Protocol for Evaluating the Extraction of Adverse Drug Reactions Information in Social Media, the ADR-PRISM Project

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

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 de Cordeliers, Paris, France
i Département d'Informatique Hospitalière, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France

Corresponding Author:
- Name, Surname: Girardeau, Yannick
- Full address: Unité de Mixte de Recherche 1138 Team 22, Institut National de la Santé et de la Recherche Médicale / Université Pierre et Marie Curie. Escalier D, 15 rue de l'école de médecine, 75006 Paris, France
  Phone +33 1 44279201
- e-mail: yannick.girardeau@aphp.fr
Abstract

Background: Social media is today seriously emerging as a source of information on post-marketing pharmacovigilance (PV), i.e. adverse events occurring with marketed drugs. Social media grew in importance in the very last years but still remains largely unexploited nowadays. Quite a few researchers published their work on either drug names or adverse effects (AE) recognition or drug adverse reactions (ADR) information extraction. However, a Gold Standard, consisting in manual annotations of the ADR by human experts from the corpus extracted from social media, was not systematically implemented and its quality is not always assessed on its own. When existing, the sample size is arbitrary only and doesn't rely on statistical arguments. This questions the statistical reliability of the results. At last, some of the extraction methods just take into account any co-occurrence of a drug entity with one or several AE entitie(s) which leads to extract general information such as the indication of the drug in addition to regular ADR.

Objective: In our work, we propose a standardized protocol for the evaluation of a software extracting information purely on ADRs.

Methods: Messages from French health forums are extracted (evaluation dataset). Drug and AE entities recognition are based on lexicons: Racine Pharma thesaurus and MedDRA terminology respectively. NLP-based techniques automate the ADR information extraction (relation between the Drug and AE entities). Several concepts of ADRs will be chosen. The corpus for evaluation is a random sample of the messages containing these concepts from the evaluation dataset. Two persons experienced in medical terminology will manually annotate the corpus (Gold Standard). The study outcomes will be the precision and recall and their confidence intervals at 95%. For each concept, these outcomes will be computed against the Gold Standard. Necessary and sufficient sample size will be calculated to ensure statistical confidence in the assessed results. Gold Standard in its-self will be evaluated through Kappa inter-annotators agreements. Further analyses will enable to explore the granularity in the terminologies.

Results: The automated ADR information extraction in the corpus for evaluation is already finished; the ADRs’ concepts as well as the sample for the Gold Standard are selected. The Gold Standard is completed, we currently proceed to the data analyses and the study results are expected in 2018.

Conclusions: This protocol is the first one that used standardized statistical methods to create the corpus for evaluation of a NLP tool, thus ensuring necessary statistical power of the assessed results. Once completed, our ADR information extraction software should interest different actors of our society (Health agencies, pharmaceutical companies, general public).

KEYWORDS
Social media, adverse events, adverse drug reactions, named-entities recognition, text mining, non-structured data, MedDRA, Racine Pharma.
Introduction

Overview

The detection of new adverse drug reactions (ADR) is mainly based on post-marketing surveillance by government agencies derived from spontaneous reporting by healthcare professionals and the public. The United States of America. Food and Drug Administrations (FDA) and the European Medicines Agency (EMA) collect potential ADR case reports in huge databases such as the U.S. Food and Drug Administration’s Adverse Event Reporting System (FAERS) [1-4] or the EudraVigilance system in European Union. Methods and tools have been developed over the years to identify new drug safety signal on these data sources [5-8].

In the last 20 years, Internet and social media have become an integral part of people’s daily life. Social media is now often used to communicate with other persons having the same health concerns and share information related to health condition, illnesses, feelings, medication use and many other aspects [9] Social media is therefore a potential provider of information on ADRs. In 2005, the International Society of Drug Bulletins already recognized such an usage: “ « Patient reporting systems should periodically sample and evaluate the scattered drug experiences patients report on the internet »” [10]. This new source of knowledge captured the interest from health informatics, statistics and public health researchers. Although in its infancy, related scientific literature increased in the last decade [1,11-13]

In this context, the French Ministry of Industry funded and launched the Adverse Drug Reactions from Patient Reports in Social Media (ADR-PRISM) project. The objective of ADR-PRISM was to make available the contents about ADR, informal and embedded in forums and discussions on the web, to the actors involved in drug safety (drug companies, agencies, and pharmacovigilance experts). In the end, the software developed in ADR-PRISM project should generate hypotheses about new or poorly-documented adverse effects. To reach its goals, the ADR-PRISM consortium gathers a company developing text mining software (Expert System), a company specialized in pharmaco-epidemiology (Kappa Santé), three academic research groups providing expertise in medical informatics and statistics (National Institute of Health and Medical Research & Cordeliers Research Centre Information Sciences to support Personalized Medicine– UMRS 1138-team 22, Laboratory in Medical Informatics and Knowledge Engineering in e-Health – LIMICS, and Biomedicine informatics Service Catalogue and Index of French Language medical websites - SIBM – CISMEF), two experts in pharmacovigilance (regional center of pharmacovigilance – CRPV), as well as Vidal Group, which supplies the drug database used in most drug prescription systems in France.

From a Natural Language Processing (NLP) perspective, we considered ADR as relationships between drug and adverse event (AE) concepts. Based on that, the NLP software developed by Expert System for ADR-PRISM includes a relation extraction module based on (named-) entity recognition combined with rules and regular expressions. Before applying it to large collections of forum discussions, we designed a protocol to assess the performance of this ADR information extraction module. We helped us with the quality standards used in clinical research. First we followed the Standards for Reporting Diagnostic Accuracy studies (STARD) [14]. Second, we adopted a methodology that guarantees the confidence in the results given for the main outcomes. We now present this protocol of evaluation. We synthesized the elements of this research in a synoptic table (Table 1).

Table 1 - Synopsis of the research

| Rational | Nowadays, patients easily use social media. On health forum, they write and share the adverse effect they feel due to their medications (further called Adverse Drug Reaction-ADR). Promising studies exist on the extraction of ADR information from social media. The authors mainly used NLP or Machine |


Learning tools. However, in none of them we could find the use the rigour of clinical research to assess the ADR extraction tool. Comparators are various and not always manual annotations which is considered as the Gold Standard yet.

The consortium ADR-PRISM has been constituted to create a tool extracting the ADR information found in the social media. Teams specialized in text mining, NLP and pharmacovigilance participates to the Consortium. Before using the ADR-PRISM’s tool on a large scale (on millions of posts), we need to evaluate the efficacy of this new tool as well as its limits as it can be done in the clinical research, to gain in robustness.

<table>
<thead>
<tr>
<th>Primary objective</th>
<th>To estimate robustly the performance of the ADR information extraction tool, against Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary endpoint</td>
<td>Drug expression and drug expression position and AE expression and AE expression positions and classification of the relationship between the drug and the AE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>To verify the quality of the Gold Standard</td>
</tr>
<tr>
<td>To estimate the performance of the ADR information extraction tool, under varying conditions</td>
</tr>
<tr>
<td>To estimate the performance of the ADR information extraction tool, in terms of AE expression extraction</td>
</tr>
<tr>
<td>To estimate the performance of the ADR information extraction tool, in terms of AE expression extraction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary endpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug expression and AE expression and classification of the relationship between the drug and the AE</td>
</tr>
<tr>
<td>Drug expression +/- drug expression position and AE expression +/- AE expression positions and classification of the relationship between the drug and the AE</td>
</tr>
<tr>
<td>Drug expression +/- drug expression position</td>
</tr>
<tr>
<td>AE expression +/- AE expression positions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recruitment</th>
</tr>
</thead>
<tbody>
<tr>
<td>283,000 posts belonging to the Kappa Santé Detec't database for which the Kappa Santé collection tool was able to extract at least one pharmaceutical or molecule name in + 300,000 posts randomly selected from the Kappa Santé Detec't database</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eligibility criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post published between 1st January 2007 and 31st January 2017</td>
</tr>
<tr>
<td>ADR concept: any explicit and positive relationship between a drug and an AE where either the drug was one of the 9 preliminary selected Drug concepts or the AE was one of the 6 preliminary selected AE concept or both the drug and AE belonged to preliminary selected concepts</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index test method</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADR information extraction tool: this tool classifies each co-occurrence of (Drug ; AE) as positive ADR or negative ADR or no ADR and maps the drug and the AE expression with,</td>
</tr>
</tbody>
</table>
respectively, a Racine Pharma and a MedDRA inputs

Reference method
Gold Standard: manual annotations of the co-occurrence of (Drug ; AE) as positive ADR or negative ADR or no ADR including the mapping of the drug and the AE expression with, respectively, a Racine Pharma and a MedDRA inputs. Manual annotations will be provided by 2 annotators with experience in medical terminology.

Sample size
349 extracted ADR concepts to obtain a 95% Confidence Interval, having a total width of 0.1, around the recall and precision estimated at 0.65.

Statistical analyses
- Recall, precision and F-measure calculations for evaluation of the performances of the ADR extraction information tool
- Inter-annotators agreement for the evaluation of the Gold Standard with a Cohen's Kappa

**Methods**

**Design and ethics**

**Design**
We based this project on retrospective data, collected among threads of discussion accessible on Social Media. The two experts in pharmacovigilance helped delineating the project. The objective is twofold: (i) focus on a pharmaceutical product of interest and identify the ADRs related to this drug, (ii) detect the emergence of any potential problem in Public Health. The two experts worked out a list of use cases further employed as basis for the analyses (Table 2).

**Table 2 List of drug and AE concepts selected as pharmacovigilance use cases [15,16]**

<table>
<thead>
<tr>
<th>Drug (active ingredient and French names)</th>
<th>Adverse effects</th>
<th>Corresponding MedDRA term</th>
<th>Media coverage or alert date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gardasil</td>
<td>Auto-immune disease, complex regional pain syndrome and postural orthostatic tachycardia syndrome</td>
<td>Autoimmune disorders (HLGT) + complex regional pain syndrome (PT) + postural orthostatic tachycardia syndrome (PT)</td>
<td>2013</td>
</tr>
<tr>
<td>Meningitec</td>
<td>Quality defect + AE</td>
<td>Cardiac disorders (SOC) + Nervous system disorders (SOC)</td>
<td>2009 (recommendation)</td>
</tr>
<tr>
<td>Methylphendate: Ritaline, Concerta, Medikinet, Quasym, Ritaline</td>
<td>Neurological and cardiac affectations</td>
<td>Cardiac disorders (SOC) + Nervous system disorders (SOC)</td>
<td>2013</td>
</tr>
</tbody>
</table>
We built up the global process of the evaluation of the ADR information extraction module in 2 phases (Figure 1).

In phase 1, we implemented an iterative process of validation and improvement of the software. With this phase, on the one hand, we aimed at correcting the most frequent errors and training the ADR information extraction module on extracting concepts of drugs and adverse events. On the other hand, we got estimates of performance indicators, including the time needed to perform a manual annotation of a corpus of social media messages.

In phase 2, we conducted the definite assessment of the performance of the ADR information extraction module approaching a methodology and conducting a statistical analysis, as rigorous as possible.

In the rest of the article, the term “AE” corresponds to medical events that are present in text and potential ADR, whereas we use the term “ADR” when a relation between a drug and an AE is established.

**Figure 1 Study overview**

<table>
<thead>
<tr>
<th>(1997-2011)</th>
<th>2013?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Méthadone</td>
<td>QT prolongation Qt interval prolongation</td>
</tr>
<tr>
<td>Sofosbuvir: Harvoni, EPCLUSA, Sovaldi</td>
<td>Bradycardia Bradycardia (PT)</td>
</tr>
<tr>
<td>Codeine: Codenfan, Codoliporane, Migralgine, Néocodion et Prontalgine</td>
<td>Respiratory disorders Respiratory disorders (HLGT)</td>
</tr>
<tr>
<td>Hydroxyzine: atarax</td>
<td>Rhythm disorder Cardiac arrhythmias (HLGT)</td>
</tr>
<tr>
<td>Nicorandil: Adancor, Ikorel</td>
<td>Skin ulceration Skin ulcer (HLT)</td>
</tr>
<tr>
<td>Midodrine: Gutron</td>
<td>Hypertension Blood pressure increased (PT)</td>
</tr>
<tr>
<td>Crizotinib: Xalkori</td>
<td>Heart failure Heart failures HLGT</td>
</tr>
<tr>
<td>Valproate de sodium: Depakine, Imaslav, Micropakine</td>
<td>Teratogenic effects Congenital, familial and genetic disorders (SOC)</td>
</tr>
<tr>
<td>Isotrétinoine: Curacné, Acnetrait, Contracné, Procuta, Roaccutane</td>
<td>Teratogenic effects + psychiatric disorders Congenital, familial and genetic disorders (SOC) and Psychiatric disorders (SOC)</td>
</tr>
<tr>
<td>Fingolimod: Gilenya</td>
<td>Leukoencephalopathy Toxic leukoencephalopathy (PT)</td>
</tr>
<tr>
<td>Aripiprazole: Abilify</td>
<td>Suicidal behaviour Suicidal and self injurious behaviour (HLGT)</td>
</tr>
</tbody>
</table>
Phase 1: pilot phase
1. Refer to Figure 2 for ADR-PRISM pipeline.
2. Sample size calculated on the following hypotheses: precision or recall = 0.65 ± 0.05 and alpha = 0.05.
3. Refer to Figure 4 for exact message selection.

Ethics
This research does not involve experiment on either humans or animals. Ethics and guarantee of data privacy constitutes an integral working group inside the ADR-PRISM project. To comply with national regulations, we first registered, to the French Data Protection Agency (CNIL for Commission Nationale Informatique et Liberté), a Normal Notification on the 23th December 2015 regarding data collection. We later submitted an Authorisation Request on 30th March 2016 regarding data analysis and validation of the approach adopted about ethics and confidentiality to the same Agency. ADR-PRISM’s consortium detailed approach is explained in [17].
An Ethics Advisory Board supported the ADR-PRISM project. The Board included scientists with different scientific expertise and its role was to give independent advice regarding ethical issues to the project consortium. The Board approved the conducted project.

**ADR-PRISM Adverse drug reaction extraction**

**Collection of Social Media Posts**
Kappa Santé Detec't Database collects discussions containing at least one word that can be either a pharmaceutical name or a molecule name [18]. Those words have been taken from the French Health Insurance database. All these discussions were extracted from five distinct websites: www.atoute.org, www.doctissimo.fr, www.e-sante.fr, www.aufeminin.com (which became www.onmeda.fr) and sante-medecine.journaldesfemmes.com. We selected these websites with search engines and net scoring tool application as described in [9]. The Google Site Search (GSS) crawls discussions on main website forums generating high numbers of posts.

**Pharmacovigilance and most sold Drugs Use Cases**
Eighteen use cases, combining one drug concept with one AE concept, were primarily provided by pharmacovigilance experts. The use cases were based on pharmacovigilance issues emerging in the last years (Table 2).
In addition to the drug concepts belonging to the 18 pharmacovigilance use cases, we constituted a group of drug concepts based on the most sold drugs in 2013. We selected the 10 most sold drugs with mandatory medical prescription for dispensation and the 9 most sold drugs with optional medical prescription for dispensation (Table 3).

**Table 3 List of drug and AE concepts selected as pharmacovigilance use cases**

5. **Optional medical prescription for dispensation**: Doliprane, Dafalgan, Efferalgan, Kardégic, Spasfon, Gaviscon, Dexéryl, Météospasmyl, Biseptine [19] (in French)

**Drugs, Adverse events and Adverse Drug Reaction Extraction**
ADR extraction was performed in two steps: first, a named-entity recognition module was used to identify drug names and AEs in posts, and then a relation extraction algorithm was applied to these entities (Figure 2).

**Figure 2 ADR pipeline**
Drugs and Adverse Events Extraction
For Drugs, Expert System has developed a named-entity recognition module capable of identifying words or tokens listed in the drug thesaurus “Racine Pharma” (5,164 inputs), maintained by the SIBM – CISMeF [20] and extracting their positions from forum posts. Racine Pharma is not currently available in open access and has already been described in connection with the project [17]. We chose the CISMeF Racine Pharma because of its exhaustiveness.

We mapped the extracted expressions from Racine Pharma with any word or token referenced in the chemical substance (5th level) in the ATC (Anatomical Therapeutic Chemical) hierarchy (Last updated version: 2016-12-19) [21]. Considering that the same active ingredient could be found under different trade names, due to marketing by different pharmaceutical companies, we pooled all mapped expressions within the same ATC chemical subgroup (4th level) to define the drug concepts.

Expert System has developed another module in order to identify and extract the position of the adverse events present in the posts. This module is based on the MedDRA (Medical Dictionary for Regulatory Activities) hierarchical terminology v15.1 (74,202 inputs in total, all levels mixed up) [22]. Fuzzy-matching and enrichment of thesaurus enabled to take into account misspelling, and colloquial language encountered in the patients’ or forums’ way of speaking.
Drug and Adverse Event Relationship Extraction

A Natural Language Processing module, developed by Expert System, performed the identification and the extraction of the specific information regarding the relationship established between a drug and an adverse event by the post's author. The module combined a set of rules and regular expressions, with a "Patient" lexicon constituted to ensure that the post's author set out a situation of a person taking the drug and experiencing the symptom, and exclude from the analysis general information regarding a drug or an AE. This lexicon held terms like “I”, “me”, “my”, “cause”, “test”, “take”, “feeling”, “because of”, “provoke”, “intolerance”, “allergies”, etc. This NLP module also takes negative sentence into account.

The algorithm can be summarized as follow:

6. Text is split into sentences based on the punctuation mark.
7. For each sentence, (named-)entity recognition modules extract drugs and AEs
8. For each pair of drug and AE co-occurring in a sentence, if the NLP module extracts specific information regarding a relation then the co-occurrence is classified as (i) explicit and positive ADR (e.g. “Abilify® causes me such fears that I cannot concentrate to read, work, etc.... »), (ii) explicit and negative ADR (e.g. « I took some Doliprane® and I didn’t feel any nausea »). If no specific information is identified, the co-occurrence is classified as (iii) no ADR (e.g. “Usually, Focalin’s adverse effects are loss of appetite, insomnia and naso-pharyngitis”).

ADR-PRISM platform

We developed a web application dedicated to manual annotations valuable as Gold Standard, for the project purposes in the phase 2 (see Figure 1). This application was based on Java Servlets and JavaScript libraries. This application was connected in JDBC to the dataset of posts selected for Gold Standard annotation. We used a self-completion mechanism to attach portions of message to Racine Pharma and MedDRA terms. We used drag-and-drop operations to fill in a table containing the manual review of the co-occurences. In this table, each line is dedicated to one co-occurrence. For each co-occurrence, the Drug was dragged-and-dropped in the first column and the AE was dragged-and-dropped in a second column. In the third column, the manual annotator was given a drop-down menu. The menu presented 3 possibilities for defining the co-occurrence: explicit and positive ADR or explicit and negative ADR or no ADR. We finally offered the possibility to export this table containing the manual annotations obtained via the application.

Phase 1 — Iterative improvement of the ADR information extraction module

Phase 1 consisted in an iterative process of validations and improvements of the tool. The output of Phase 1 is a first calculation of recall, precision and F-measure, assessing the capacity to extract information regarding drugs, AEs and ADRs as compared to the Gold Standard [23].

Phase 1 Corpus

Phase 1 Dataset

At the time of the first project’s phase, the Kappa Santé Detec’t Database collected all threads of discussions concerning 50 drugs of interest. Around 20 discussions have been selected at random for each drug. This subset constituted the dataset for the improvement of the Drug, Adverse Events and ADR information extraction modules.

Drug Concepts selection

We chose 12 drug concepts: 3 among the 18 pharmacovigilance use cases, 3 among the most sold drugs in 2013, 3 among the most frequent recognized drug expressions in the phase 1 dataset and 3 at random among all recognized expressions in the phase 1 dataset. Each drug concept had to be different from each other. For both the pharmacovigilance use
cases and the most sold drugs, our choice was guided also by the frequencies of the recognized named-entities to cope with the necessary sample size.

**AE Concepts selection**

We chose 9 AE concepts at preferred term (PT) level in the MedDRA hierarchy: 3 among the 18 pharmacovigilance use cases, 3 among the most frequent recognized entities in the phase 1 dataset and 3 at random among all recognized entities in the phase 1 dataset. Each AE concept had to be different from each other. For the pharmacovigilance use cases, our choice was guided also by the frequencies of the recognized entities to cope with the necessary sample size.

**Sample size calculations**

For both Drug and AE recognition, we hypothesized to observe a precision and recall of 0.5 of the ADR information extraction tool as compared to manual review. Based on these expected values, we needed at least 45 occurrences of one concept to obtain a 95% Confidence Interval having a total width of 0.3 around the estimates of the recall and precision.

**Random sample from Dataset**

In the full dataset for the improvement of the ADR information extraction modules, we randomly sampled, for each Drug concept and each AE concept, 45 occurrences among all the occurrences extracted by the modules. Pooling the two samples then constituted the corpus used for iterative improvements. We added, to this corpus for iterative improvements, all messages in which the ADR information module extracted at least one ADR (Figure 3).

**Figure 3 Phase 1 messages selection flowchart**

Manual review and partial manual annotations

The manual review on the corpus for iterative improvement was constituted by manual annotation of the messages by the two experts in pharmacovigilance specialists. We expected that the annotators would annotate all drug and AE relationships even if the drug and the AE were not found in the same sentence and even if the drug or the AE has not been selected as a concept. All along the process, both annotators could see the drug, adverse
events and ADR information extracted by the tool in order to validate or invalidate these extractions. In case help was needed, they could refer to an annotator guideline developed to standardize manual annotations. We expected that the annotators would annotate all drugs, AEs and drug-AE relationships that had not been extracted yet by the ADR information extraction tool. The manual annotation process took place on the software computerized by Expert System for in-house testing purposes [24].

**Statistical Analyses**

In order to assess the capacity to extract information regarding drugs, AEs and ADR as compared to the manual review, we calculated the recall, precision and F-measure for each concept as well as globally on the corpus for iterative improvements.

If the ADR information extraction tool extracted a relationship between the same Drug expression at the same position in the post, AND the same AE expression at the same position in the post, as did the Gold Standard, then we counted the extracted ADR as a true positive.

**Phase 1 results**

The full dataset for the improvement of the ADR information extraction tool included 967 discussions split into 325,435 messages posted between 2002 and 2014.

The corpus for iterative improvements pooled 561 messages containing the random selection of 45 occurrences of each Drug concept with 401 messages containing the random selection of 45 occurrences of each AE concept and the 1129 distinct messages containing all ADRs extracted by the ADR information extraction module (n=1,385 AE-Drug relationships).

The two experts in pharmacovigilance manually annotated the corpus twice. In the end, the Phase 1 resulted in an overall precision, recall and F-measures of 0.98, 0.90 and 0.94 for Drug; 0.98, 0.69, 0.81 for AEs and 0.78, 0.63 and 0.70 for ADRs. Results by concepts were published in [23].

**Phase 2 — Evaluation**

**Corpus for Evaluation**

*Evaluation Dataset*

For this phase 2, the selection of dataset is given in Figure 4.

**Figure 4 Phase 2 messages selection flowchart**
At the 28th October 2016, Kappa Santé Detec’t database contained about 23 millions of posts. Drugs, for which marketing authorization had been withdrawn before the last extracted post, were kept because some adverse effects appear at long-term. Based on the use cases (see definition above), we restricted the time period keeping the posts published between the 1st January 2007 and the 28th October 2016. At the date of 28th October 2016, restricted Kappa Santé Detec’t database contained approximately 21 million posts. Kappa Santé software for webdiscussions collection was able to extract at least one pharmaceutical or molecule name in 283,000 posts on the 21 millions. Due to time and calculation duration constraints, the Drug, AE and ADR extraction modules developed for ADR-PRISM, were executed on both all the 283,000 posts and a random sample of 300,000 posts among the 21 million minus 283,000 posts. We chose 300,000 additional posts to assess the proportion of potential drug names missed by Kappa Santé software for website posts collection.
The result of the execution of the ADR information extraction module is considered as the evaluation dataset.

For the evaluation of the ADR information extraction module, drug concepts, AE concepts and ADR concepts were selected as follows:

**Drug Concepts selection**
We chose 9 drug concepts at chemical subgroup level (4th level) of ATC hierarchy: 3 among the 18 pharmacovigilance use cases, 3 among the most frequent recognized drug expressions in the evaluation dataset and 3 among the most sold drugs in 2013. Each drug concept had to be different from each other. For both the pharmacovigilance use cases and the most sold drugs, our choice was guided also by the frequencies of the recognized named-entities to cope with the necessary sample size.

**AE Concepts selection**
We chose 6 AE concepts at preferred term (PT) level in the MedDRA hierarchy: 3 among the 18 pharmacovigilance use cases and 3 among the most frequent recognized entities in the evaluation dataset. Each AE concept had to be different from each other. For the pharmacovigilance use cases, our choice was guided also by the frequencies of the recognized entities to cope with the necessary sample size.

**ADR Concepts definition**
We defined an ADR concept as any explicit and positive relationship between a drug and an AE where either the drug was one of the 9 Drug concepts or the AE was one of the 6 AE concept defined above.

**Corpus for the Evaluation (random sample from evaluation dataset)**
Finally, we drew a random sample among all the ADR concepts extracted by the ADR information extraction module. This constituted our corpus for evaluation.

**Gold Standard**
The manual annotations of the corpus for evaluation by annotators with experience in medical terminology are valuable as Gold Standard. We expect that the annotators will annotate all drug and AE relationships even if the drug and the AE are not found in the same sentence. This is explained in an annotator guideline established to standardize manual annotations.

To assess inter-expert agreement, both annotators will annotate the same 10% subset of the evaluation corpus. In case of disagreement, the two annotators will meet each other to generate a consensus.

All along the process, both annotators will be double-blinded: first, from the ADRs identified by the ADR information extraction module and second, from the other annotator’s annotations.

The manual annotation process will take place on an interface computerized by the SIBM – CISMeF (cf ADR-PRISM platform presented before).

**Statistical Analyses**

**Primary Endpoint Analysis**
In order to assess the efficacy of the ADR information extraction module as compared to the Gold Standard, we will globally calculate the recall, precision and F-measure.

In a post, if the ADR information extraction module identifies the same expression from Racine Pharma at the same position, AND the same AE expression at PT level from MedDRA hierarchy at the same position AND the same type of relationship, as did the Gold Standard, then we will count the extracted ADR as a true positive (Table 4).

**Table 4 Primary outcomes definitions**
**Gold Standard's Evaluation**

We will assess the inter-annotators agreements by a hierarchical Kappa, calculated on the 10% evaluation corpus subset annotated in common [25]. A hierarchical Kappa enables to take into account the situation where the two annotators disagree in either the level or the expression in ATC or MedDRA terminologies but agree on higher levels than the annotated ones.

We will also provide separate calculations for inter-annotators agreements by hierarchical Kappa on drug and AE expressions.

**Secondary Endpoint Analyses**

We will complete these principal results with the following analyses.

We will take into account MedDRA and ATC granularities by reproducing precision, recall and F-measure at each level of the hierarchies (SOC, HLGT, HLT and anatomical main group, therapeutic subgroup, pharmacological subgroup, chemical subgroup, respectively).

We will provide a relaxed definition of true positive ADRs combining the three following conditions: (i) the positions in the post of the extracted expressions for Drug and AE both match the positions of the drug and AE expressions manually annotated, (ii) the extracted and the manually annotated expressions for AE from MedDRA hierarchy are found in the same SOC levels and (iii) the classification of the identified relationship match the
classification of the manually annotated relationship. We will calculate recall, precision and F-
measure with this definition for ADRs, Drug and AE identification separately.
We will provide a global estimation of the performance of ADR information extraction module
by taking into account other types of messages. These messages are those without
information on at least one ADR concept, i.e. messages in which (i) ADR information is
extracted but is not concerned by the drug or AE selected concepts, (ii) no ADR information
is extracted but only co-occurences of information about one drug and one AE, (iii) only AE
information is extracted, (iv) only drug information is extracted, (vi) neither drug nor AE
information is extracted. After having sampled each of types of messages described above,
we will then calculate global recall or precision or F-measure on all types of messages by
marginal calibration.

All analyses will be performed in ‘R’ Software [26].

Sample Size Calculation
Recall can be estimated from a random sample of ADR concepts annotated by the Gold
Standard while precision can be estimated from a random sample of ADR concepts extracted
by the ADR information extraction module. However creating the Gold Standard on the whole
evaluation dataset before sampling in it for the recall, is an overwhelming task. At the same
time, a vast majority of messages doesn’t contain any information on Drug concept as it is
shown by (roughly) the only 283,000 messages on a total of 21 millions (1.35%) containing at
least one pharmaceutical or molecule name in the restricted Kappa Santé Deteç't database
(only 283,000 messages on a total of 21 millions). Thus, a fortiori, we expect a very low
proportion of ADR concepts identified by the ADR information extraction module in the
evaluation dataset. Hence, the recall estimated from a random sample of ADR concepts
annotated by the Gold Standard is mathematically approximated by the recall estimated on
ADR concepts sampled for the precision.
As mentioned above, the first phase about the iterative improvement of the tool enabled to
show precision and recall values about 0.78 and 0.63 respectively, for the ADR information
extraction module [23].
To evaluate the efficacy of ADR information extraction tool, as compared to the Gold
Standard, by hypothesizing a global recall of 0.65, we will need at least 349 identified ADR
concepts to obtain a 95% Confidence Interval, having a total width of 0.1 around the
estimates of the recall and precision.

Again, we synthesized our methodology for evaluation in Table 1.

Results
At the time we publish this protocol, we already completed several steps. We constituted the
corpus for evaluation, randomly sampling in the evaluation dataset. The NLP tool identified
and extracted the information about ADR inside this corpus. We selected the ADR concepts
and constituted the samples of the entities necessary to set up the Gold Standard. The two
annotators completed the annotations process.
Data analyses for assessing both the performance of the ADR information extraction module
as compared to the Gold Standard, and the inter-annotators’ agreement are ongoing and
study results are expected in 2018.

Discussion
We presented here the protocol of evaluation of the ADR information extraction software
developed in the project. We adopted a methodology that guarantees the confidence in the
results given for the main outcomes.
Study Strengths
One of our major concerns is to apply standards that have been developed over the years in clinical research to the evaluation of the ADR extraction module. For both phases of this study, but especially for phase 2 – evaluation, we chose to follow the STARD statement [14] used in clinical research protocols to assess diagnostic tests. First, we paid particular attention to the Gold Standard and its validation. Two experts in medical terminologies will perform manual annotations in a “double-blind” manner, from both the other annotator and the ADR identified by the ADR information extraction module. This constitutes a valuable Gold Standard. In some of former publications (see Supplementary file 1), a Gold Standard was implemented [27-41] but it is not systematical [42-47]. In the latest, the extracted ADRs were compared to either known AEs from FAERS [27,43,45,46] or drug label declared to FDA [42,47] or even AE described in websites [44]. We will also ensure ourselves about the quality of this Gold Standard by evaluating it with inter-annotator’s agreement. If in most of studies, the authors gauged the inter-annotator’s agreement [28,29,31,34,35,38,39], it is not always the case [27,28,30,33,36,37,40,48,49], The quality then remains questionable.
Second, by calculating a sample size of messages collected from social media, to assess recall, precision and F-measure, we will guarantee the statistical power to place reliance on our study results. This is the first time to our knowledge that designs and methods usually used in clinical research are applied to assess a NLP software. We claim that IT methods that may be used to support decision require rigorous evaluation based on clinical research standards level.

The chosen terminologies are another crucial aspect of this work. On the one hand, the “Racine Pharma” thesaurus, with 5.164 inputs, exhaustively covers a large range of drug names and active ingredients. On the other hand, the MedDRA hierarchy is daily used by drug safety experts who benefit from its expressivity. By using expressive terminologies, we expect to increase the sensitivity of the ADR-PRISM software. Our choice to map Racine Pharma to ATC was guided by two aspects. First, ATC like MedDRA, have hierarchical organizations. Thus, we will be able to provide outcomes of the performance of the ADR information extraction tool at the different levels of these terminologies. Second, ATC like MedDRA are widely used and internationally agreed reference terminologies. Hence, we expect to provide strong and reproducible outcomes.

The ADR information extraction module is not only based on drug and AE information identification but also on rules and regular expressions. As such, we expect to discard non-informative sentences, addressing generalities or drug indications. We would also be able to extract unexpected adverse “positive effects”, as for example, headaches that would be reduced by a drug without indication for the treatment of this kind of pain. Only few studies have been able to take this aspect into account [31,33,34,36,38,50].

Study Limitations
Despite its positive aspects, the study exhibits several limitations.

The mapping between the terms listed in the drug thesaurus “Racine Pharma” and the terms referenced in the chemical substance in the ATC hierarchy is incomplete. On the 5.164 inputs in the thesaurus, 852 inputs (16.5%) could not be mapped with an ATC term as, for example some phytotherapies (St John’s wort herbal tea, Silver birch juice, extract of liquorice root, arum triphyllum compound, arnica, etc). This could lead to lessen the recalls calculated according to the ATC levels.

In social media’ posts, slangs and colloquial languages are frequent likewise syntactic rules are approximate. We chose to use fuzzy-matching and enrichment of thesaurus to take into
account this bias. But this work can’t be exhaustive and it could lead to a decrease of the recall of the ADR information extraction module.

Regarding the Gold Standard, two experts in medical terminologies will perform manual annotations. However contrary to the phase 1 manual review, none of them can be considered as a pharmacovigilance expert.

The medical informatics community needs shared open corpora to evaluate their methods and tools. Recent efforts have led to make several datasets more accessible and the evaluation of the methods more standardized: like the MIMIC corpus [51], and the i2b2 challenges [52]. Recently, a set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—has established a set of principles that we refer to as the FAIR (Findable, Accessible, Interoperable, Re-usable) Data Principles [53]. The FAIR principles support the idea that automated methods are being developed to search for relevant data sources, to analyse the datasets, and to mine the data for scientific discovery. Drug safety could benefit from the FAIR initiatives. Extending the assessment of the ADR-PRISM software to another international dataset could be very useful in order to ensure external validity of our results and to compare our tool to others. Gonzales et al. [30] proposed an open-access validation corpus. However, the ADR information extraction module is mainly developed for French language at the moment and timelines are too short to obtain this open-access corpus.

**Conclusion**

The objective of this article is to present the scientific approach developed in the first stage of the ADR-PRISM project. In this stage, our principal objective is to evaluate the performance of the ADR information extraction software against a Gold Standard constituted by human annotations. To address this question, we adopted rigor from clinical research to assess sound and trustworthy measurements of precision, recall and F-measure. With the statistical theory, we calculate a sample size. This now guarantees enough ADR information and sufficient narrowness of the confidence interval, to scientifically conclude on our principal objective. We also avoid unnecessarily large extractions that mean manual annotations process that are wasteful and time-consuming.
Acknowledgements

This work was labelled in 2013 by the competitiveness cluster Cap Digital and funded by the DGE and territorial collectivities (Ile de France and Haute Normandie) under the 16th FUI (for French Unique Interministerial Fund) request for proposal through the ADR-PRISM project.

Conflicts of Interest

Abbreviations

ADR: adverse drug reaction
ADR-PRISM: adverse drug reaction from patient reports in social media
ATC: anatomical therapeutic chemical (in ATC classification)
AE: adverse event or adverse effect
CNIL: data protection authority (French commission nationale de l'informatique et des libertés)
CRPV: regional pharmacovigilance centres (French centre régional de pharmacovigilance)
EMA: European medicine agency
FAERS: U.S. FDA adverse event reporting system
FAIR: findable, accessible, interoperable, re-usable
FDA: U.S. food and drug administrations
FUI: unique interministerial fund (French fond unique interministeriel)
GSS: Google site search
HLGT: high level group term (in MedDRA)
HLT: high level term (in MedDRA)
I2b2: informatics for integrating biology and the bedside
IT: information technology
JDBC: Java database connectivity
LIMICS: laboratory in medical informatics and knowledge engineering in e-Health (French laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé)
LLT: lowest level term (in MedDRA)
MedDRA: medical dictionary for regulatory activities
MIMIC: multiparameter intelligent monitoring in intensive care
NER: named-entities recognition
NLP: natural language processing
PT: preferred term (in MedDRA)
PV: pharmacovigilance
SIBM-CISMeF: biomedicine IT service – catalogue and index of French language medical websites (French Service d'Informatique Biomédicale-Catalogue et Index des Sites Médicaux de langue Française)
SOC: system organ class (in MedDRA)
STARD: standards for reporting diagnostic accuracy studies
UMRS: health research mixed unit (French unité mixte de recherche en santé)
References


15. ANSM ANdSDmedPdS. ANSM - Répertoire des Spécialités Pharmaceutiques. Archived by WebCite® at http://www.webcitation.org/70UXwdakk
16. VIDAL APHP. Base de données VIDAL pour professionnels de santé de l'APHP (Assistance Publique Hôpitaux de Paris).


