **Négativité et peurs** », **7,6%** des messages chacun) et un sur l'usage chez les adultes (**5,2%** des messages). Les autres concernaient des usages cohérents dans le cadre de l'AMM.

La CHD appliquée au topic d'« **effets secondaires** » (fig.4) a mis en évidence plusieurs troubles liés à la prise du traitement tels que la perte de poids ou encore les vomissements.

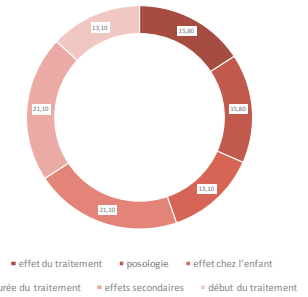


Figure 4 – Thématiques de la CDH du topic « effets secondaires »

Concernant les topics du mésusage (fig. 5), la méthode a permis d'observer l'utilisation de la molécule par les adultes ou les étudiants pour un usage récréatif ou pour augmenter leurs capacités.

De plus, les patients comparaient les effets de la molécule à certaines drogues telles que les amphétamines. Enfin, le topic sur l'usage par les adultes a mis en avant des erreurs ou des retards de diagnostic de TDAH.

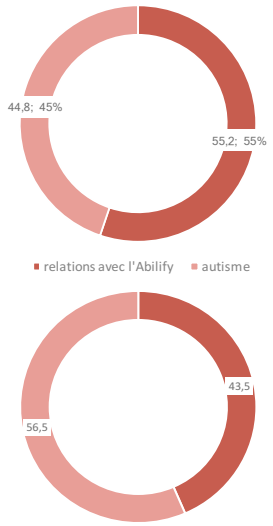


Figure 5 – Thématiques de la CDH des deux topics « mésusage »

CONCLUSION

Cette étude a permis d'identifier :

- Des **effets indésirables**, ce thème constituant un thème d'importance au sein des discussions ;
- Des pratiques non identifiables via les systèmes de soins classiques :
  - l'usage du **méthylphénidate** par des populations spécifiques telles que les adultes et les étudiants.
  - des utilisations hors AMM : comme **substitution aux drogues** et en tant que **psychostimulant**.

La surveillance des réseaux sociaux apparait comme un outil d'alerte complémentaire dans la surveillance des médicaments.

RÉFÉRENCES

[1] www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm452304

[2] www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000258

[3] Abdellaoui R et al. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help?. JMIR public health and surveillance, 2017, vol. 3, no 2.

JMIR Publications
The Leading eHealth Publisher

2018

Web-Based Signal Detection Using Medical Forums Data in France: Comparative Analysis

Marie-Laure Kürzinger, MSc; Stéphane Schück, MSc, MD; Nathalie Texier, PharmD; Redhouane Abdellaoui, MSc; Carole Faviez, MSc; Julie Pouget, MSc; Ling Zhang, MSc; Stéphanie Tcherny-Lessenot, MSc, MD; Stephen Lin, MD; Juhaeri Juhaeri, PhD

Original Paper

Web-Based Signal Detection Using Medical Forums Data in France: Comparative Analysis

Marie-Laure Kürzinger[1], MSc; Stéphane Schück[2], MSc, MD; Nathalie Texier[2], PharmD; Redhouane Abdellaoui[3], MSc; Carole Faviez[3], MSc; Julie Pouget[4], MSc; Ling Zhang[5], MSc; Stéphanie Tcherny-Lessenot[1], MSc, MD; Stephen Lin[5], MD; Juhaeri Juhaeri[6], PhD

[1]Epidemiology and Benefit Risk Evaluation, Sanofi, Chilly-Mazarin, France
[2]Kappa Santé, Paris, France
[3]Kap Code, Paris, France
[4]Information Technology and Solutions, Sanofi, Lyon, France
[5]Global Pharmacovigilance, Sanofi, Bridgewater, NJ, United States
[6]Epidemiology and Benefit Risk Evaluation, Sanofi, Bridgewater, NJ, United States

Corresponding Author:
Marie-Laure Kürzinger, MSc
Epidemiology and Benefit Risk Evaluation
Sanofi
1 avenue Pierre Brossolette
Chilly-Mazarin, 91385
France
Phone: 33 1 69 74 59 42
Email: marie-laure.kurzinger@sanofi.com

Abstract

Background: While traditional signal detection methods in pharmacovigilance are based on spontaneous reports, the use of social media is emerging. The potential strength of Web-based data relies on their volume and real-time availability, allowing early detection of signals of disproportionate reporting (SDRs).

Objective: This study aimed (1) to assess the consistency of SDRs detected from patients' medical forums in France compared with those detected from the traditional reporting systems and (2) to assess the ability of SDRs in identifying earlier than the traditional reporting systems.

Methods: Messages posted on patients' forums between 2005 and 2015 were used. We retained 8 disproportionality definitions. Comparison of SDRs from the forums with SDRs detected in VigiBase was done by describing the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, receiver operating characteristics curve, and the area under the curve (AUC). The time difference in months between the detection dates of SDRs from the forums and VigiBase was provided.

Results: The comparison analysis showed that the sensitivity ranged from 29% to 50.6%, the specificity from 86.1% to 95.5%, the PPV from 51.2% to 75.4%, the NPV from 68.5% to 91.6%, and the accuracy from 68% to 87.7%. The AUC reached 0.85 when using the metric empirical Bayes geometric mean. Up to 38% (12/32) of the SDRs were detected earlier in the forums than that in VigiBase.

Conclusions: The specificity, PPV, and NPV were high. The overall performance was good, showing that data from medical forums may be a valuable source for signal detection. In total, up to 38% (12/32) of the SDRs could have been detected earlier, thus, ensuring the increased safety of patients. Further enhancements are needed to investigate the reliability and validation of patients' medical forums worldwide, the extension of this analysis to all possible drugs or at least to a wider selection of drugs, as well as to further assess performance against established signals.

(J Med Internet Res 2018;20(11):e10466) doi:10.2196/10466

KEYWORDS

adverse event; internet; medical forums; pharmacovigilance; signal detection; signals of disproportionate reporting; social media

HEALTHCARE DATA INSTITUTE

2018

LIVRE BLANC : Les réseaux sociaux et la santé

LES RÉSEAUX SOCIAUX ET LA SANTÉ :

UN ENJEU POUR LE SUIVI DES PATIENTS ET LA RECHERCHE SCIENTIFIQUE

CO-AUTEURS (PAR ORDRE ALPHABÉTIQUE) :

- **Lina Autelitano,** Chef de projet digital, Direction Digitale, Pierre Fabre Médicament & Santé, membre du Healthcare Data Institute
- **Caroline Henry,** Avocat Associé, Pons & Carrère, membre du conseil d'administration du Healthcare Data Institute
- **Adel Mebarki,** Directeur général de Kap Code, membre du Healthcare Data Institute
- **Patrick Olivier,** Directeur général IVBAR France, Vice-Président du Healthcare Data Institute
- **Francisco Orchard,** data scientist Epiconcept, membre du Healthcare Data Institute

Coordination des travaux : Jean-Baptiste Fantun

PubMed

2019

How to improve vaccine acceptability (evaluation, pharmacovigilance, communication, public health, mandatory vaccination, fears and beliefs).

Dutilleul A, Morel J, Schilte C, Launay O; Autran B, Béhier JM, Borel T, Bresse X, Chêne G, Courcier S, Dufour V, Faurisson F, Gagneur A, Gelpi O, Gérald F, Kheloufi F, Koeck JL, Lamarque-Garnier V, Lery T, Ménin G, Molimard M, Opinel A, Roger C, Rouby F, Schuck S, Simon L, Soubeyrand B, Truchet MC.

Therapie. 2019 Feb;74(1):131-140. doi: 10.1016/j.therap.2018.12.005. Epub 2018 Dec 20.

How to improve vaccine acceptability (evaluation, pharmacovigilance, communication, public health, mandatory vaccination, fears and beliefs).

Dutilleul A[1], Morel J[2], Schilte C[3], Launay O[4]; participants of Giens XXXIV Round Table "hot topic No. 1".

Collaborators (24)

Autran B[5], Béhier JM[6], Borel T[7], Bresse X[8], Chêne G[9], Courcier S[10], Dufour V[11], Faurisson F[12], Gagneur A[13], Gelpi O[3], Gérald F[14], Kheloufi F[15], Koeck JL[16], Lamarque-Garnier V[17], Lery T[18], Ménin G[19], Molimard M[20], Opinel A[3], Roger C[10], Rouby F[15], Schuck S[21], Simon L[22], Soubeyrand B[23], Truchet MC[24].

Author information

1 Sanofi France, 69007 Lyon, France.
2 CHU de Montpellier, 34295 Montpellier, France.
3 Institut Pasteur, 75015 Paris, France.
4 Inserm, CIC Cochin Pasteur, hôpital Cochin, université Paris-Descartes, 75679 Paris, France. Electronic address: odile.launay@aphp.fr.
5 COREVAC-AVIESAN, hôpital Pitié-Salpêtrière, 75013 Paris, France.
6 Celgene, 92066 Paris-la-Défense, France.
7 Leem, 75017 Paris, France.
8 MSD vaccins, 69007 Lyon, France.
9 Inserm, 33076 Bordeaux, France.
10 GSK, 92500 Rueil-Malmaison, France.
11 PMI, ville de Paris, Infovac-France, 75116 Paris, France.
12 Inserm, 75013 Paris, France.
13 Université de Sherbrooke, J1H5N4 Sherbrooke, Canada.
14 ACS-France, 06300 Nice, France.
15 CHU Timone, AP-HM, 13005 Marseille, France.
16 Direction centrale du service de santé des armées, 94114 Arcueil, France.
17 EVAL Santé, 78290 Croissy-sur-Seine, France.
18 Janssen Cilag, 92787 Issy-les-Moulineaux, France.
19 Sanofi Pasteur Europe, 69007 Lyon, France.
20 CHU de Bordeaux, 33076 Bordeaux, France.
21 Kappa Santé, 75002 Paris, France.
22 Le pharmacien de France, presse professionnelle pharmaceutique, 75009 Paris, France.
23 Blossom vaccinology, 69001 Lyon, France.
24 Pfizer, 75668 Paris, France.

Abstract

A flagship recommendation of the citizen's steering committee on immunization, the mandatory immunization for infants extended to 11 vaccines, introduced in January 2018, is part of a set of recommendations that must be considered as a whole, each component being indispensable to the achievement of objectives: restore confidence in vaccination and increase immunization coverage. Roundtable # 6 participants identified a decade of concrete initiatives that could address, at least in part, the committee's recommendations, including: developing information systems and data generation; simplify the vaccination journey and increase vaccination opportunities; developing training of health professionals; learning vaccines at school; using motivational interviewing in educational intervention; undertaking local initiatives; improving supply and communicate on the value of vaccines. To carry out these actions, it has been proposed that a joint ministerial task-force bringing together the different stakeholders at the national level should be set up to promote their implementation and follow-up, and at regional level, the establishment of an Agences régionales de santé awareness plan making vaccination a priority.

Copyright © 2018. Published by Elsevier Masson SAS.

KEYWORDS: Education; Hesitancy; Immunization; Information systems; Motivational interviewing; Vaccine

PMID: 30660377 DOI: 10.1016/j.therap.2018.12.005

JMIR Publications
The Leading eHealth Publisher

2019

The Adverse Drug Reactions From Patient Reports in Social Media Project: Protocol for an Evaluation Against a Gold Standard

Armelle Arnoux-Guenegou, MSc, PhD ; Yannick Girardeau, MSc, MD ; Xiaoyi Chen, MSc, PhD ; Myrtille Deldossi, MSc ; Rim Aboukhamis, M Pharm ; Carole Faviez, MSc ; Badisse Dahamna, MSc ; Pierre Karapetiantz, MSc ; Sylvie Guillemin-Lanne, MSc ; Agnès Lillo-Le Louët, MD, Prof Dr ; Nathalie Texier, M Pharm ; Anita Burgun, MD, PhD ; Sandrine Katsahian, MD, PhD, Prof Dr

ABSTRACT

Background: Social media is a potential source of information on postmarketing drug safety surveillance that still remains unexploited nowadays. Information technology solutions aiming at extracting adverse reactions (ADRs) from posts on health forums require a rigorous evaluation methodology if their results are to be used to make decisions. First, a gold standard, consisting of manual annotations of the ADR by human experts from the corpus extracted from social media, must be implemented and its quality must be assessed. Second, as for clinical research protocols, the sample size must rely on statistical arguments. Finally, the extraction methods must target the relation between the drug and the disease (which might be either treated or caused by the drug) rather than simple co-occurrences in the posts.

Objective: We propose a standardized protocol for the evaluation of a software extracting ADRs from the messages on health forums. The study is conducted as part of the Adverse Drug Reactions from Patient Reports in Social Media project.

Methods: Messages from French health forums were extracted. Entity recognition was based on *Racine Pharma* lexicon for drugs and Medical Dictionary for Regulatory Activities terminology for potential adverse events (AEs). Natural language processing–based techniques automated the ADR information extraction (relation between the drug and AE entities). The corpus of evaluation was a random sample of the messages containing drugs and/or AE concepts corresponding to recent pharmacovigilance alerts. A total of 2 persons experienced in medical terminology manually annotated the corpus, thus creating the gold standard, according to an annotator guideline. We will evaluate our tool against the gold standard with recall, precision, and f-measure. Interannotator agreement, reflecting gold standard quality, will be evaluated with hierarchical kappa. Granularities in the terminologies will be further explored.

Results: Necessary and sufficient sample size was calculated to ensure statistical confidence in the assessed results. As we expected a global recall of 0.5, we needed at least 384 identified ADR concepts to obtain a 95% CI with a total width of 0.10 around 0.5. The automated ADR information extraction in the corpus for evaluation is already finished. The 2 annotators already completed the annotation process. The analysis of the performance of the ADR information extraction module as compared with gold standard is ongoing.

Conclusions: This protocol is based on the standardized statistical methods from clinical research to create the corpus, thus ensuring the necessary statistical power of the assessed results. Such evaluation methodology is required to make the ADR information extraction software useful for postmarketing drug safety surveillance.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2019

Utilisation des médias sociaux dans l'étude de la qualité de vie des patients atteints de cancer et traités par immunothérapie : une étude infodémiologique

S.Schück, F.-E.Cotté, P.Voillot, B.Falissard, C.Tzourio, P.Foulquié, A.-F. Gaudin, H.Lemasson, C.Faviez

Objectifs

Dans de nombreux cancers, l'immunothérapie est devenue un standard de prise en charge au profil d'efficacité et de tolérance différent de celui des chimiothérapies. Le vécu et la qualité de vie (QdV) des patients traités par immunothérapie en vie réelle restent méconnus. Les médias sociaux sont de plus en plus un moyen d'expression et d'échange des patients sur leur maladie et leurs préoccupations. L'objectif était de décrire dans les médias sociaux les thématiques de discussion des patients traités ou ayant été traités par immunothérapie et en particulier d'étudier les messages mentionnant leur QdV.

Méthode

L'analyse portait sur les messages publiés par des patients ou leurs proches entre janvier 2011 et août 2018 sur de nombreux médias sociaux généralistes ou spécialisés en santé. Les messages d'intérêt ont été identifiés à l'aide d'une liste exhaustive de mots clés relatifs à l'immunothérapie (*anti-CTLA4, anti-PD1,* [noms des traitements et spécialités], etc.) incluant les possibles fautes d'orthographes. Les thématiques des messages étaient analysées de deux manières :

– manuellement par deux relecteurs indépendants ;

– automatiquement à l'aide d'un modèle de sujet (*topic model*).

Les mentions relatives à la QdV présentes dans les messages étaient classées selon huit dimensions prédéfinies à partir des principaux auto-questionnaires validés en oncologie (e.g. QLQ-C30, FACT-G).

Résultats

Au total, 267 messages ont été identifiés sur 19 sources différentes. Cela correspondait à 150 patients, dont une majorité de femmes (58 %). Le type de cancer était disponible pour 123 patients et les localisations les plus fréquents étaient le cancer du poumon (46 ; 37 %) et le mélanome (41 ; 33 %). Pour 44 % des patients, les messages étaient postés par un proche, en majorité un conjoint ou un enfant. L'occurrence des messages a augmenté de manière importante à partir de l'année 2017 (59 % des premiers messages de patients postés après 2017). Les thématiques de discussion identifiées manuellement concernaient : l'accès à l'immunothérapie (modalité d'accès, essais cliniques et traitement à l'étranger), le contexte de la prise d'une immunothérapie (espoir, information et description du traitement) et les effets de l'immunothérapie (efficacité et effets indésirables). L'application d'un *topic model* a permis d'identifier automatiquement 10 thématiques. Certaines identifiées manuellement étaient retrouvées (accès, essais cliniques, effets secondaires, parcours thérapeutique) alors que d'autres ont émergé (soutien des communautés de patients, fréquence des soins, cancer d'un proche). Pour 137 patients (91 %), l'expression d'au moins une dimension prédéfinie de QdV a été identifiée : la santé générale (115 ; 77 %), les symptômes (76 ; 51 %), l'état émotionnel (49 ; 33 %), les activités courantes (22 ; 15 %), l'état physique (13 ; 9 %), la situation professionnelle (9 ; 6 %), l'état cognitif (2 ; 1 %) et la vie sociale (2 ; 1 %).

Conclusion

Cette étude montre la diversité des thématiques abordées sur les médias sociaux par les patients atteints de cancer et traités par immunothérapie ainsi que leurs proches. La QdV apparaît comme un sujet central mentionné par la quasi-totalité des patients. L'ensemble des dimensions prédéfinies de QdV sont spontanément évoquées, en particulier la santé générale, les symptômes et l'état émotionnel.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2019

Livre blanc : Les chatbots en santé

Sanofi, Kap Code et Orange business service

Une initiative de SANOFI en partenariat avec Kap Code FROM DATA TO HEALTH et orange Business Services

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2019

Identification des facteurs de l'inconfort digestif à partir des témoignages sur les réseaux sociaux français.

C.Faviez, B.Le Nevé, F.Schäfer, J.-F.Jeanne, P.Voillot, M.Najm, G.Fagherazzi, S.Schück.

et dont les causes sont souvent mal identifiées. On estime par ailleurs que plus de 50 % des français sont actifs sur les réseaux sociaux. Un grand nombre d'entre eux partage des informations, des opinions et des expériences au sein de communautés, en particulier dans le domaine de la santé. L'objectif de cette étude consistait à identifier les facteurs associés à l'inconfort digestif par les utilisateurs à partir de messages publiés sur les réseaux sociaux.

Méthode

Les messages relatifs à l'inconfort digestif publiés entre janvier 2003 et août 2018 ont été extraits à partir de forums généralistes et spécialisés francophones. Les facteurs de l'inconfort digestif ont été identifiés à l'aide d'une méthode d'analyse automatique mixte combinant une analyse syntaxique à l'application d'un topic model (LDA). L'analyse syntaxique consistait dans un premier temps à identifier les messages contenant à la fois un symptôme de l'inconfort digestif et un terme de causalité (verbe ou conjonctions de coordination). La position des termes de causalité permettait de définir des segments de texte et des phrases sur lesquels les causes de l'inconfort étaient susceptibles d'être retrouvées. Un second topic model était ensuite appliqué sur ces segments de textes.

Résultats

Après nettoyage, 198 866 messages issus de 14 forums généralistes et spécialisés ont été extraits. La détection de termes de causalité au sein des messages contenant des symptômes (29 935 messages) a permis d'identifier 35 220 termes de causalité au sein de 20 500 messages, soit dans 10 % du corpus total. Un topic model a été appliqué sur les segments de phrases de ces messages associés aux termes de causalité. Le nombre de topics a été fixé arbitrairement à 30 afin qu'un grand nombre de thématiques puisse émerger. La revue manuelle de ces topics a permis d'en caractériser 10 contenant des facteurs de l'inconfort digestif. Ces 10 topics d'intérêt ont pu être regroupés en 7 catégories : les facteurs psychologiques (stress, psychologique, angoisse ; 14,5 % des messages contenant des symptômes), la nutrition (repas, nourriture, digérer ; 10,8 %), les allergènes/intolérances perçues (éviter, gluten, fibres ; 9,5 %), les pathologies gastro-intestinales (hernie, ulcère, bactérie ; 9 %), les facteurs gynécologiques (enceinte, hormones, pilule ; 6,3 %), les facteurs sociaux (famille, aime, perdre ; 5,2 %) et enfin les complications médicales (opération, urgences, ablation ; 3,6 %).

Conclusion

De nombreux internautes échangent fréquemment sur leur inconfort digestif ainsi que sur ce qu'ils estiment être la cause de cet inconfort. La méthode innovante qui a été mise en place a permis d'identifier un ensemble de facteurs considérés comme des causes de l'inconfort par les internautes. Ces causes sont diverses et incluent le contexte psychologique (facteurs psychologiques et sociaux), les facteurs médicaux (pathologies gastro-intestinales, facteurs gynécologiques et complications médicales) et l'alimentation (nutrition et allergènes/intolérances perçues). Ce travail a montré que l'information issue des réseaux sociaux permettait de compléter l'information existante concernant l'inconfort digestif et la compréhension de ses causes en incluant des informations de vie réelle issues de témoignages d'internautes.

Previous article in issue

Next article in issue

Copyright © 2019 Published by Elsevier Masson SAS

Advertise Contact and support Terms and conditions Privacy policy

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2019

Étude de l'utilisation de l'inhalateur connecté Connect'inh en vie réelle et de sa pertinence pour répondre aux besoins des patients asthmatiques et atteints de bronchopneumopathie chronique obstructive.

S.Renner, A.Nigar, A.Mebarki, J.Olivier, V.Marie-Joseph, N.Texier, J.Koman, S.Schück

Connect'inh est un dispositif connecté développé par la start-up Kap Code et destiné aux patients atteints de pathologies respiratoires chroniques comme l'asthme ou la bronchopneumopathie chronique obstructive (BPCO). Cette étude a été réalisée en condition réelle d'utilisation afin de s'assurer de la pertinence de l'objet pour répondre aux besoins des patients mais aussi pour évaluer au cours du temps le degré d'appropriation de l'objet par les utilisateurs finaux de Connect'inh.

Méthode

L'étude a été menée avec des patients atteints de pathologies respiratoires (asthme et BPCO) recrutés via les réseaux sociaux (Twitter, Facebook, LinkedIn, Google). Les patients ont été suivis pendant trois mois. Au total, 42 personnes (18 hommes et 24 femmes) ont participé à l'étude : 14 % avaient entre 18 et 25 ans, 43 % entre 25 et 34 ans, 26 % entre 35 et 49 ans et 17 % plus de 50 ans. Trois questionnaires ont été administrés en ligne : le premier questionnaire avait pour but d'explorer le profil des malades et leurs attentes concernant un objet connecté, avant la prise en main de l'objet Connect'inh. Un deuxième questionnaire a été administré une semaine après la prise en main de l'objet puis, un troisième après trois mois d'utilisation. Une analyse statistique a ensuite permis de classifier et mettre en regard les attentes des patients et la réponse aux besoins en vie réelle grâce à l'objet Connect'inh.

Résultats

L'analyse des questionnaires préalables a pu montrer que 63 % des patients estimaient que le suivi de leur pathologie chronique respiratoire n'était pas suffisant, et qu'ils souhaitaient un outil adaptable permettant un suivi personnalisé (suivi des crises, information sur les traitements de fond ou de crise) couplé à des informations sur l'environnement (notification sur la qualité de l'air) ; 100 % des patients ont estimé qu'un inhalateur connecté comme Connect'inh pouvait répondre à leurs besoins. Les patients n'ont pas été réticent à l'utilisation d'un objet connecté ou à l'utilisation de la géolocalisation via l'application, et ce, quel que soit l'âge des participants aux questionnaires. Au final, après une semaine d'utilisation, 63 % des patients bêta-testeurs estimaient que Connect'inh répondait à leurs besoins, et 78 % recommanderaient le produit à une autre personne souffrant d'une pathologie respiratoire chronique.

Conclusion

Cette étude a permis de s'assurer de la bonne appropriation d'un objet connecté et sa pertinence en condition réelle d'utilisation par des patients atteints de pathologies respiratoires. L'analyse des données met en évidence que les patients asthmatiques et/ou atteints de BPCO sont aujourd'hui en majorité demandeurs de davantage de suivi pour leur pathologie respiratoire, et qu'un objet connecté pourrait être un moyen d'y parvenir. Dans les conditions réelles d'utilisation, les patients ont estimé que Connect'inh pouvait répondre à leurs besoins. Un suivi à plus long terme sera nécessaire pour s'assurer de l'appropriation de l'inhalateur connecté dans la durée et s'intéresser à l'impact d'un tel dispositif sur la fréquence et la gravité des crises.

Previous article in issue

Next article in issue

Copyright © 2019 Published by Elsevier Masson SAS

rtise Contact and support Terms and conditions Privacy policy

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2019

Identification des facteurs de l'inconfort digestif à partir des témoignages sur les réseaux sociaux français.

C.Faviez, B.Le Nevé, F.Schäfer, J.-F.Jeanne, P.Voillot, M.Najm, G.Fagherazzi, S.Schück.

Kap Code FROM DATA TO HEALTH — DANONE NUTRICIA RESEARCH

IDENTIFICATION DES FACTEURS DE L'INCONFORT DIGESTIF À PARTIR DES TÉMOIGNAGES SUR LES FORUMS DE DISCUSSION FRANÇAIS

Carole Faviez[1], Boris Le Nevé[2], Florent Schäfer[2], Jean-François Jeanne[2], Paméla Voillot[1], Matthieu Najm[1], Guy Fagherazzi[3], Stéphane Schück[1]

[1]Kap Code, Paris, France
[2]Danone Research, Palaiseau, France
[3]Centre de recherche en Épidémiologie et Santé des Populations, UMR1018 INSERM, Institut Gustave Roussy, université Paris-Sud Saclay, Villejuif, France

OBJECTIF ET MÉTHODOLOGIE

L'inconfort digestif est une affection fréquente qui touche une part grandissante de la population. **Les causes de cet inconfort sont souvent mal identifiées. On estime par ailleurs que plus de 50% des français sont actifs sur les réseaux sociaux.** Un grand nombre d'entre eux partage des informations et leurs expériences, en particulier dans le domaine de la santé.
L'objectif de cette étude consistait, à partir de messages publiés sur les forums de discussion relatifs à l'inconfort digestif, à identifier les thèmes abordés par les utilisateurs ainsi que les facteurs qu'ils associaient à leurs troubles.

Les messages relatifs à l'inconfort digestif publiés **entre janvier 2003 et août 2018** ont été extraits depuis un ensemble de forums généralistes et spécialisés francophones et anonymisés.
L'âge et le sexe des utilisateurs étaient identifiés via la détection d'expressions régulières au sein des messages.
Les thématiques de discussions étaient identifiées à l'aide d'un topic model (*Correlated Topic Model*).

Les facteurs de l'inconfort digestif ont été identifiés à l'aide d'une méthode d'analyse automatique mixte combinant une analyse syntaxique à l'application d'un topic model (*LDA*). L'analyse syntaxique consistait dans un premier temps à identifier les messages contenant à la fois un symptôme de l'inconfort digestif et un terme de causalité (verbe ou conjonctions de coordination).
La position des termes de causalité permettait de définir des segments de texte et des phrases sur lesquels les facteurs de l'inconfort étaient susceptibles d'être retrouvés (Fig. 1).
Un topic model était ensuite appliqué sur ces segments de textes. Les thèmes identifiés étaient manuellement revus afin de déterminer les facteurs de l'inconfort.

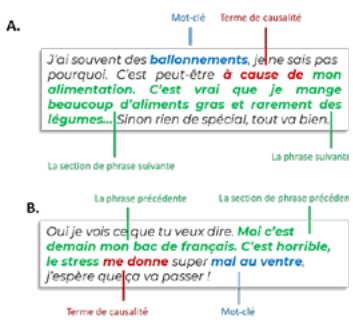


Figure 1 – Exemple explicatif. (A) terme de causalité associé à la droite. (B) terme de causalité associé à la gauche

RÉSULTATS

Après nettoyage, **198 866 messages publiés entre 2003 et 2018 sur 14 forums différents ont été retenus.**
Les utilisateurs étaient principalement des femmes ayant moins de 30 ans (Tab.1).

Tableau 1 – Caractéristiques des internautes

Genre	Nombre d'utilisateurs
Femmes	12 071
Hommes	2 370
Inconnu	22 548
Groupes d'âge	**Nombre d'utilisateurs**
0-20	1 401
20-30	1 883
30-40	825
40-50	396
50-60	211
60+	86
Inconnu	32 187

Dix-huit thèmes ont été identifiés (Fig.2). Des analyses détaillées ont été réalisées sur certaines catégories par application d'un second topic model. Au sein de la catégorie « alimentation », des sous-thèmes relatifs à la digestion difficile, aux régimes alimentaires, au microbiote intestinal et aux intolérances alimentaires ont émergé. Le nombre de messages relatifs à ces deux dernières catégories était en augmentation sur la dernière année. L'analyse du thème « stress et symptômes » a mis en évidence des sous-thèmes variés : le stress comme cause des troubles digestifs, les solutions pour l'éviter (acupuncture, activité sportive, traitements médicamenteux) et son impact sur la vie sociale/la qualité de vie.

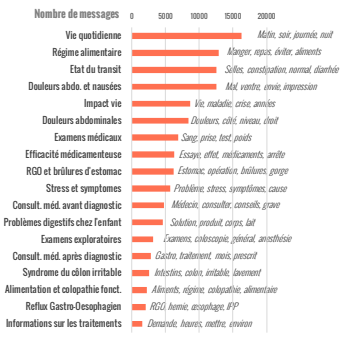


Figure 2 – Thèmes identifiés dans le corpus et mots caractéristiques

29 935 messages contenant des symptômes digestifs ont été identifiés. **35 220 termes de causalité** ont été détectés au sein de 20 500 messages (Fig. 3).

Un topic model a été appliqué sur les segments de phrases de ces messages associés aux termes de causalité.

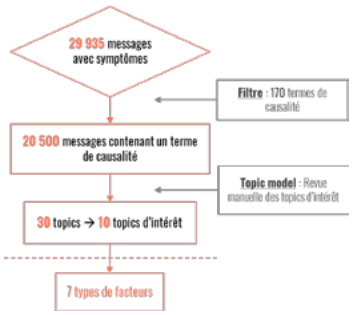


Figure 3 – Flowchart de l'analyse des facteurs

Le nombre de topics a été fixé arbitrairement à 30 afin qu'un grand nombre de thématiques puisse émerger. La revue manuelle de ces topics a permis d'en retenir 10 correspondant à 7 types de facteurs de l'inconfort digestif (Tab. 2).
Les principaux facteurs identifiés correspondaient à l'alimentation (Nutrition et Allergènes / Intolérances perçues) et aux facteurs psychologiques.

Tableau 2 – Facteurs identifiés

Facteurs	% messages	Mots caractéristiques
Facteurs psychologiques	14.5%	stress, psychologique, crise, pure, angoisse
Nutrition	10.8%	repas, ouverture, digérer, mangeurs, café
Allergènes/Intolérances perçues	9.5%	évités, gluten, fibres, intolérance, laitiers
Pathologies gastro-intestinales	9%	hernie, ulcère, hémorroïde, gastrite, bactérie
Facteurs gynécologiques	6.3%	enceinte, règles, pilule, cycle, hormones
Facteurs sociaux	5.2%	famille, parents, sexe, perdre, fille
Complications médicales	3.6%	opératoire, urgences, hôpital, chirurgien, ablation

CONCLUSION

Cette étude a mis en évidence la présence d'un grand nombre de messages relatifs à l'inconfort digestif sur les forums de discussion francophones. Les internautes échangent sur leurs troubles, recherchent des solutions et décrivent leur parcours de soins.
La méthode innovante qui a été mise en place a permis d'identifier un ensemble de facteurs considérés comme des causes de l'inconfort par les internautes. Ces facteurs sont divers et incluent le contexte psychologique, l'alimentation et les facteurs médicaux.
Les informations de vie réelle issues de témoignages d'internautes pourraient permettre de compléter la connaissance existante sur les causes et les conséquences de l'inconfort digestif.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2019

Étude de l'utilisation de l'inhalateur connecté Connect'inh en vie réelle et de sa pertinence pour répondre aux besoins des patients asthmatiques et atteints de bronchopneumopathie chronique obstructive.

S.Renner, A.Nigar, A.Mebarki, J.Olivier, V.Marie-Joseph, N.Texier, J.Koman, S.Schück

Kap Code

ÉTUDE DE L'UTILISATION DE L'INHALATEUR CONNECTÉ CONNECT'INH EN VIE RÉELLE ET DE SA PERTINENCE POUR RÉPONDRE AUX BESOINS DES PATIENTS ASTHMATIQUES ET ATTEINTS DE BPCO

Simon Renner[1], Archange Nigar[1], Adel Mebarki[1], Juliette Olivier[1], Vanessa Marie-Joseph[1], Nathalie Texier[1], Jason Koman[1], Stéphane Schück[1]

[1]Kap Code, Paris, France

INTRODUCTION

Connect'inh est un **dispositif connecté** développé par la start-up Kap Code et destiné aux **patients atteints de pathologies respiratoires chroniques comme l'asthme ou la BPCO.** Cette étude a été réalisée en condition réelle d'utilisation afin de s'assurer de la pertinence de l'objet pour répondre aux besoins des patients. L'étude a aussi pour but d'évaluer le degré d'appropriation de l'objet par les utilisateurs cibles de Connect'inh au cours du temps.

MATÉRIEL & MÉTHODES

L'étude a été conduite auprès de **62 patients** atteints de pathologies respiratoires chroniques et **recrutés via les réseaux sociaux (Twitter, Facebook, LinkedIn, Google).** Les patients ont été suivis pendant 1 semaine et l'étude s'est déroulée en 2 étapes :

- Un premier questionnaire avait pour but d'explorer le profil des malades et leurs attentes concernant un objet connecté, avant la prise en main de l'objet Connect'inh.
- Un deuxième questionnaire était administré une semaine après la prise en main de l'objet.

Des analyses statistiques ont ensuite permis de **comparer dans un contexte de vie réelle les attentes des patients quant à l'amélioration de leur prise en charge aux réponses apportées par Connect'inh.**

RÉSULTATS

Parmi les 62 patients (fig.1), **97%** étaient atteints d'asthme, associé dans **61%** des cas à des allergies. Plus de **74%** d'entre eux étaient malades depuis plus de 5 ans.

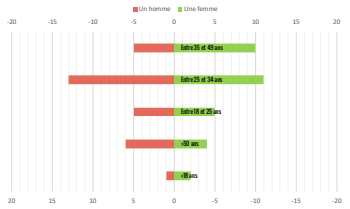


Figure 1 – Pyramides des âges des patients

L'analyse des questionnaires a permis de montrer que **40%** des participants à l'étude rencontraient des difficultés au niveau de leur prise en charge thérapeutique. Parmi les patients sous traitement (tab.1), **89%** utilisaient un inhalateur à chaque crise.

TRAITEMENT	EFFECTIFS
De crise et de fond	37
De crise	17
De fond	8

Tableau 1 – Effectifs par traitements (crise ou fond)

40% des patients utilisaient des objets connectés en santé (dont 4% pour leur pathologie respiratoire).
L'analyse du premier questionnaire met en évidence que **100%** des patients estimaient qu'un inhalateur connecté comme Connect'inh pourrait répondre à leurs besoins et les aiderait à améliorer leur prise en charge (fig.2).

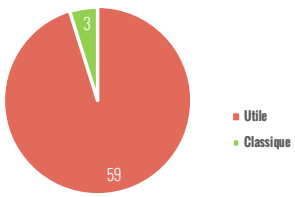


Figure 2 – Qualification de Connect'inh avant sa réception

L'analyse du second questionnaire, après 1 semaine d'utilisation de Connect'inh, a montré que dans **63%** des cas, ce dernier répondait mieux à leurs besoins que d'autres alternatives (fig. 3 et 4). **78%** recommanderaient le produit.

Figure 3 – Utilité du dispositif

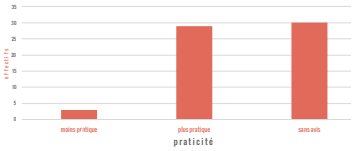


Figure 4 – Praticité du dispositif

Ces patients étaient satisfaits après utilisation du produit dans plus de **76%** des cas (fig. 5).

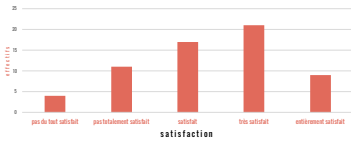


Figure 5 – Satisfaction générale des patients utilisateurs de Connect'inh

De façon générale, les utilisateurs appréciaient avant tout d'avoir la possibilité d'un suivi au long terme de leur pathologie via une solution pratique et simple d'utilisation (fig. 6)

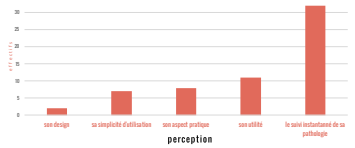


Figure 6 – Perception générale après 1 semaine d'utilisation

CONCLUSION

Cette étude a permis de s'assurer de la bonne appropriation d'un objet connecté et sa pertinence en condition réelle d'utilisation par des patients atteints de pathologies respiratoires.

L'analyse des données met en évidence que ces patients sont aujourd'hui en majorité demandeurs de davantage de suivi, et qu'un objet connecté pourrait être un moyen d'y parvenir.

Dans les conditions réelles d'utilisation, les patients ont estimé que Connect'inh pouvait répondre à leurs besoins.
Un suivi à plus long terme est en cours pour s'assurer de l'appropriation de l'inhalateur connecté dans la durée et s'intéresser à l'impact d'un tel dispositif sur la fréquence et la gravité des crises.

CONNECT'INH

*Les données de l'étude ont été mises à jour depuis la publication de l'abstract

ISPOR
Improving healthcare decisions

2019

HEALTH-RELATED QUALITY OF LIFE (HRQOL) OF CANCER PATIENTS TREATED WITH IMMUNOTHERAPY: USE OF SOCIAL MEDIA IN FRENCH LANGUAGE TO EXPLORE RELEVANCE OF CONCEPTS COVERED BY CANCER GENERIC HRQOL MEASURES MESSAGES FROM FRENCH-SPEAKING PLATFORMS

C. Faviez, B. Bennett, P. Voillot, S. Schück, B. Falissard, C.T. Tzourio, P. Foulquié, A.F. Gaudin, H. Lemasson, V. Grumberg, L. McDonald, F.E. Cotte

Objectives

Immune-checkpoint inhibitors (ICI) are emerging as the standard of care for many cancers. However, HRQoL of patients treated with ICI in the real-world remain largely unknown. Social media is increasingly used by patients to express their views about their illness and treatment experience. The objective was to assess the conceptual similarity between cancer generic HRQoL measures and cancer patients' experience described in social media.

Methods

Patient messages were retrieved from 19 different French social media sources between Jan-2011 and Aug-2018. Messages of interest were extracted using automatic processes and manual searches. Extracted messages were analysed by two independent reviewers. HRQoL statements were classified according to predefined HRQoL dimensions in the *European Organisation for Research and Treatment of Cancer* (EORTC) QLQ-C30 and FACT-G.

Results

Overall, 137 ICI-treated patients posted HRQoL-related messages on social media. Cited dimensions of HRQoL were global health (115 patients), symptoms (76; mean: 2.1 per patient), emotion (49), role (22), physical functioning (13), professional situation (2) and cognitive state (2). Both the QLQ-C30 and FACT-G cover global health dimension expressed by patients in their messages. A total of 13 symptoms were identified from messages; the QLQ-C30 captured 5, whereas the FACT-G captured 3. Ten themes related to emotional functioning were identified, 2 were covered by the QLQ-C30 and 3 by the FACT-G. Seven themes were retrieved describing patients' role; the QLQ-C30 captured 5, whereas the FACT-G captured 3. In terms of physical functioning, 4 themes were identified; the QLQ-C30 captured all themes, whereas the FACT-G captured 1.

Conclusions

Conceptual similarity between HRQoL themes from social media and the most commonly used generic HRQoL measures was generally suboptimal. Careful interpretation is required due to the relatively low sample size. A larger study on social media would be needed to assess correlation with tumor specific questionnaires of HRQoL.

JMIR Publications
Advancing Digital Health & Open Science

2020

Exploring the Health-Related Quality of Life of Patients Treated With Immune Checkpoint Inhibitors: Social Media Study

Cotté FE, Voillot P, Bennett B, Falissard B, Tzourio C, Foulquié P, Gaudin AF, Lemasson H, Grumberg V, McDonald L, Faviez C, Schück S

ABSTRACT

Background: Immune checkpoint inhibitors (ICIs) are increasingly used to treat several types of tumors. Impact of this emerging therapy on patients' health-related quality of life (HRQoL) is usually collected in clinical trials through standard questionnaires. However, this might not fully reflect HRQoL of patients under real-world conditions. In parallel, users' narratives from social media represent a potential new source of research concerning HRQoL.

Objective: The aim of this study is to assess and compare coverage of ICI-treated patients' HRQoL domains and subdomains in standard questionnaires from clinical trials and in real-world setting from social media posts.

Methods: A retrospective study was carried out by collecting social media posts in French language written by internet users mentioning their experiences with ICIs between January 2011 and August 2018. Automatic and manual extractions were implemented to create a corpus where domains and subdomains of HRQoL were classified. These annotations were compared with domains covered by 2 standard HRQoL questionnaires, the EORTC QLQ-C30 and the FACT-G.

Results: We identified 150 users who described their own experience with ICI (89/150, 59.3%) or that of their relative (61/150, 40.7%), with 137 users (91.3%) reporting at least one HRQoL domain in their social media posts. A total of 8 domains and 42 subdomains of HRQoL were identified: Global health (1 subdomain; 115 patients), Symptoms (13; 76), Emotional state (10; 49), Role (7; 22), Physical activity (4; 13), Professional situation (3; 9), Cognitive state (2; 2), and Social state (2; 2). The QLQ-C30 showed a wider global coverage of social media HRQoL subdomains than the FACT-G, 45% (19/42) and 29% (12/42), respectively. For both QLQ-C30 and FACT-G questionnaires, coverage rates were particularly suboptimal for Symptoms (68/123, 55.3% and 72/123, 58.5%, respectively), Emotional state (7/49, 14% and 24/49, 49%, respectively), and Role (17/22, 77% and 15/22, 68%, respectively).

Conclusions: Many patients with cancer are using social media to share their experiences with immunotherapy. Collecting and analyzing their spontaneous narratives are helpful to capture and understand their HRQoL in real-world setting. New measures of HRQoL are needed to provide more in-depth evaluation of Symptoms, Emotional state, and Role among patients with cancer treated with immunotherapy.

J Med Internet Res 2020;22(9):e19694

JMIR Publications
Advancing Digital Health & Open Science

2020

Assessing Patient Perceptions and Experiences of Paracetamol in France: An Infodemiology Study Using Social Media Data Mining

Stéphane Schück; Avesta Roustamal; Anaïs Gedik; Paméla Voillot; Pierre Foulquié; Catherine Penfornis; Bernard Job

ABSTRACT

Background:

Frequently, individuals are turning to social media to discuss medical conditions and medication, sharing their experiences and information and asking questions among themselves. These online discussions can provide valuable insights into individuals' perceptions of medical treatment, and increasingly, studies are focusing on the potential use of this information to improve healthcare management.

Objective:

The objective of this infodemiology study was to identify social media posts mentioning paracetamol-containing products, to develop a better understanding of the patients' opinions and perceptions of the drug.

Methods:

Posts containing at least one mention of paracetamol were extracted from 18 French forums between January 2003 and March 2019 with the use of the Detec't webcrawler. Posts were then analyzed using the automated Detec't tool which uses machine-learning and text-mining methods to inspect social media posts and extract relevant content.

Results:

Overall, 44,283 posts were analyzed from 20,883 different users. Post volume over the study period showed a peak in activity between 2009 and 2012, as well as a spike in 2017 in the General group. The number of posts tended to be higher during winter each year. Posts were made predominantly by women (71%), with 12% made by men and 17% by individuals of unknown gender. The mean age of web users was 39 (±19) years. In the General group, pain was the most common medical concept discussed (22,257 posts, 50%) and paracetamol risk was the most common discussion topic, addressed in 8,902 posts (20.36%). Doliprane® was the most common medication mentioned (14,058 posts, 32%) within the General group, and tramadol was the most commonly mentioned drug in combination with paracetamol in the General group (1,038 posts, 5%). The most common unapproved indication mentioned within the Paracetamol Only group was fatigue (190 posts, with 16% positive for an unapproved indication), with reference to dependence made by 0.79% of the web users, accounting for 1.33% of the posts in the Paracetamol Only group. Dependence mentions in the Paracetamol and Opioids group were provided by 3.64% of web users, accounting for 5.44% of total posts. Reference to overdose was made by 245 web users across 291 posts within the Paracetamol Only group. The most common potential adverse event (PAE) detected was nausea (2.38% of posts) within the Paracetamol Only group

Conclusions:

The use of social media mining with the Detec't tool provided valuable information on the perceptions and understanding of the web users, highlighting areas where providing more information for the general public on paracetamol, as well as other medications, may be of benefit.

JMIR Publications
Advancing Digital Health & Open Science

2020

Mapping and Modeling of Discussions Related to Gastrointestinal Discomfort in French-Speaking Online Forums: Results of a 15-Year Retrospective Infodemiology Study

Schäfer F, Faviez C, Voillot P, Foulquié P, Najm M, Jeanne JF, Fagherazzi G, Schück S, Le Nevé B

ABSTRACT

Background: Gastrointestinal (GI) discomfort is prevalent and known to be associated with impaired quality of life. Real-world information on factors of GI discomfort and solutions used by people is, however, limited. Social media, including online forums, have been considered a new source of information to examine the health of populations in real-life settings.

Objective: The aims of this retrospective infodemiology study are to identify discussion topics, characterize users, and identify perceived determinants of GI discomfort in web-based messages posted by users of French social media.

Methods: Messages related to GI discomfort posted between January 2003 and August 2018 were extracted from 14 French-speaking general and specialized publicly available online forums. Extracted messages were cleaned and deidentified. Relevant medical concepts were determined on the basis of the Medical Dictionary for Regulatory Activities and vernacular terms. The identification of discussion topics was carried out by using a correlated topic model on the basis of the latent Dirichlet allocation. A nonsupervised clustering algorithm was applied to cluster forum users according to the reported symptoms of GI discomfort, discussion topics, and activity on online forums. Users' age and gender were determined by linear regression and application of a support vector machine, respectively, to characterize the identified clusters according to demographic parameters. Perceived factors of GI discomfort were classified by a combined method on the basis of syntactic analysis to identify messages with causality terms and a second topic modeling in a relevant segment of phrases.

Results: A total of 198,866 messages associated with GI discomfort were included in the analysis corpus after extraction and cleaning. These messages were posted by 36,989 separate web users, most of them being women younger than 40 years. Everyday life, diet, digestion, abdominal pain, impact on the quality of life, and tips to manage stress were among the most discussed topics. Segmentation of users identified 5 clusters corresponding to chronic and acute GI concerns. Diet topic was associated with each cluster, and stress was strongly associated with abdominal pain. Psychological factors, food, and allergens were perceived as the main causes of GI discomfort by web users.

Conclusions: GI discomfort is actively discussed by web users. This study reveals a complex relationship between food, stress, and GI discomfort. Our approach has shown that identifying web-based discussion topics associated with GI discomfort and its perceived factors is feasible and can serve as a complementary source of real-world evidence for caregivers.

JMIR Publications
Advancing Digital Health & Open Science

2020

Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study

Koman J, Fauvelle K, Schuck S, Texier N, Mebarki A

ABSTRACT

Background: Seeking medical information can be an issue for physicians. In the specific context of medical practice, chatbots are hypothesized to present additional value for providing information quickly, particularly as far as drug risk minimization measures are concerned.

Objective: This qualitative study aimed to elicit physicians' perceptions of a pilot version of a chatbot used in the context of drug information and risk minimization measures.

Methods: General practitioners and specialists were recruited across France to participate in individual semistructured interviews. Interviews were recorded, transcribed, and analyzed using a horizontal thematic analysis approach.

Results: Eight general practitioners and 2 specialists participated. The tone and ergonomics of the pilot version were appreciated by physicians. However, all participants emphasized the importance of getting exhaustive, trustworthy answers when interacting with a chatbot.

Conclusions: The chatbot was perceived as a useful and innovative tool that could easily be integrated into routine medical practice and could help health professionals when seeking information on drug and risk minimization measures.

J Med Internet Res 2020;22(11):e15185

JMIR Publications
Advancing Digital Health & Open Science

2020

Comparison of topics and concerns discussed on Chinese and French social media during the COVID-19 lockdown

Stéphane Schück; Pierre Foulquié; Adel Mebarki; Carole Faviez; Mickaïl Khadhar; Nathalie Texier; Sandrine Katsahian; Anita Burgun; Xiaoyi Chen

ABSTRACT

Background:

During the coronavirus disease 2019 (COVID-19) pandemic, numerous countries, including China and France, have implemented lockdown measures that have been shown to be effective in controlling the epidemic. However, little is known about the impact of these measures on the population as expressed on social media from different cultural contexts.

Objective:

To assess and compare the evolution of the topics discussed on Chinese and French social media during the COVID-19 lockdown.

Methods:

We extracted posts containing "COVID-19"- or "lockdown"-related keywords in the most commonly used micro-blogging social media platforms, i.e., Weibo (China) and Twitter (France), from one week before to the lifting of the lockdown. A topic model was applied independently for three periods: pre-lockdown, early lockdown and mid-to-late lockdown, to assess the evolution of the topics discussed on Chinese and French social media.

Results:

6 395, 23 422 and 141 643 Chinese Weibo messages, and 34 327, 119 919, and 282 965 French tweets were extracted in the pre-lockdown, early lockdown and mid-to-late lockdown periods in China and France, respectively. Four categories of topics were discussed in a continuously evolving way in all three periods: epidemic news and everyday life, scientific information, public measures and solidarity & encouragement. The most represented category over all periods in both countries was epidemic news and everyday life. Scientific information was far more discussed on Weibo than in French tweets. Misinformation circulated through social media in both countries; however, it was more concerned with the virus and epidemic in China, whereas it was more concerned with the lockdown measures in France. Regarding public measures, more criticisms were identified in French tweets than on Weibo. Advantages and data privacy concerns regarding tracing apps were also addressed in French tweets.

Conclusions:

This study is the first to compare the social media content in Eastern and Western countries during the unprecedented COVID-19 lockdown. Our results describe common and different public reactions behaviors and concerns, and can help characterize country-specific public needs and appropriately address them during an outbreak.

ISPOR
Improving healthcare decisions

2020

Social media as a tool to monitor factors for HPV vaccine hesitancy in France

Simon Renner, Tom Marty, Pamela Voillot, Pierre Foulquie, Adel Mebarki, and Stéphane Schück

Social media as a tool to monitor factors for HPV vaccine hesitancy in France

Simon Renner[1], Tom Marty[1], Pamela Voillot[1], Pierre Foulquie[1], Adel Mebarki[1], and Stéphane Schück[1]
[1]Kap Code, 28 rue d'Enghien, 75010 Paris ; Kap Code ; France

INTRODUCTION

Vaccination against human papillomavirus (HPV) infection is recommended as it is a global public health issue. In 2019, France came in last position in Europe, with only 21% vaccination coverage. On the other hand, social networks are a privileged medium for more than 32 million active Internet users. These platforms hold a growing role as a place where Internet users share their health, their concerns and often their position regarding vaccination. Analyzing and understanding these testimonies would help identify the drivers associated with an improved HPV vaccination coverage.

MATERIAL & METHODS

The study was conducted via the EVANEX© immunization observatory. Posts associated with HPV vaccination and written in French (dataset HPV) were extracted from 23 medical forums and social networks, between 2006 and 2019. The extraction terms either referred to an act of HPV immunization "papillomavirus vaccine" or directly to a product name (Gardasil® or Cervarix®). A manual annotation of posts (presence or not of vaccine hesitancy) enabled the creation of a Gold standard aimed at developing an algorithm for the detection of vaccine hesitancy. This algorithm was then applied to the entire HPV dataset. A first corpus of analysis (pre-corpus of vaccine hesitancy, Figure 1) was created. A manual annotation of this corpus was performed in order to identify the different determinants of hesitation (corpus of vaccine hesitancy, Figure 1) and cluster these factors into major hesitation topics.

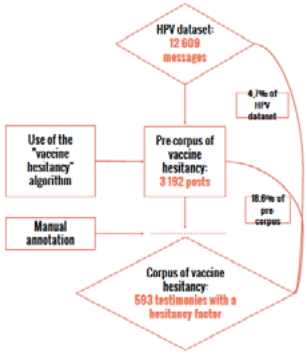


Figure 1: Data Collection

RESULTS

The HPV dataset included 12 609 unique posts written by 5,117 different web users. A pre-corpus of 3 192 posts specific to vaccine hesitancy was determined algorithmically. The manual analysis of these 3,192 testimonies allowed the characterization of 593 posts with an hesitancy (18.6% of the pre-corpus, 4.7% of the HPV dataset) (Figure 1).

« [...] It's been about a year and a half since I lost my virginity. My mother wants me to get the HPV vaccine but I don't want to tell her I'm no longer a virgin, so is there a risk for me to get vaccinated?" – Translated from French

"Hello, I am 17 years old and I have already had sex with only one person, this person has only had sex with me too. I am not vaccinated against cervical cancer and I would like to know if it is still possible?" – Translated from French

Figure 2: Examples of posts collected

Posts (Figure 2) were grouped into 3 clusters, according to the factors of vaccine hesitation identified.
The first is based on the influence of sexuality on immunization (339 posts) (7.5% pre-corpus). Within this group, teenage girls questioned both the need to be vaccinated but also the possible risk of immunization once sexually active (n=113; Figure 3), or if a relationship occurs between two injections (n=70). Fear of adverse events due to sexual activity was important.

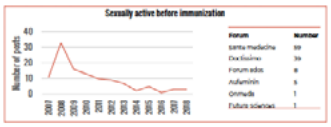


Figure 3: Sexually Active - Volumetrics

In addition, partners are questioning themselves on the same issues (n=30). Fear of disclosing one's sex life to parents is also a source of concern (n=126; Figure 4). Teenage girls are looking for a solution to inform their doctor about their sexual activity without telling parents, who are also present during the medical visit.

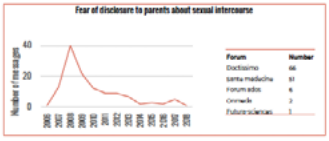


Figure 4 : Fear of disclosure to parents about sexual intercourse - Volumetrics

The second cluster relates to the lack of information (196 posts) (5.6% pre-corpus). This lack of information concerns overall (n=60, Figure 5) vaccine information: obligation, efficacy, utility, reimbursement, etc.

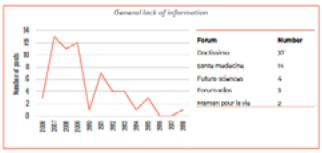


Figure 5: General lack of information - Volumetrics

We observe lack of information relating to recommendations on injection schedules and booster shots (n=16), or the age limit for immunization, especially if the patient is not sexually active (n=15).

The third cluster shows the influence of different external sources on vaccine hesitancy (n=59, 1.8% of the pre-corpus). These posts, for people who are being vaccinated against HPV or not, highlight a fear of adverse effects (n=30; Figure 6) or negative aspects of injections (n=20) after browsing through a website, testimonials or an online discussion.
Nine testimonies reported a medical source explicitly advising against HPV vaccination despite parents' and adolescent's desire to be vaccinated.

Figure 6 : Information on side effects, resulting issues - Volumetrics

Belonging to groups 2 and 3, mothers (n=39) justified reluctance to have their daughter vaccinated by the lack of clear and concordant information.

CONCLUSION

This study highlighted the existence of a strong community of web users debating and seeking information about HPV immunization.
The analysis of these testimonies allowed the identification of two main drivers of reluctance to immunization: impact of sexual activity and lack of information.
A detailed analysis of these factors of hesitancy would make it possible to identify the associated levers. On local scales, depending on the behaviors and characteristics of the population, this type of infodemiological study could serve public health strategies and policies to improve immunization coverage.
This work on HPV vaccine hesitation was carried out before the Haute Autorité de Santé recommendation to extend immunization to teenage boys in 2019. This study could be conducted on vaccine hesitancy expressed by teenage boys on the web, in order to analyze causes of hesitation.

ISPOR
Improving healthcare decisions

2020

Patients Comments on Social Networks about Paracetamol Misuses

Avesta Roustamal; Anaïs Gedik; Pierre Foulquié; Paméla Voillot; Adel Mebarki; Nathalie Texier; Stéphane Schück

PATIENTS' COMMENTS ON SOCIAL NETWORKS ABOUT PARACETAMOL MISUSES

Avesta Roustamal[1]; Anaïs Gedik[1]; Pierre Foulquié[1]; Paméla Voillot[1]; Adel Mebarki[1]; Nathalie Texier[1]; Stéphane Schück[1]
[1]Kap Code, Paris, France

INTRODUCTION

Paracetamol is one of the Essential Medicines according to WHO. It is one of the most common painkiller but misuses and overdoses occur.

Recently, the ANSM has launched a public consultation in order to raise awareness of the toxicity risks linked to paracetamol in case of misuse.
The social media are an important way for patients to share experiences regarding health issues, treatments and illness [Van Stekelenborg, J. et al. 2019]. These discussions on social media could be a potential new source of data to detect early potential signals. Within this context, Kap Code has developped Detect' an automatic tool for analyzing social media based on artificial intelligence and text mining methods that appears useful to analyze posts concerning paracetamol.
The main objective is to realize an infodemiology study of patients expressing on social media about paracetamol-containing products. More specifically, the goal is to detect misuse cases including overdose, unapproved indications and dependance discussed by web users about paracetamol.

METHODOLOGY

The posts in french language from social media sources were extracted with Detect' tool by detecting in social media discussions key words related to paracetamol containing products such as «Paracetamol»,«Doliprane», « Actifed »... The posts were then, collected and agregated. The project does not aim to collect adverse events as the data were not treated individually. The data were cleaned in order to apply the algorithms available with Detect' tools. These methods of analysis were used by ANSM in collaboration with Kap Code[P.Foulquié et al. 2018].
The monitoring posts volume were calculated by applying a Markov chain-based algorithm on posts in order to qualify the volume of activity (number of posts published by web users that can be weak, moderate or high) for each time period. This methodology allowed the identification of unusual activity period (that could correspond to alerts). The age and gender of web users were determined with the detection of regular expressions within the posts. A drug-intake algorithm based on a random forest method was applied on the corpus of posts in order to identify whether each drug identified in the post had been taken by the web user. The three next algorithms of misuse detection were applied on posts where the drug intake was detected. The overdose results were obtained with the detection of words from overdose lexical field and the digits in the posts were compared to the daily dose recommended. The identification of the reason of paracetamol use was performed by the detection of medical concepts discussed through MedDRA which was enriched with web user vocabulary. The unapproved indications were obtained by comparing the medical concepts (MedDRA) detected as the indication found in the posts and the approved indications for paracetamol products(described in the drug summary**). Dependence was studied by the identification in the posts of dependence lexical field.

RESULTS

The number of posts extracted with the key words of only paracetamol containing products is 33 196 posts and published from 2003 and 2019 (March) on 18 social media sources (*). 17 070 web users dealt with paracetamol in their messages. The proportion of web users is 70% of women, 12% of men and 18% of unknown gender (FIGURE 1). The mean age of web users is 39 years old ± 20. The variable activity seen in the post volume over the study period shows a peak in post-volume activity between 2009 and 2012 (FIGURE 2). This spike may correspond with the withdrawal of dextropropoxyphene-containing medicines from the market in France. The number of posts tended to be higher during the winter months and lower during the summer and can be due to the increase of colds during this period. Regarding the misuse cases, the algorithm of unapproved indication detection identified 190 posts. The unapproved indications the most represented in the posts are Fatigue and Anxiety (0.28% of posts dealing with drug intake and quoting at least one medical concept) and Suicide (0.18%) as shown on the figure 3. The dependence of paracetamol was detected in 171 posts (1.33% of the posts detected with drug intake) and provided by 136 users (0.79% of users). Most of the posts detected with paracetamol dependence, deal with long term paracetamol treatment which induces the dependence and the associated risks. Some examples of posts of dependence are presented on figure 4. Reference to paracetamol overdose was made by 245 web users (which represents 1.43% of total web users) across 291 posts (2.26% of total corpus). The chart on the figure 5 shows the percentage of overdose above daily dose detected in the posts and the table introduces the drugs quoted for paracetamol overdose. The range of overdose described in the posts varies between 4.8 grams and 60 grams of paracetamol taken by web users. They describe an overdose for different reasons such as maximum daily dose unknown by the patients or intentionally substance abuse for recreational use or suicide attempt.

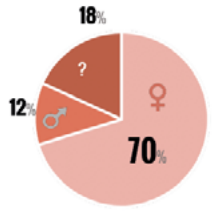


Figure 1 : Proportion of web users by genders

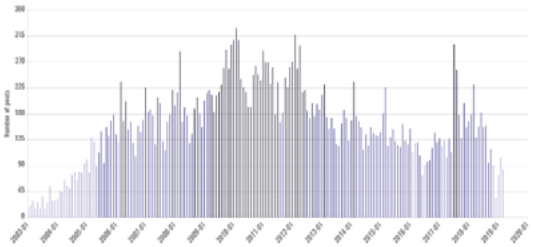


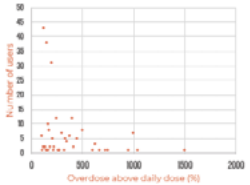
Figure 2 : Monitoring posts volume

Figure 3 : Unapproved indications

«Last Saturday,I stopped all antalgics (codein,tramadol,paracetamol) for limiting the addiction and thus, to get myself off drugs.»

«Finally,I had paracetamol poisoning. Since,I am scared to take it again but the addiction takes over. »

Figure 4 : Examples of posts detected with dependence



DRUGS	NUMBER OF USERS
DOLIPRANE	88
PARACETAMOL	75
DAFALGAN	61
EFFERALGAN	20
PANADOL	3
ACETAMINOPHENE	1
CLARADOL	1
PERFALGAN	1
TOTAL	250

«Hello, last night I swallowed 18g of Doliprane (I weight around 55kg and I'm a teenager) [...] I have a slight abdominal cramp but nothing worst. What should I do ? What are the next symptoms ? Is it serious ?»

« Hello sunday I took at least 12g of Dafalgan, I can't stand I feel totally depressed. I know I am taking a risk but which one ? Thank you for your help »

Figure 5 : Paracetamol overdose

CONCLUSION AND PERSPECTIVES

This retrospective study permitted to highlight the importance to consider social media as a source for detecting early potential signals from real life data. Real communities are constituted, patients mostly sharing their experience, seeking for support or describing drug misuse such as the use of paracetamol for reducing fatigue or anxiety (both 0.28% of the messages). Web users share also information about their drug addiction (1.33% of posts) and their overdose experience (291 posts).Thanks to these analyses, we will be able to identify the drug consumer behavior and the misuses as soon as possible.
These analyses will be applied prospectively in order to identify the drug consumer behavior and the misuses as soon as possible and thus, to set up mitigation plan and to monitor risks.

REFERENCES

1. Kürzinger ML, Schück S, Texier N, et al. Web-based signal detection using medical forums: data in France. J Med Internet Res (2018)
2. Van Stekelenborg, J., Ellenius, J., Maskell, S. et al. Recommendations for the Use of Social Media in Pharmacovigilance: Lessons from IMI WEB-RADR Drug Safety (2019)
3. P. Foulquié, P. Voillot, C. Faviez, A. Mebarki, X. Chen, N. Texier, S. Schück, Étude de l'usage du méthylphénidate sur les réseaux sociaux, Revue d'Épidémiologie et de Santé Publique, Volume 66, Supplement 4, (2018)

ISPOR
Improving healthcare decisions

2020

Studying fake news on Twitter during the COVID-19 pandemia in France

Gedik Anais; Foulquié Pierre; Renner Simon; Mebarki Adel; Texier Nathalie; Schück Stéphane

Studying fake news on Twitter during the COVID19 pandemia in France

Gedik A[1]; Foulquié P[1]; Renner S[1]; Mebarki A[1]; Texier N[1]; Schück S[1]
[1]Kap Code, Paris, France

INTRODUCTION

From the beginning of the global Covid-19 pandemic, misinformation have risen and have been spread on different social media platforms such as Facebook or Twitter [1]. Therefore, the latter has reported an 8% increase of its daily users on a global scale between the end of 2019 and the beginning of 2020*. This epidemic of misinformation is a global issue and disrupts public health *"We're not just fighting an epidemic; we're fighting an infodemic"*, said WHO Director-General Tedros Adhanom Ghebreyesus at the Munich Security Conference on February 15.

Social networks, especially Twitter, are a well-known source of real-world and real-time data [2]. The nature of shared information on these platforms can be of all kinds, true or false, and the accounts sharing it are also diverse, belonging to official sources and lambda citizens.

Covid-19 has been an important topic covered on Twitter that has given French people the chance to share their opinions on various matters related to it. Even though this social network has helped to create links between the users, it led to a propagation and a spread, consciously or not, of misinformation, commonly known as fake news.

The goal of our study was to identify and characterize fake news shared on Twitter related to Covid-19 pandemic.

METHODOLOGY

The dataset of this study is constituted of active French accounts during the pandemic. Through the Twitter API, we were able to extract French tweets related to Covid-19 between March and June 2019 using popular hashtags and key words such as *"#coronavirusfrance"* or *"#confinement"*. The dataset was cleaned to keep only tweets that have been retweeted and/or quoted at least once, retweets and quotes.

A specific list of word combinations (unigram or n-grams) enabled the identification of tweets reporting false information. Elements of this list were gathered from words, expressions and tweet contents found in fact-checking resources released by media associations such as *Les décodeurs du monde* or *AFP***. A primary filter was applied using this list, however while the detection of some fake news being biased, especially those concerning miracle cures or treatments such as "Gargle can help fight the virus", we applied a second filter based on misinformation lexical fields ("intox", "conspiracy", etc.). In other words, a tweet spreading a fake news will absolutely contain at least one element of our two lists mentioned above.

Once the tweets relating to fake news were identified, fake news were grouped manually into clusters depending on the subject covered in the tweets. With this step, an analysis of tweets contents was performed. These clusters also made it possible to determine twitter accounts involved in the transmission of each fake news. Every twitter account was weighted on its popularity and retweeting rate. To detect users' communities in each fake news group the Louvain clustering algorithm [3] was used.

For each fake news, a propagation model has been built in order to observe and analyze the exchanges between the different user accounts and detected communities. Retweets and quotes between users were represented by a network graph where each node represented a user and each edge an interaction (retweets or quote).

In this study, the analysis of the propagation network was done on the biggest group of fake news. Finally, a categorization of the different clusters of users was conducted by analyzing the nature of their exchanges as well as the tweet contents (articles, hashtags, shared URLs, videos...).

RESULTS

2.5 million tweets related to Covid-19 were extracted and analyzed, of which twenty fake news were deducted. These latter were identified and clustered into 5 groups (figure 1) thanks to the association rules of lexical fields such as "intox/Buzyn/chloroquine" or "creation/virus/labo ".

The first group, corresponding to the biggest proportion of fake news (39%) (figure 1), refers to the former French Minister of Health, Agnès Buzyn, and her husband, the former director of the National Institute of Health and Medical Research, Yves Lévy the decision makers of not to generalize the massive prescription of hydroxychloroquine as a way of preventing serious cases and deaths. Tweets of this group accuse them of sabotaging French infectiologist Didier Raoult's work on hydroxychloroquine with suspicion of conflicts of interest. Analyzing tweet contents, especially shared URLs, we found that more than half of the tweets deny this fake news by sharing articles from *Les décodeurs du monde****.

The second group identified (21%) (figure 1) mentions the creation of the virus in laboratories. This group is composed by several subgroups covering the potential origins of the virus. More than half of the tweets mention the creation of the virus within the laboratory P4 in the Institute of virology in Wuhan-China, built with the help of France, particularly Yves Lévy. Then comes that the Pasteur Institute would have launched the virus and would own a patent, and finally that of a Harvard researcher, Charles Lieber, who allegedly sold the virus to China.

The category "Other Fake News» (figure 1) corresponds to fake news represented respectively with a proportion of less than 5%, such as the fact that the heat can eradicate the virus, with the heat waves of summer or by drinking hot beverages (figure 3).

Focusing on the first group of fake news, more than 150 users' communities were found using the Louvain algorithm. Only accounts retweeted or quoted more than 5 times in each community were kept to build the network graph.

Analyzing the propagation network of fake news and the tweet contents enabled the categorization of accounts and different communities. The transmission network (Figure 2) reveals that the two protagonists of this fake news, Agnès Buzyn and Didier Raoult, hold the most retweeted accounts surrounded by media organizations accounts ("lemondefr", "decodeurs"). These latter interact with lambda citizens' accounts and would be part of the community denying the fake news relying on the fact checking articles URLs shared on their tweets. Many users' clusters are built around two central accounts belonging either to a journalist or a media source such as BFM. Small communities are on the periphery and seem to have no interaction with the media sphere. We also detected small communities politically involved in spreading the fake news coming from certain political parties and politically engaged users. These small communities, after analyzing the content of their tweets are separated into two categories. Those that are at the center of the network like the *Pcroquet* accounts, supporting the Professor Raoult, posting hashtags and mentions related to Didier Raoult, and mainly made of people that share the same nationalist ideologies analyzing its account description. And those that are on the periphery such as the community composed by the account *UPR_Rhone*.

If the typology of each cluster's twitter accounts is quite heterogeneous in terms of influence (number of post retweets, number of followers, etc.), it is homogeneous in terms of the tweet's content (hashtags, shared URLs, etc.)

Figure 1 : Proportion of fake news groups detected

Figure 2: Propagation network of the first group of fake news

Potential involment of the Buzyn/Lévy couple in the non prescription of chloroquine :

"Conflict of interest: this decision by Agnès Buzyn, who did her husband's business Agnès Buzyn made a decision in 2017 that suited her husband Yves Lévy, director of Inserm, which fuelled suspicions of a conflict of interest"

Other fake news category :
The heat can eradicate the virus

"Not many cases in Africa without containment, will the virus melt like snow in the sun with the great heat? I'm starting to believe it.. CoronavirusFrance "

Figure 3 : Example of fake news tweets

CONCLUSION AND PERSPECTIVES

Analyzing the proportion and model the propagation network of fake news relayed on Twitter can help fight misinformation. The study of networks has highlighted different clusters and accounts spreading fake news, with various degrees of influence.

It is then possible to quantify the presence of misinformation on social networks and its impact. These methods also allow to reveal the form and the nature of the fake news posts by analyzing their contents. Thus, we could determine the profile of accounts spreading fake news (bots...) and all the rhetorical elements used.

This kind of algorithm, which would prospectively detect fake news could help could help health authorities fight against them.

REFERENCES

[1] Brennen S., Simon F.M., Howard P., Nielsen R. Types, Sources, and Claims of COVID-19 Misinformation. Reuters Institute for the Study of Journalism, Oxford, UK. 2020.

[2] Ghosh P., Schwartz G., Narouze S. Twitter as a powerful tool for communication between pain physicians during COVID-19 pandemic. Reg. Anesth. Pain Med. .2020. doi: 10.1136/rapm-2020-101530.

[3] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics, vol. 2008, no. 10, Article ID P10008, 2008.

ISPOR
Improving healthcare decisions

2020

Detection of Quality of Life Impact from Health-related Messages in Social Media

Tom Marty, Mickaïl Khadar, Simon Renner, Pierre Foulquié, Pamela Voillot, Adel Mebarki, Nathalie Texier and Stéphane Schück

Detection of Quality of Life Impact from Health-related Messages in Social Media

Tom Marty[1], Mickaïl Khadar[1], Simon Renner[1], Pierre Foulquié[1], Pamela Voillot[1], Adel Mebarki[1], Nathalie Texier[1] and Stéphane Schück[1]
[1]Kap Code, 28 rue d'Enghien, 75010 Paris ; Kap Code ; France

INTRODUCTION

Every day, people share health-related messages and opinions online. Undearstanding and analyzing these insights could be useful for medical applications. Quality of life, medically defined as "An individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by the person's physical health, psychological state, personal beliefs, social relationships and their relationship to salient features of their environment". Patients and caregivers make extensive use of social media where they seek information and support and share their experience, difficulties and expectations. Mining social media to understand the burden of pathologies and treatments is a new approach to better understand patients' concerns and unmet needs. This study proposes a NLP algorithm able to detect health-related quality of life (HRQoL) impacts on social media posts.

MATERIAL & METHODS

Messages relating a patient experience with a drug or disease were collected on 19 social medias using a Web crawler, respectively 1000 and 400 messages, randomly taken from a 20 000 health-related messages batch. Data were harvested with an internal developed tool, Detec't©, based on a name entity recognition module using a medical lexicon.

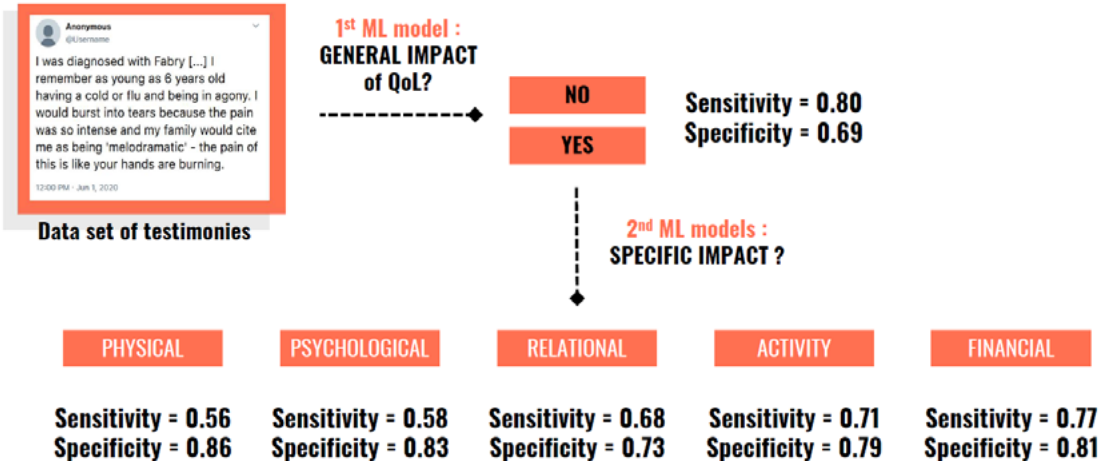
Two reviewers manually annotated messages following guidelines based on EQ-5D and SF-36 QoL measurements. Focus was put on physical, psychic, activity, relational and financial dimensions of HRQoL. The annotation homogeneity was measured with Cohen's Kappa coefficient. For each HRQoL dimension, several domains and subdomains were created and annotated : 5 general dimensions (Physical, Psychic health, Activity, Relational, Financial) and 9 subdomains for physical impact; 27 subdomains for psychic health; 32 subdomains for activity-related impact; 9 subdomains for relational impact; and 6 domains for financial impact. This methodology was defined in order to better characterize a HRQoL impact in a patient message: for example one subdomain of the "Physical" dimension is "Physical pain", and "Depression syndrome" is a subdomain of "Psychic health". A first model detected impact presence. Then each impact was identified by a specific model (figure 1). Through annotation, patients' expressions of impact were collected in order to composed specific lexical fields, then used to generate features. Other features were based on the message content, such as expressed sentiment, grammar and conjugation. The rational of this process was to be able to adapt to the many expressions of patients. Indeed, psychic impacts and physical impacts are different and so are their expressions. Hence, having specific models is a way to minimize interpretation bias Analysis corpus with annotations allowed obtaining a Gold Standard.

RESULTS

Training set represented 1000 posts related to pathologies and 400 to treatments. Extreme gradient boosting was the chosen model for both impact detection and specific dimension identification. Models were trained using cross-validation and hyperparameter optimization. Over-sampling was used to augment infrequent dimensions. This allowed us to detect a general impact with a sensitivity of 0.8 and a specificity of 0.7, physical (0.56 ; 0.857), psychic (0.58 ; 0.828), activity (0.71 ; 0.79), relational (0.675 ; 0.73), financial (0.77 ; 0.814) dimension. (figure 1).

CONCLUSION

We provided evidence that social medical listening can be used to assess the impact and burden of one or more diseases and treatments on patients' HRQoL. We developed an algorithm, based on medically validated questionnaires, able to identify impacts of HRQoL in online patient's messages. These findings can provide public health experts, healthcare professionals, pharmaceutical companies and Health authorities with patient-generated information on their experiences with treatments, diseases and needs for appropriate medical care in timely and real-life conditions. Social media studies could be a complementary source of real world evidence, to understand how diseases and therapies represent a burden to patients and their caregivers/relatives..



AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2020

Application de tracking StopCovid : identification des thématiques de discussions et des opinions par l'analyse des réseaux sociaux

Paméla Voillot, Pierre Arwidson, Pierre Foulquié, Anne-Juliette Serry, Adel Mebarki, Nathalie Texier, Stéphane Schück

Kap Code

APPLICATION DE TRACKING **STOPCOVID** : IDENTIFICATION DES THÉMATIQUES DE DISCUSSIONS ET DES OPINIONS PAR L'ANALYSE DES RÉSEAUX SOCIAUX

Santé publique France

Paméla Voillot[1], Pierre Arwidson[2], Pierre Foulquié[1], Anne-Juliette Serry[2], Adel Mebarki[1], Nathalie Texier[1], Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France
[2]Santé publique France, 12, rue du Val d'Osne, 94 415 Saint-Maurice, France

INTRODUCTION

L'application StopCovid, mise en ligne le 2 juin 2020, permet de prévenir les individus qui ont été proches d'une personne testée positive au Covid-19 pendant au moins 15 minutes et à moins d'un mètre [1]. Celle-ci est téléchargeable sur la base du volontariat. Le fonctionnement de l'application a été approuvée par la CNIL et le code source est public. Malgré ces précautions, cette application divise la population. **L'application a été téléchargée plus de 2,3 millions de fois à la mi-août. À ce jour, 1 169 usagers se sont déclarés positifs sur l'application permettant de repérer 72 contacts à risque.**

Les réseaux sociaux sont un espace privilégié pour plus de 32 millions d'internautes. Ces plateformes occupent une place grandissante comme lieu d'expression où les internautes échangent sur leur santé, leurs préoccupations et où la parole est libérée [2].

L'identification des opinions et des thématiques de discussions des internautes autour de l'application durant le confinement pourrait aider à **mieux appréhender le comportement de la population envers cet outil numérique en post-confinement.**

MATÉRIEL ET MÉTHODES

L'étude a été menée via l'outil Detec't®. Les messages associés à l'**application StopCovid** (dataset StopCovid) ont été extraits de **Twitter entre le 13/04/2020 et le 06/05/2020.**

La méthode d'analyse de sentiment de Microsoft Azure Cognitive Services a été utilisée pour identifier la perception générale des messages vis-à-vis de StopCovid (neutre, négative ou positive).

Par la suite et afin de caractériser le contenu de ces messages Twitter, l'algorithme Biterm Topic Model (BTM) a été appliqué sur l'ensemble du corpus afin de permettre une identification automatique des différentes thématiques de discussions abordées [3].

RÉSULTATS

Le dataset StopCovid contenait 7 110 tweets rédigés par 4 248 internautes différents avec une majorité de jeunes postants (moyenne = 27 ans) (figure 1).

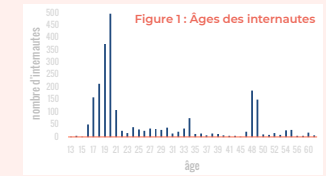


Figure 1 : Âges des internautes

Nous avons pu observer un accroissement du nombre de messages rédigés à ce sujet fin avril, concomitant avec l'avis de la commission supérieure du numérique du 24 avril 2020 sur les conditions de mise en œuvre de l'application StopCovid (figure 2).

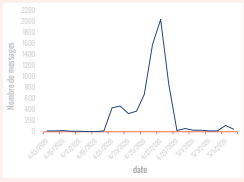


Figure 2 : Dates de publication des tweets concernant l'application StopCovid

L'analyse de la perception générale des messages a révélé une majorité de neutralité (53,29% des messages) (figure 3).

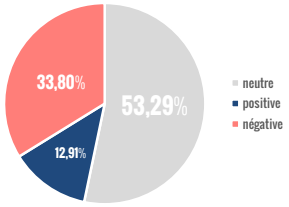


Figure 3 : Perception des messages des internautes concernant l'application StopCovid (%)

Ils se rapportent à des partages d'informations générales et diverses telles que les participants au projet ou encore l'objectif d'une telle application. Cette neutralité est suivie d'un grand nombre de messages avec une perception négative (33,80%) remettant principalement en question l'intérêt même du dispositif. Toutefois, 12,9% des messages avaient une perception positive, mentionnant l'utilité sanitaire du projet.

Par la suite, l'application de l'algorithme BTM a permis de mettre en avant 9 grandes thématiques de discussions dont 8 d'intérêt (figure 4).

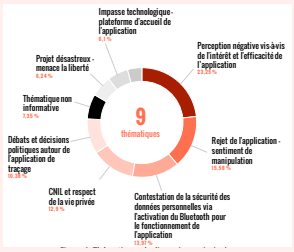


Figure 4 : Thématiques de discussions principales

La principale thématique se réfère aux opinions négatives des internautes vis-à-vis de l'intérêt et de l'efficacité de l'application (23,25% des messages) (figure 5).

« Cessez ce caprice de StopCovid ! Vraiment ce gadget n'aura aucune efficacité...Comptez vous procurer un smartphone à chaque citoyen ?! Arrêtez les frais ! Dépensez utilement et intelligemment > Masques Tests... »

Figure 5 : Ex. de tweet concernant la thématique « Perception négative vis-à-vis de l'intérêt et l'efficacité de l'application »

La seconde thématique la plus représentée dans le corpus se réfère au sentiment de manipulation par culpabilisation pour imposer l'utilisation de StopCovid (15,58%) (figure 6).

« Je refuse l'injonction culpabilisante que nos politiques, technos et autre relais médiatiques ne vont pas tarder à nous imposer sur StopCovid. Ce sera sans moi. »

Figure 6 : Ex. de tweet concernant la thématique « Rejet de l'application – sentiment de manipulation »

La troisième thématique met en avant l'insécurité des données personnelles via l'utilisation permanente du Bluetooth pour le bon fonctionnement de l'application (13,97%) (figure 7).

«Tiens, une nouvelle faille de sécurité dans une implémentation Bluetooth ! Nous maintenons : imposer Bluetooth risque de mettre en danger la vie numérique les utilisateurs de StopCovid!»

Figure 7 : Ex. de tweet concernant la thématique « Contestation de la sécurité des données personnelles via l'activation du Bluetooth pour le fonctionnement de l'application »

Les thématiques suivantes s'en remettent au respect de la vie privée (12,9%), aux débats politiques (10,38%), aux menaces de liberté (6,24%), puis à l'impasse technologique concernant la plateforme d'accueil de l'application (6,1%) et à la mobilisation générale pour son développement (4,23%).

CONCLUSION

Cette étude a permis d'observer l'existence d'une communauté d'internautes en débat et en échange d'information concernant l'application de tracking StopCovid.

L'analyse du contenu des messages sur Twitter a principalement permis de mettre en avant des perceptions et opinions négatives vis-à-vis du dispositif. Cependant, certaines limites inhérentes à toutes les études sur les médias sociaux demeurent. Aucun média ne s'est avéré représentatif de la population générale [4]; par exemple, seuls 34 % des internautes français utilisent activement Twitter.

Malgré cela, nous pouvons observer que l'efficacité, l'insécurité des données personnelles et l'atteinte à la vie privée sont au cœur des échanges sur les réseaux sociaux à ce sujet.

Ce type d'étude infodémiologique pourrait servir aux stratégies de Santé publique afin de comprendre le ressenti des utilisateurs français et d'améliorer leur information.

RÉFÉRENCES

[1] Info Coronavirus COVID-19 - STOPCOVID | Gouvernement, https://www.gouvernement.fr/info-coronavirus/stopcovid , 31 août 2020.

[2] Fox S, Duggan M. PewInternet. 2013 Jan 15. Health Online 2013 URL: http://www.pewinternet.org/2013/01/15/health-online-2013/[accessed 2018-02-27] [WebCite Cache ID 6xXiCgyLK].

[3] Blei DM, Lafferty JD, 2009. "Topic models," in Text Mining: Classification, Clustering, and 670 Applications, Vol. 10, eds A. N. Srivastava and M. Sahami (Boca Raton, FL: Chapman and 671 Hall/CRC), 34.

[4] Hootsuite & We Are Social. Digital 2020: Global Digital Overview. DataReportal – Global Digital Insights https://datareportal.com/reports/digital-2020-global-digital-overview

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2020

Réseaux sociaux et coronavirus : caractérisation des contenus échangés en France sur Twitter pendant la crise sanitaire du Covid-19

Pierre Foulquié, Léa Châteauneuf, Paméla Voillot, Simon Renner, Adel Mebarki, Nathalie Texier et Stéphane Schück

Kap Code

RÉSEAUX SOCIAUX ET CORONAVIRUS : CARACTÉRISATION DES CONTENUS ÉCHANGÉS EN FRANCE SUR TWITTER PENDANT LA CRISE SANITAIRE DU COVID-19

Pierre Foulquié[1], Léa Châteauneuf[1], Paméla Voillot[1], Simon Renner[1], Adel Mebarki[1], Nathalie Texier[1] et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

INTRODUCTION

Suite aux premiers cas de coronavirus enregistrés sur son territoire en février 2020, **la France a pris différentes mesures sanitaires pour répondre à la crise mondiale**. Ces mesures, dont le confinement mis en place le 17 mars, ont impacté le quotidien des français de façon inédite. Les réseaux sociaux, notamment Twitter, source reconnue de données de vie réelle, ont permis aux français de communiquer et d'échanger des informations pendant cette crise. **Cette étude infodémiologique se propose d'étudier la nature et l'évolution des contenus échangés sur twitter à partir de plusieurs méthodes de fouille de texte.**

MATÉRIEL ET MÉTHODES

Un corpus de tweets français liés au Covid-19 a été constitué grâce à l'API Twitter. Les tweets extraits contenaient des mots-clés ou des hashtags évocateurs du coronavirus (*#coronavirusfrance*, *#covid19fr* etc.) et du confinement (*#restezchezvous*). Trois périodes d'intérêt ont été définies : du 10 au 31 mars, le mois d'avril et le mois de mai.

Le contenu des tweets est caractérisé par trois méthodes et les résultats de chaque période sont comparés.
Un modèle de sujet adapté aux textes courts (*biterm topic model* [1]) a été utilisé pour déterminer les thématiques des discussions. Ce modèle classe les messages selon les thématiques qu'ils abordent. Celles-ci se présentent sous la forme de listes de mots apparaissant ensemble dans les messages. Une classification manuelle des adresses web partagées a été opérée, basée sur la nature du site internet (presse, site gouvernemental etc.). Les proportions de ces regroupements dans le total des contenus web partagées est étudiée, ainsi que leur évolution au cours des trois périodes d'intérêt, dans le but de distinguer les types de contenus échangés.

Une identification des terminologies du dictionnaire MedDRA (*Medical Dictionnary for Regulatory Activities*), enrichie de vocabulaire patient [2], a permis d'extraire le vocabulaire médical et les éventuels symptômes du coronavirus et du confinement. Les terminologies détectées sont ensuite regroupées par catégories médicales, grâce à la structure arborescente du dictionnaire MedDRA. Par exemple, « expectorer » et « tousser » sont regrouper dans la catégorie « Toux ».

RÉSULTATS

Le corpus constitué via l'**API Twitter** contient **2,5 millions de tweets** : 933 481 en mars, 833 108 en avril et 774 778 en mai. La figure 1 montre l'évolution temporelle hebdomadaire nombre de messages. Le pic à la gauche du graphique correspond à la première semaine de confinement, qui a généré beaucoup de contenu. Le tableau 1 présente les cinq thèmes les plus abordés, chaque mois. Au mois de mars, ils correspondent à des inquiétudes et difficultés liées au confinement (37% des tweets), ainsi que des commentaires de la gestion de la crise par l'État (27%). Ces thèmes persistent en avril. Ils sont accompagnés par des thématiques plus positives décrivant des solidarités (4,68%) et des idées d'occupations (8,80%). Au mois de mai, les conditions du déconfinement sont largement discutées, avec par exemple la réouverture des écoles (18,09%) et le port du masque (22,55%).

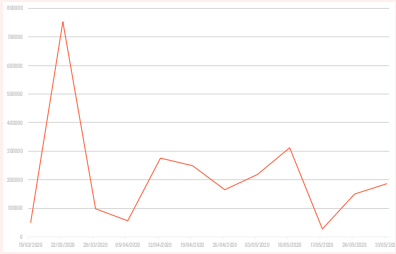


Figure 1 – Évolution hebdomadaire du volume de tweets

	Mars		Avril		Mai	
Thèmes	Thème	P	Thème	P	Thème	P
	Difficultés liées au confinement	37,48%	Difficultés liées au confinement	29,43%	Port du masque quotidien	22,55%
	La crise sanitaire en France	27,00%	Non-respect des mesures de santé publique	15,64%	Écoles	18,09%
	Initiatives suite au confinement	13,52%	La gestion de la crise par L'Etat	10,54%	L'utilité du masque	10,44%
	Mesures de santé pub. FR	9,16%	Positiver pdt le confinement	8,80%	Lutte contre l'épidémie	9,26%
	Le personnel soignant	6,37%	Relations sociales pdt le confinement	6,37%	Retour en entreprise	5,37%
URL	Type	P	Type	P	Type	P
	Hébergeurs de vidéos	20,90%	Hébergeurs de vidéos	19,25%	Presse quotidienne nat.	23,92%
	Réseaux sociaux	13,76%	Presse quotidienne nat.	15,78%	Télévision FR	14,49%
	Presse quotidienne rég.	12,59%	Presse quotidienne rég.	10,03%	Presse quotidienne rég.	11,94%
	Presse quotidienne nat.	11,31%	Télévision FR	6,92%	Hébergeurs de vidéos	8,42%
	Site Service Public	8,50%	Réseaux sociaux	5,83%	Station de radio	6,79%

Tableau 1
Thèmes de discussion abordés (haut) et type de sites partagés (bas)

Tableau 2
Catégories de terminologies médicales exprimées

	Mars		Avril		Mai	
Terminologies médicales	Concept	P	Concept	P	Concept	P
	Mort	9,33%	Mort	11,06%	Mort	10,26%
	Ennui	2,92%	Fatigue	2,25%	Anxiété	1,73%
	Fatigue	1,90%	Ennui	2,11%	Fatigue	1,69%
	Humeur dépressive	1,58%	Humeur dépressive	1,77%	Dépendance	1,37%
	Grippe	1,39%	Usage abusif de substances	1,29%	Étouffement	1,29%

Les sites de presse et d'informations constituaient un tiers des adresses web échangées en mars et en avril, et plus de la moitié en mai. Youtube, avec des vidéos d'ordre informatif, humoristique ou créative, est le site le plus partagé en mars (20% des liens). Cette proportion diminue à chaque période. Une tendance similaire est constatée pour les réseaux sociaux (autres que Twitter). Les sites de services publics sont présents au mois de mars, notamment du fait des décisions de confinement, du partage des gestes barrières et d'attestation de sorties. Cette proportion diminue dès avril. À chaque période, des terminologies médicales sont identifiées dans environ 10% des tweets. Les catégories les plus représentées sont regroupées dans le tableau 2 Les terminologies liées à la mort proviennent des bilans journaliers sur l'évolution de l'épidémie et relayées par les internautes. Si peu de symptômes du coronavirus sont observés, des termes évocateurs de troubles psychologiques sont identifiés : **Anxiété, Ennui, et Dépression**. Des comportements à risque sont également identifiables par les catégories Dépendance et Usage abusif de substances. Ces résultats suggèrent un effet du climat de crise et du confinement sur les individus.

CONCLUSION

L'étude du contenu échangé sur Twitter de mars à mai 2020 permet de qualifier l'usage dont en ont fait les français pendant la crise sanitaire. Les utilisateurs partagent leurs inquiétudes, leurs opinions, et partagent des supports d'information. Ils partagent également des troubles, notamment psychologiques et imputables au confinement. Chaque période analysée contient une ou plusieurs décisions de santé publique majeures (confinement, prolongation du confinement, déconfinement) ayant influencé le contenu des échanges.
Ces résultats souffrent de limites inhérentes à toutes études réalisées à partir des réseaux sociaux.
La représentativité d'abord, puisque seuls 34% des internautes français utilisent activement Twitter [3].
Dans un contexte d'épidémie, les réseaux sociaux sont néanmoins des indicateurs du niveau d'information de la population. Les analyser permet donc d'identifier et de répondre à ses attentes et inquiétudes et d'orienter les politiques publiques sanitaires.

RÉFÉRENCES

[1] Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web (pp. 1445-1456).

[2] Abdellaoui, R., Schück, S., Texier, N., & Burgun, A. (2017). Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help?. JMIR public health and surveillance, 3(2), e36

[3] Hootsuite & We Are Social. Digital 2020: Global Digital Overview. DataReportal – Global Digital Insights https://datareportal.com/reports/digital-2020-global-digital-overview

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2020

Segmentation des internautes issus des réseaux sociaux au cœur de la controverse sur la vaccination

Paméla Voillot, Avesta Roustamal, Anaïs Gedik, Pierre Foulquié, Simon Renner, Adel Mebarki et Stéphane Schück

Kap Code

SEGMENTATION DES INTERNAUTES ISSUS DES RÉSEAUX SOCIAUX AU CŒUR DE LA CONTROVERSE SUR LA VACCINATION

Paméla Voillot[1], Avesta Roustamal[1], Anaïs Gedik[1], Pierre Foulquié[1], Simon Renner[1], Adel Mebarki[1] et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

Evanex

INTRODUCTION

L'élargissement de l'obligation vaccinale dès 2018 a permis d'améliorer le taux de couverture vaccinale en France [1]. En effet, l'éradication de certaines maladies comme la rougeole nécessite un niveau de couverture vaccinale de 95% [2]. **L'obtention d'une couverture vaccinale suffisante est un objectif majeur de santé publique.** Néanmoins, une part de la population reste toujours réticente à cette mesure. **L'hésitation vaccinale persistante a été classée par l'OMS comme l'une des plus grandes menaces sur la santé.** En complémentarité avec les bases de données publiques, les réseaux sociaux, avec plus de 32 millions d'internautes, sont une source d'échanges et de partage d'information où la parole est libérée [3] et permet d'exprimer ses doutes et interrogations sur la vaccination. La caractérisation de ces internautes pourrait aider à mieux comprendre l'attitude de la population envers la vaccination. L'objectif de cette étude était d'identifier la typologie des internautes qui s'expriment sur les réseaux sociaux au sujet de la vaccination.

MATÉRIEL ET MÉTHODES

Dans le cadre de la mise en place d'un observatoire sur la vaccination appelé **Evanex**®, les messages d'internautes issus des réseaux sociaux ont été extraits à partir d'un *webcrawler* développé par la société Kap Code. L'analyse des messages a été réalisée au moyen de divers outils de *Machine Learning*. La méthode de classification hiérarchique ascendante (CAH) a été utilisée pour former des catégories d'internautes [4]. Celle-ci se base sur plusieurs variables telles que l'activité de l'internaute sur les réseaux sociaux (nombre de messages, fréquence de publication, etc.), les sujets abordés dans ses messages (méthode de *topic modeling*) et les concepts médicaux exprimés (détection de termes médicaux à partir du dictionnaire médical MedDRA, enrichi avec du vocabulaire d'internaute). Les individus qui présentaient des centres de gravité similaires après le calcul des variables, ont été regroupé ensemble par la méthode de Ward. Les groupes obtenus sont interprétés à partir de leurs centres de gravité.

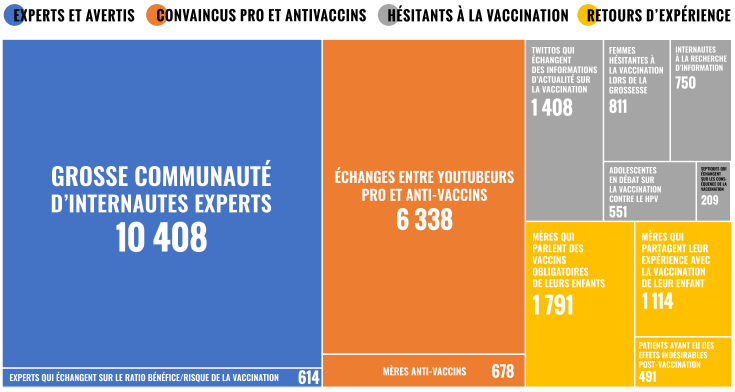


Figure 1 – Segmentation des internautes

RÉSULTATS

L'analyse a été réalisée sur **88 894** messages provenant de **37 forums français** (Doctissimo, Au Féminin...) et a mis en évidence une diversité d'internautes actifs sur des discussions autour de la vaccination (**25163 internautes**). La segmentation a permis d'identifier **12 catégories d'internautes regroupées en 4 grands groupes d'utilisateurs** (figure 1). Le groupe rassemblant le plus d'individus (n=11 022; 44%) est composé d'internautes « experts » et averties qui échangent sur les risques et bénéfices de la vaccination. Ils échangent également des informations scientifiques factuelles. Ces internautes constituent une communauté très active et aguerrie avec une moyenne de rédaction de 5 messages et 14 concepts médicaux par post. Le deuxième groupe est composé des convaincus pro et anti-vaccins (n=7016, 28%) qui soutiennent une certitude dogmatique. On y retrouve des Youtubeurs pro ou anti-vaccins ainsi que des mères anti-vaccins. Ce groupe est définit par un nombre important d'expression de termes médicaux tels que l'autisme, les effets indésirables et l'aluminium. Le troisième groupe d'internautes est composé d'hésitants (n=3729, 15%). Ils s'expriment sur des forums spécifiques comme Forum Ados ou le Journal des Femmes. Les hésitants sont majoritairement des mères qui se questionnent sur les obligations vaccinales. Il est également composé d'adolescentes qui débâtent sur les risques de la vaccination anti-HPV, avec des termes médicaux exprimés comme « tumeur maligne », « sclérose en plaque » et « grossesse ». Le dernier groupe est représenté par des internautes qui décrivent un retour d'expérience (n=3396; 13%). Ce groupe est caractérisé par des individus qui partagent les effets indésirables post-vaccination vécus personnellement ou par un proche ainsi que des mères témoignant d'expériences de vaccination de leurs enfants.

CONCLUSION

Cette étude démontre que la vaccination est un sujet qui continue à diviser. Les internautes partagent leurs inquiétudes mais sont également à la recherche d'informations. Les réseaux sociaux sont une source à surveiller et à exploiter pour contribuer aux stratégies de Santé publique d'amélioration de la couverture vaccinale.

RÉFÉRENCES

[1] Qu'est-ce que la couverture vaccinale ?, Dossier thématique, https://www.santepubliquefrance.fr/determinants-de-sante/vaccination/articles/qu-est-ce-que-la-couverture-vaccinale, 29 Juin 2019

[2] Les programmes de vaccination sont de plus en plus confrontés aux hésitations de la population, Communiqué de Presse OMS, 18 Août 2015, https://www.who.int/fr/news-room/detail/18-08-2015-vaccine-hesitancy-a-growing-challenge-for-immunization-programmes

[3] Fox S, Duggan M. PewInternet. 2013 Jan 15. Health Online 2013 URL: http://www.pewinternet.org/2013/01/15/health-online-2013/[accessed 2018-02-27] [WebCite Cache ID 6xXiCgyLK].

[4] S. Schück et al, Que nous apportent les réseaux sociaux quant à la crise sanitaire du Levothyrox® d'août 2017 ?,Revue d'Épidémiologie et de Santé Publique, Volume 66, Supplement 4,2018,Page S225, ISSN 0398-7620

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2020

Analyse temporelle du volume et de la perception relatifs aux masques à partir de tweets français pendant la période de confinement

Pierre Foulquié, Ludovic Figarella, Léa Châteauneuf, Simon Renner, Adel Mebarki, Nathalie Texier et Stéphane Schück

Kap Code

ANALYSE TEMPORELLE DU VOLUME ET DE LA PERCEPTION RELATIFS AUX MASQUES À PARTIR DE TWEETS FRANÇAIS PENDANT LA PÉRIODE DE CONFINEMENT

Pierre Foulquié[1], Ludovic Figarella[1], Léa Châteauneuf[1], Simon Renner[1], Adel Mebarki[1], Nathalie Texier[1] et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

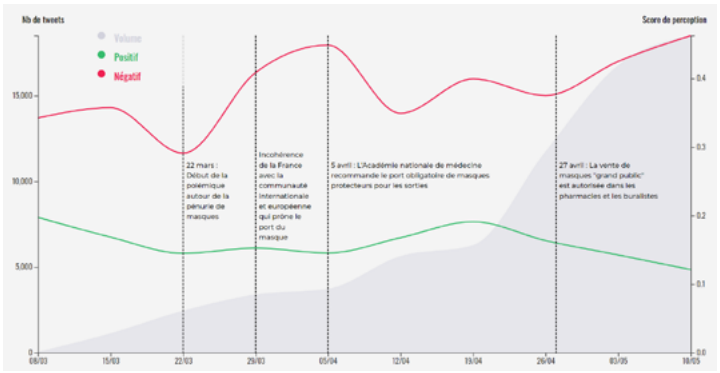
Detec't

INTRODUCTION

En 2020, en pleine pandémie de Covid-19, la gestion du matériel sanitaire, dont font partie les masques, a été au cœur de plusieurs polémiques : **efficacité du port du masque comme moyen de prévention, communication du gouvernement ou encore gestion du stock d'État.** De tels sujets, suscitant des divergences de discours, sont analysables à l'aide de réseaux sociaux comme Twitter. Ces canaux permettent à la population de réagir à l'actualité, d'y exprimer ses opinions et sa perception des sujets de santé publique. **Cette étude propose une analyse temporelle de l'apparition du sujet des masques et de l'évolution des opinions associées.**

MATÉRIEL ET MÉTHODES

Un corpus de tweets français liés au Covid-19 a été constitué grâce à l'**API Twitter**. Les tweets extraits contenaient des mots-clés ou des hashtags évocateurs du coronavirus (*#coronavirusfrance, #covid19fr, ...*) et du confinement (*#restezchezvous*), postés dans une période allant du **10 mars au 11 mai 2020**. Elle correspond à la semaine précédant le confinement jusqu'à la date du déconfinement en France.
Un second filtre textuel a ensuite été appliqué sur les tweets dans le but de conserver uniquement les messages mentionnant les mots "*masque*" ou "*masques*". Ceci permet d'obtenir un corpus de tweets liés au coronavirus et aux masques.
Pour chaque tweet, les proportions de sentiments positif ou négatif exprimées ont été estimées par la méthode d'analyse de sentiment de Microsoft Azure Cognitive Services [1]. **Trois séries temporelles bijournalières ont pu être construites**. La première série temporelle est constituée du volume cumulé de tweets postés au sujet des masques, toutes les douze heures. Elle correspond à l'évolution de l'importance de ce sujet parmi les discussions des internautes. Les deux autres séries temporelles sont constituées des proportions de messages avec un sentiment positif et négatif, toutes les douze heures. Ces deux courbes visent à créer un indicateur de la perception des utilisateurs concernant les sujets connexes aux masques. Elles ne sont pas des indicateurs de proportions d'utilisateurs pour ou contre le port du masque. En effet, un tweet peut exprimer une indignation (identifié comme un sentiment négatif par la méthode) concernant la non-obligation de porter le masque dans l'espace public.



Graphique 1 – Évolution du volume de tweets et de la perception des utilisateurs sur la question des masques.
Les pontillés verticaux indiquent des évènements susceptibles d'avoir impacté la perception de ce sujet par le public

RÉSULTATS

773 115 messages en français liés au coronavirus ont été extraits de Twitter entre le 10 mars au 11 mai. **La détection des termes liés aux masques a permis d'isoler 18 508 messages**, soit 2.4% du corpus total. Les dates importantes dans le débat sur le masque sont identifiables à partir des évolutions des trois séries temporelles (Graphique 1).
Au début de la crise sanitaire, en même temps que l'instauration du confinement, il est décidé que les masques seront en priorité pour le personnel soignant et le ministre de la Santé, Oliver Véran, annonce que la France dispose d'un stock assez important pour eux. Ceci est rapidement contesté et, à partir du 22 mars, des premiers débats émergent quant à la responsabilité de la pénurie. A compter de cette date, une hausse importante de messages négatifs sur le sujet des masques est observée, passant de 29% à 45% en deux semaines.
Ce pic de 45% a lieu la semaine du 5 avril, durant laquelle le port obligatoire d'un masque, même alternatifs (c'est-à-dire autre que FFP2), a été recommandé par l'Académie Nationale de Médecine.
Sur les trois dernières semaines de confinement, le volume de messages s'est accru avec 12 245 messages (deux tiers de notre corpus final). La proportion de nouveaux messages négatifs augmente tandis que le nombre de tweets positifs diminue, en passant de 16% à 12%. Ces tendances correspondent aux inquiétudes grandissantes des Français à l'approche du déconfinement, à l'obligation du port du masque est obligatoire dans les transports et commerces.

CONCLUSION

La question des masques a pris une place grandissante au cours de la période de confinement. Bien qu'en pénurie et jugé inutile au début de la crise sanitaire, le masque a progressivement été accepté par la majorité de la population comme essentiel pour lutter contre l'épidémie. L'analyse proposée dans cette étude suggère des conclusions similaires. Le volume de contenus échangés sur le sujet et les fluctuations de perceptions sont directement corrélés aux annonces et évènements relatifs à la problématique des masques.
Ces résultats souffrent de limites inhérentes à toutes études réalisées à partir des réseaux sociaux. La représentativité d'abord, puisque seuls 34% des internautes français utilisent activement Twitter [2]. Une deuxième limite est liée à l'imperfection de méthodes de fouilles de texte comme l'analyse de sentiment. Bien que validées scientifiquement, une telle méthode ne peut refléter l'opinion complète d'un individu sur une question complexe comme celle que nous traitons ici.
Ce travail démontre néanmoins si la possibilité de suivre l'impact d'une politique de santé publique et sa perception par une partie de la population en utilisant des données de vie réelle issues de témoignages sur Twitter.

RÉFÉRENCES

[1] https://azure.microsoft.com/fr-fr/services/cognitive-services/text-analytics/#features.

[2] Hootsuite & We Are Social. Digital 2020: Global Digital Overview. DataReportal – Global Digital Insights https://datareportal.com/reports/digital-2020-global-digital-overview.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2020

Analyse des réseaux sociaux pour identifier les motifs de l'hésitation vaccinale anti-HPV : une étude infodémiologique

Simon Renner, Tom Marty, Pamela Voillot, Pierre Foulquié, Adel Mebarki, et Stéphane Schück

Kap Code

ÉTUDE DES RÉSEAUX SOCIAUX POUR IDENTIFIER LES MOTIFS DE L'HÉSITATION VACCINALE ANTI-HPV : UNE ÉTUDE INFODÉMIOLOGIQUE

Simon Renner[1], Tom Marty[1], Pamela Voillot[1], Pierre Foulquié[1], Adel Mebarki[1], et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

Evanex

INTRODUCTION

La recommandation vaccinale contre l'infection à papillomavirus humain (HPV) est un **enjeu mondial de santé publique**. En 2019, la **couverture vaccinale de 21%** place la France en dernière position européenne. Parallèlement, les **réseaux sociaux** sont un espace privilégié pour plus de **32 millions d'internautes actifs**. Ces plateformes occupent une place grandissante comme **lieu d'expression ou les internautes échangent sur leur santé**, leurs préoccupations et souvent sur leur position vis-à-vis de la vaccination. Analyser et comprendre ces messages permettrait d'identifier les **leviers associés à une amélioration de la couverture vaccinale anti-HPV**.

MATÉRIEL ET MÉTHODES

L'étude a été menée via l'observatoire de la vaccination EVANEX®. Les messages associés à la vaccination anti-HPV et écrits en français (dataset HPV) ont été extraits à partir de **23 forums médicaux et de réseaux sociaux, entre 2006 et 2019**. Les termes d'extraction faisaient soit référence à un acte de vaccination contre le HPV « vaccin contre le papillomavirus » soit directement à un nom de produit (Gardasil® et Cervarix®). Une annotation manuelle de messages (présence ou non d'une hésitation) a permis la création d'un Gold standard visant à développer un algorithme de détection d'une hésitation vaccinale. Cet algorithme a ensuite été appliqué sur l'ensemble du dataset HPV. Un premier corpus d'analyse (pré-corpus d'hésitation vaccinale, Figure 1) a pu être constitué. Une annotation manuelle de l'ensemble de ce corpus d'intérêt a été effectuée afin d'identifier les différents facteurs d'hésitation (corpus d'hésitation vaccinale, Figure 1) et les regrouper par grandes thématiques d'hésitation.

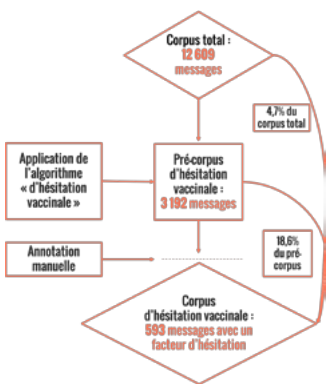


Figure 1 : Acquisition des données

RÉSULTATS

Le dataset HPV contenait **12 609 messages** uniques rédigés par 5 117 internautes différents. Un **pré-corpus de 3 192 messages propres à l'hésitation vaccinale** a été déterminé algorithmiquement. L'analyse manuelle de ces 3 192 messages a permis de caractériser **593 messages d'hésitation vaccinale** (18.6% du pré-corpus, 4.7% du dataset HPV) (Figure 1).

[...] sa fait a peu pré 1 an et demi que je ne suis plu vierge ma mère veut absolument que je fasse le vaccin HPV mais je ne ve pas lui dire que je ne suis plus vierge eske cela comprte un risque que je me fasse vacciner ?

Bonjour, j'ai 17 ans et j'ai déjà eu des rapports avec une seule personne, cette personne n'a eu des rapports qu'avec moi aussi. Je n est pas été vacciner contre le cancer du col de l'utérus et j'aimerais savoir si c'est toujours possible ?

Figure 2 : Exemples de messages extraits

Les messages (Figure 2) ont été regroupés en **3 principaux groupes**, selon les facteurs d'hésitation vaccinale exprimés. Le premier a pour motif l'influence de la vie sexuelle sur la vaccination (339 messages, 7.5% pré-corpus). Au sein de celui-ci, les adolescentes s'interrogent à la fois sur la nécessité de se faire vacciner mais aussi sur une éventuelle dangerosité du vaccin une fois sexuellement actif (n=113 ; Figure 3) ou si un rapport advient entre deux injections (n=70). La crainte de l'apparition d'effets indésirables en raison de leur activité sexuelle est importante.

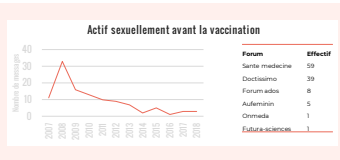


Figure 3 : Actif sexuellement - Volumétrie

Parallèlement les partenaires s'interrogent sur les mêmes thématiques (n=30). La peur de l'aveu d'une vie sexuelle aux parents est également source de préoccupations (n=126 ; Figure 4). Les adolescentes sont à la recherche d'un moyen d'informer leur médecin de leur activité sexuelle sans le signaler à leurs parents, présents lors de la consultation médicale.

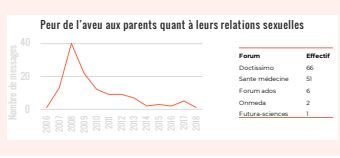


Figure 4 : Peur de l'aveu aux parents - Volumétrie

Le second groupe concerne le manque d'informations (196 messages 5.6% pré-corpus). Ce manque d'informations peut-être global (n=60, Figure 5) sur les modalités vaccinales : obligation vaccinale, efficacité, utilité, remboursement ...

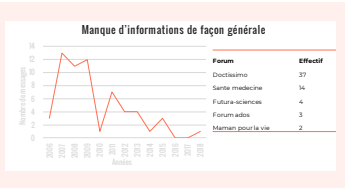


Figure 5 : Manque d'informations générale – Volumétrie

Le manque d'informations porte également sur les recommandations sur les schémas d'injections et les rappels (n=16) ou l'âge limite de vaccination, notamment si une adolescente est toujours vierge (n=15).
Le troisième groupe exprime l'influence qu'ont les différentes sources extérieures sur l'hésitation vaccinale (n=59, 1.8% du pré-corpus). Ces messages, de personnes en cours de vaccination anti-HPV ou non, mettent en avant une peur des effets indésirables (n=30 ; Figure 6) ou des aspects négatifs des injections (n=20) après consultation d'un site internet, de témoignages ou une discussion.
9 messages évoquent une source médicale déconseillant explicitement la vaccination anti-HPV malgré la volonté des parents et de l'adolescente de se faire vacciner.

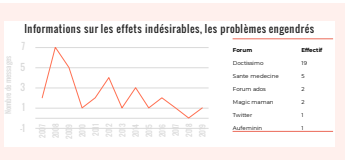


Figure 6 : Volumétrie impact sources extérieures

Appartenant aux groupes 2 et 3, les mères (n=39) justifient leur hésitation à faire vacciner leur fille par le manque d'informations claires et concordantes.

CONCLUSION

Cette étude a permis de mettre en avant l'existence d'une forte communauté d'internautes en débat et en recherche d'informations concernant la vaccination anti-HPV. **L'analyse du contenu des messages a permis d'isoler deux motifs principaux d'hésitation vaccinale : l'impact des relations sexuelles et le manque d'informations.** La compréhension fine de ces facteurs d'hésitation permettrait d'identifier les leviers associés. A des échelles locales, en fonction des comportements et caractéristiques des populations, ce type d'étude infodémiologique pourrait servir les stratégies et politiques de sante publique d'amélioration de la couverture vaccinale. Ces travaux sur l'hésitation vaccinale anti-HPV ont été effectués avant la recommandation de la HAS d'ouvrir la vaccination aux garçons. Ils pourraient être continués spécifiquement sur l'hésitation vaccinale exprimée sur internet par les adolescents afin de décrypter leurs motifs d'hésitation.

AFCROs
LES ENTREPRISES DE LA RECHERCHE CLINIQUE

2020

Conception d'un algorithme permettant de détecter l'hésitation vaccinale anti-papillomavirus humain au sein de messages issus des réseaux sociaux

Pierre Foulquié, Anaïs Gedik, Simon Renner, Paméla Voillot, Adel Mebarki et Stéphane Schück

Kap Code

CONCEPTION D'UN ALGORITHME PERMETTANT DE DÉTECTER L'HÉSITATION VACCINALE ANTI-PAPILLOMAVIRUS HUMAIN AU SEIN DE MESSAGES ISSUS DES RÉSEAUX SOCIAUX

Pierre Foulquié[1], Anaïs Gedik[1], Simon Renner[1], Paméla Voillot[1], Adel Mebarki[1] et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

Detec't

INTRODUCTION

La France est un des pays où l'hésitation vaccinale anti-papillomavirus humain (HPV) est la plus forte au monde. Cette hésitation s'observe particulièrement sur les réseaux sociaux, interface ou les internautes peuvent s'exprimer librement sur leur santé. L'amélioration de l'acceptabilité vaccinale HPV passe par la compréhension des déterminants de l'hésitation. **Un algorithme d'analyse sémantique capable d'identifier les messages exprimés sur les réseaux sociaux contenant une hésitation vaccinale anti-HPV permettrait d'analyser et de comprendre ce phénomène.**

MATÉRIEL ET MÉTHODES

Un corpus de messages associés à la vaccination anti-HPV, postés entre 2006 et 2019, a été extrait dans le cadre du projet Detec't [1] à partir de **17 sources francophones.** Les 23 mots-clés d'extraction évoquaient plusieurs sujets associés à la sphère du papillomavirus : **la vaccination anti-HPV, la sexualité et l'anatomie.** Une annotation d'un échantillon du corpus a été effectuée par 3 annotateurs qui disposaient d'une charte d'annotation basée sur la définition de l'hésitation vaccinale de l'OMS [2]. Elle a permis de classer les messages comme exprimant de l'hésitation ou non et d'extraire des expressions des différentes perceptions vaccinales (anti-vaccin, pro-vaccin). Le gold standard (GS) ainsi créé a été réparti en **2 jeux de données.** Un premier jeu, dit « d'entraînement » et contenant 85% des données, a été utilisé pour entraîner le modèle. Le deuxième jeu, désigné comme jeu « de validation » et constitué des 15% restants, a servi à la validation de la méthode.

À partir du jeu d'entrainement, plusieurs variables ont été déterminées à l'aide des formes syntaxiques des messages (N-grams), de la présence des mots de champs lexicaux spécifiques (anti-vaccin, pro-vaccin, etc.) et du *word embedding* (représentation contextuelle des mots via un modèle Glove [3]).

Par la suite et afin de mettre en place un modèle performant en termes de précision de détection d'hésitation vaccinale, une recherche de la meilleure combinaison entre différents classifieurs (*support vector classification*, *logistic regression*, *random forest*, et *extreme gradient boosting*) et les différentes variables identifiées précédemment a été effectuée.

RÉSULTATS

1 370 messages contenant une mention de vaccination anti-HPV ont été extraits pour annotation. Les sources les plus présentes dans l'échantillon étaient les forums Doctissimo et Santé Médecine, avec respectivement 615 et 193 messages, suivis de Twitter (n= 395). Le terme « Gardasil » est le mot d'extraction ayant permis de recueillir le plus de messages dans cet échantillon (n= 967).

L'annotation a permis d'identifier 497 messages (36%) représentant une hésitation vaccinale et 891 (64%) une perception positive, négative ou neutre de la vaccination. Les jeux de données d'entrainement et de validation étaient composés respectivement de 1 164 et 206 messages. La meilleure combinaison de variables identifiée est constituée des 300 premiers N-grams en terme de pouvoir prédictif de l'hésitation vaccinale. Ces formes syntaxiques sont composées d'un mot unique ou d'associations de deux ou trois mots telles que « savoir » ; « mère veut » ; « rapports sexuels ».

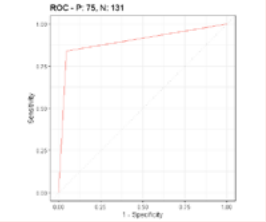
Suite aux tests, le meilleur modèle choisi est un classifieur binaire *Random Forest*, discriminant correctement l'hésitation des autres perceptions dans 91% des cas. Les performances du modèle, calculées à partir de la matrice de confusion (Tableau 1), sont regroupées dans le Tableau 2. Les messages classés comme évocateur d'hésitation vaccinale exprimaient une réelle hésitation dans 91 % des cas (valeur prédictive positive). Parmi tous les messages annotés comme exprimant une hésitation vaccinale, 84 % des messages ont été identifiés par notre modèle (sensibilité). Le Graphique 1 présente la courbe ROC du modèle.

		Référence	
		oui	non
Prédiction	oui	63	6
	non	12	125

Tableau 1 - Matrice de confusion du modèle

Précision	Sensibilité	Spécificité	Valeur prédite positive	Valeur prédite négative
91,26%	84,00%	95,42%	91,30%	91,24%

Tableau 2 – Performances du modèle



Graphique 1 – Courbe ROC du modèle

CONCLUSION

Développer un algorithme d'analyse sémantique capable d'identifier une hésitation vaccinale anti-HPV au sein de messages issus des réseaux sociaux pourrait se révéler être un **nouvel outil d'aide à l'identification des déterminants de l'hésitation dans le cadre d'une couverture insuffisante.**

Les performances de l'algorithme sur des données n'ayant pas servi au développement de ce dernier démontre que ce type d'outil est efficace pour identifier puis analyser les messages d'internautes exprimant une hésitation vaccinale anti-HPV.

Ce travail ouvre de nombreuses possibilités de travaux futurs. Des méthodes complémentaires, permettant par exemple d'identifier des causalités, pourraient permettre d'identifier les facteurs de cette hésitation. De plus, l'étude des utilisateurs exprimant une hésitation vaccinale permettrait d'établir des profils types et d'étudier l'évolution temporelle de cette hésitation. Ceci pourrait ouvrir la voie à l'instauration d'outils de monitoring de la vaccination sur les réseaux sociaux dans un objectif de santé publique.

RÉFÉRENCES

[1] Abdellaoui, R., Schück, S., Texier, N., & Burgun, A. (2017). Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help?. JMIR public health and surveillance, 3(2), e36.

[2] Ten threats to global health in 2019, https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

[3] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

DDW Digestive Disease Week®

2020

The Use of Web Forums Data to Evaluate Online Conversations Associated with Gastrointestinal Discomfort: A Retrospective 15-Year Study of 200 000 Messages from French-Speaking Platforms

Boris Le Nevé, Carole Faviez, Florent Schafer, Jean-François Jeanne, Paméla Voillot, Pierre Foulquié, Guy Fagherazzi, Stéphane Schück

Tu1636

THE USE OF WEB FORUMS DATA TO EVALUATE ONLINE CONVERSATIONS ASSOCIATED WITH GASTROINTESTINAL DISCOMFORT: A RETROSPECTIVE 15-YEAR STUDY OF 200 000 MESSAGES FROM FRENCH-SPEAKING PLATFORMS

Boris Le Nevé, Carole Faviez, Florent Schafer, Jean-François Jeanne, Paméla Voillot, Pierre Foulquié, Guy Fagherazzi, Stéphane Schuck

Introduction: Gastrointestinal (GI) discomfort (ex: bloating, flatulence) is very common and can significantly affect well-being and quality of life. Our aim was to explore online conversations on this topic. **MATERIAL & METHODS.** Messages related to GI discomfort were extracted from French-speaking generalist and specialized forums from January 2003 to August 2018 using the Detec't Extractor [1]. Messages were cleaned, deidentified and relevant medical concepts were identified using MedDRA 15.0 (Medical Dictionary for Regulatory Activities). Topic modeling was performed using a correlated topic model (CTM) based on the Latent Dirichlet Allocation (LDA). Users' age category and gender were identified respectively by linear regression and application of a SVM (Support Vector Machine). **Results:** A total of 198 866 messages were extracted from 14 major French-speaking web forums. Active users with available gender were mostly women (11630 vs 1651 men) under 30 years. Average number of messages per user was 4.57. Six classes of topics were identified through topic modelling: "medical consultations/exams", "diet", "symptoms", "quality of life", "treatments" and "stress and symptoms". Within the topic "diet", the most frequently discussed concept was "nausea and vomiting linked to food intake". **CONCLUSIONS.** GI discomfort is an actively discussed topic on French-speaking web forums. Gender and age of most active users tend to mirror the higher prevalence of functional gastrointestinal disorders like irritable bowel syndrome in young women. Our innovative approach has shown that identifying discussion topics associated to GI discomfort online is feasible and can serve as a complementary source of real-world evidence for caregivers. [1]. Abdellaoui R, Schück S, Texier N, et al. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help? JMIR public health and surveillance, 2017, vol. 3, no 2.

JMIR Publications
Advancing Digital Health & Open Science

2020

Concerns Discussed on Chinese and French Social Media During the COVID-19 Lockdown: Comparative Infodemiology Study Based on Topic Modeling

Stéphane Schück, Pierre Foulquié, Adel Mebarki, Carole Faviez, Mickaïl Khadhar, Nathalie Texier, Sandrine Katsahian, Anita Burgun, Xiaoyi Che

Abstract

Background:
During the COVID-19 pandemic, numerous countries, including China and France, have implemented lockdown measures that have been effective in controlling the epidemic. However, little is known about the impact of these measures on the population as expressed on social media from different cultural contexts.

Objective:
This study aims to assess and compare the evolution of the topics discussed on Chinese and French social media during the COVID-19 lockdown.

Methods:
We extracted posts containing COVID-19–related or lockdown-related keywords in the most commonly used microblogging social media platforms (ie, Weibo in China and Twitter in France) from 1 week before lockdown to the lifting of the lockdown. A topic model was applied independently for three periods (prelockdown, early lockdown, and mid to late lockdown) to assess the evolution of the topics discussed on Chinese and French social media.

Results:
A total of 6395; 23,422; and 141,643 Chinese Weibo messages, and 34,327; 119,919; and 282,965 French tweets were extracted in the prelockdown, early lockdown, and mid to late lockdown periods, respectively, in China and France. Four categories of topics were discussed in a continuously evolving way in all three periods: *epidemic news and everyday life*, *scientific information*, *public measures*, and *solidarity and encouragement*. The most represented category over all periods in both countries was *epidemic news and everyday life*. *Scientific information* was far more discussed on Weibo than in French tweets. Misinformation circulated through social media in both countries; however, it was more concerned with the virus and epidemic in China, whereas it was more concerned with the lockdown measures in France. Regarding *public measures*, more criticisms were identified in French tweets than on Weibo. Advantages and data privacy concerns regarding tracing apps were also addressed in French tweets. All these differences were explained by the different uses of social media, the different timelines of the epidemic, and the different cultural contexts in these two countries.

Conclusions:
This study is the first to compare the social media content in eastern and western countries during the unprecedented COVID-19 lockdown. Using general COVID-19–related social media data, our results describe common and different public reactions, behaviors, and concerns in China and France, even covering the topics identified in prior studies focusing on specific interests. We believe our study can help characterize country-specific public needs and appropriately address them during an outbreak.

JMIR Form Res 2021;5(4):e23593

doi:10.2196/23593

JMIR Publications
Advancing Digital Health & Open Science

2020

Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study

Jason Koman, Khristina Fauvelle, Stéphane Schuck , Nathalie Texier, Adel Mebarki

Abstract

Background:
Seeking medical information can be an issue for physicians. In the specific context of medical practice, chatbots are hypothesized to present additional value for providing information quickly, particularly as far as drug risk minimization measures are concerned.

Objective:
This qualitative study aimed to elicit physicians' perceptions of a pilot version of a chatbot used in the context of drug information and risk minimization measures.

Methods:
General practitioners and specialists were recruited across France to participate in individual semistructured interviews. Interviews were recorded, transcribed, and analyzed using a horizontal thematic analysis approach.

Results:
Eight general practitioners and 2 specialists participated. The tone and ergonomics of the pilot version were appreciated by physicians. However, all participants emphasized the importance of getting exhaustive, trustworthy answers when interacting with a chatbot.

Conclusions:
The chatbot was perceived as a useful and innovative tool that could easily be integrated into routine medical practice and could help health professionals when seeking information on drug and risk minimization measures.

J Med Internet Res 2020;22(11):e15185

doi:10.2196/15185

JMIR Publications
Advancing Digital Health & Open Science

2020

Exploring the Health-Related Quality of Life of Patients Treated With Immune Checkpoint Inhibitors: Social Media Study

François-Emery Cotté, Paméla Voillot, Bryan Bennett, Bruno Falissard, Christophe Tzourio, Pierre Foulquié, Anne-Françoise Gaudin, Hervé Lemasson 1, Valentine Grumberg, Laura McDonald, Carole Faviez, Stéphane Schück

Abstract

Background:
Immune checkpoint inhibitors (ICIs) are increasingly used to treat several types of tumors. Impact of this emerging therapy on patients' health-related quality of life (HRQoL) is usually collected in clinical trials through standard questionnaires. However, this might not fully reflect HRQoL of patients under real-world conditions. In parallel, users' narratives from social media represent a potential new source of research concerning HRQoL.

Objective:
The aim of this study is to assess and compare coverage of ICI-treated patients' HRQoL domains and subdomains in standard questionnaires from clinical trials and in real-world setting from social media posts.

Methods:
A retrospective study was carried out by collecting social media posts in French language written by internet users mentioning their experiences with ICIs between January 2011 and August 2018. Automatic and manual extractions were implemented to create a corpus where domains and subdomains of HRQoL were classified. These annotations were compared with domains covered by 2 standard HRQoL questionnaires, the EORTC QLQ-C30 and the FACT-G.

Results:
We identified 150 users who described their own experience with ICI (89/150, 59.3%) or that of their relative (61/150, 40.7%), with 137 users (91.3%) reporting at least one HRQoL domain in their social media posts. A total of 8 domains and 42 subdomains of HRQoL were identified: Global health (1 subdomain; 115 patients), Symptoms (13; 76), Emotional state (10; 49), Role (7; 22), Physical activity (4; 13), Professional situation (3; 9), Cognitive state (2; 2), and Social state (2; 2). The QLQ-C30 showed a wider global coverage of social media HRQoL subdomains than the FACT-G, 45% (19/42) and 29% (12/42), respectively. For both QLQ-C30 and FACT-G questionnaires, coverage rates were particularly suboptimal for Symptoms (68/123, 55.3% and 72/123, 58.5%, respectively), Emotional state (7/49, 14% and 24/49, 49%, respectively), and Role (17/22, 77% and 15/22, 68%, respectively).

Conclusions:
Many patients with cancer are using social media to share their experiences with immunotherapy. Collecting and analyzing their spontaneous narratives are helpful to capture and understand their HRQoL in real-world setting. New measures of HRQoL are needed to provide more in-depth evaluation of Symptoms, Emotional state, and Role among patients with cancer treated with immunotherapy.

J Med Internet Res 2020;22(9):e19694

doi:10.2196/19694

JMIR Publications
Advancing Digital Health & Open Science

2021

Assessing Patient Perceptions and Experiences of Paracetamol in France: An Infodemiology Study Using Social Media Data Mining

Stéphane Schück, Avesta Roustamal, Anaïs Gedik, Paméla Voillot, Pierre Foulquié, Catherine Penfornis, Bernard Job

Abstract

Background: Frequently, individuals are turning to social media to discuss medical conditions and medication, sharing their experiences and information and asking questions among themselves. These online discussions can provide valuable insights into individuals' perceptions of medical treatment, and increasingly, studies are focusing on the potential use of this information to improve healthcare management.

Objective: The objective of this infodemiology study was to identify social media posts mentioning paracetamol-containing products, to develop a better understanding of the patients' opinions and perceptions of the drug.

Methods: Posts containing at least one mention of paracetamol were extracted from 18 French forums between January 2003 and March 2019 with the use of the Detec't webcrawler. Posts were then analyzed using the automated Detec't tool which uses machine-learning and text-mining methods to inspect social media posts and extract relevant content.

Results: Overall, 44,283 posts were analyzed from 20,883 different users. Post volume over the study period showed a peak in activity between 2009 and 2012, as well as a spike in 2017 in the General group. The number of posts tended to be higher during winter each year. Posts were made predominantly by women (71%), with 12% made by men and 17% by individuals of unknown gender. The mean age of web users was 39 (±19) years. In the General group, pain was the most common medical concept discussed (22,257 posts, 50%) and paracetamol risk was the most common discussion topic, addressed in 8,902 posts (20.36%). Doliprane® was the most common medication mentioned (14,058 posts, 32%) within the General group, and tramadol was the most commonly mentioned drug in combination with paracetamol in the General group (1,038 posts, 5%). The most common unapproved indication mentioned within the Paracetamol Only group was fatigue (190 posts, with 16% positive for an unapproved indication), with reference to dependence made by 0.79% of the web users, accounting for 1.33% of the posts in the Paracetamol Only group. Dependence mentions in the Paracetamol and Opioids group were provided by 3.64% of web users, accounting for 5.44% of total posts. Reference to overdose was made by 245 web users across 291 posts within the Paracetamol Only group. The most common potential adverse event (PAE) detected was nausea (2.38% of posts) within the Paracetamol Only group

Conclusions: The use of social media mining with the Detec't tool provided valuable information on the perceptions and understanding of the web users, highlighting areas where providing more information for the general public on paracetamol, as well as other medications, may be of benefit.

PubMed

2021

Acceptance of a Covid-19 vaccine is associated with ability to detect fake news and health literacy

I Montagni, K Ouazzani-Touhami, A Mebarki, N Texier, S Schück, C Tzourio, CONFINS group

> J Public Health (Oxf). 2021 Mar 9;fdab028. doi: 10.1093/pubmed/fdab028. Online ahead of print.

Acceptance of a Covid-19 vaccine is associated with ability to detect fake news and health literacy

I Montagni [1], K Ouazzani-Touhami [1 2], A Mebarki [3], N Texier [3 4], S Schück [3 4], C Tzourio [1], CONFINS group

Affiliations + expand

PMID: 33693905 PMCID: PMC7989386 DOI: 10.1093/pubmed/fdab028

Free PMC article

Abstract

Background: During the Covid-19 pandemic fake news has been circulating impacting on the general population's opinion about a vaccine against the SARS-CoV-2. Health literacy measures the capacity of navigating health information.

Methods: We used data from a prospective national online cohort of 1647 participants. Descriptive statistics, Chi2 and ANOVA independence tests and two multivariable multinomial regression models were performed. Interactions between each variable were tested.

Results: Detection of fake news and health literacy scores were associated with intention to get vaccinated against SARS-CoV-2 ($p < 0.01$). The risk of being "anti-vaccination" or "hesitant", rather than "pro-vaccination", was higher among individuals reporting bad detection of fake news, respectively OR = 1.93 (95%CI = [1.30;2.87]) and OR = 1.80 (95%CI = [1.29;2.52]). The risk of being in "hesitant", rather than "pro-vaccination" was higher among individuals having a bad health literacy score (OR = 1.44; 95%CI = [1.04;2.00]). No interaction was found between detection of fake news and health literacy.

Conclusions: To promote acceptance of a vaccine against SARS-CoV-2, it is recommended to increase individuals' ability to detect fake news and health literacy through education and communication programs.

Keywords: Covid-19; fake news; health literacy; misinformation; vaccination.

© The Author(s) 2021. Published by Oxford University Press on behalf of Faculty of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.