

## Drug safety and big clinical data Detection of drug-induced anaphylactic shock events

Guillaume Bouzille, Marie-Noëlle Osmont, Louise Triquet, Natalia Grabar, Cécile Rochefort-Morel, Emmanuel Chazard, Elisabeth Polard, Marc Cuggia

► **To cite this version:**

Guillaume Bouzille, Marie-Noëlle Osmont, Louise Triquet, Natalia Grabar, Cécile Rochefort-Morel, et al.. Drug safety and big clinical data Detection of drug-induced anaphylactic shock events. *Journal of Evaluation in Clinical Practice*, Wiley, 2018, 24 (3), pp.536-544. 10.1111/jep.12908 . hal-01833093

**HAL Id: hal-01833093**

**<https://hal-univ-rennes1.archives-ouvertes.fr/hal-01833093>**

Submitted on 13 Jul 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Drug safety and big clinical data: detection of drug-induced anaphylactic shock events

Running head: Detecting anaphylaxis from big clinical data

Guillaume Bouzillé, MD MSc<sup>1234\*</sup>, Marie-Noëlle Osmont, PharmD<sup>5</sup>, Louise Triquet, PharmD<sup>5</sup>, Natalia Grabar, PhD<sup>67</sup>, Cécile Rochefort-Morel, MD<sup>8</sup>, Emmanuel Chazard MD<sup>9</sup> PhD, Elisabeth Polard, PharmD<sup>5</sup> and Marc Cuggia MD PhD<sup>1234</sup>

<sup>1</sup>INSERM, U1099, Rennes, F-35000, France

<sup>2</sup>Université de Rennes 1, LTSI, Rennes, F-35000, France

<sup>3</sup>CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France

<sup>4</sup>CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France

<sup>5</sup>CHU Rennes, Centre Régional de Pharmacovigilance, Rennes, F-35000, France

<sup>6</sup>CNRS, UMR 8163, F-59000 Lille, France

<sup>7</sup>Université de Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France

<sup>8</sup>CHU Rennes, Service de pneumologie et d'allergologie, Rennes, F-35000, France

<sup>9</sup>Département de Santé Publique, Université de Lille EA 2694, CHU Lille, F-59000 Lille, France

\* Corresponding author

E-mail: [guillaume.bouzille@univ-rennes1.fr](mailto:guillaume.bouzille@univ-rennes1.fr)

Phone: +33 6 15 71 12 30

Fax: N/A

Authors contacts:

- Guillaume Bouzillé: [guillaume.bouzille@univ-rennes1.fr](mailto:guillaume.bouzille@univ-rennes1.fr)
- Marie-Noëlle Osmont: [Marie-noelle.OSMONT@chu-rennes.fr](mailto:Marie-noelle.OSMONT@chu-rennes.fr)
- Louise Triquet: [Louise.TRIQUET@chu-rennes.fr](mailto:Louise.TRIQUET@chu-rennes.fr)
- Natalia Grabar: [natalia.grabar@univ-lille3.fr](mailto:natalia.grabar@univ-lille3.fr)
- Cécile Rochefort-Morel: [cecile.rochefort@chu-rennes.fr](mailto:cecile.rochefort@chu-rennes.fr)
- Emmanuel Chazard: [emmanuel.chazard@univ-lille2.fr](mailto:emmanuel.chazard@univ-lille2.fr)
- Elisabeth Polard: [elisabeth.polard@chu-rennes.fr](mailto:elisabeth.polard@chu-rennes.fr)
- Marc Cuggia: [marc.cuggia@univ-rennes1.fr](mailto:marc.cuggia@univ-rennes1.fr)

Work was done at the Clinical Data Center of Rennes Academic Hospital.

**Keywords:** Adverse Drug Reaction Reporting Systems, Drug-Related Side Effects and Adverse Reactions, Electronic Health Records, Information Storage and Retrieval

## **Abstract**

**Rationale, aims and objectives:** The spontaneous reporting system currently used in pharmacovigilance is not sufficiently exhaustive to detect all adverse drug reactions (ADRs). With the widespread use of electronic health records (EHRs), biomedical data collected during the clinical care process can be reused and analyzed to better detect ADRs. The aim of this study was to assess whether querying a Clinical Data Warehouse (CDW) could increase the detection of drug-induced anaphylaxis.

**Methods:** All known cases of drug-induced anaphylaxis that occurred or required hospitalization at Rennes Academic Hospital in 2011 (n=19) were retrieved from the French pharmacovigilance database, which contains all reported ADR events. Then, from the Rennes Academic Hospital CDW, a training set (all patients hospitalized in 2011) and a test set (all patients hospitalized in 2012) were extracted. The training set was used to define an optimized query, by building a set of keywords (based on the known cases) and exclusion criteria to search structured and unstructured data within the CDW in order to identify at least all known cases of drug-induced anaphylaxis for 2011. Then, the real performance of the optimized query was tested in the test set.

**Results:** Using the optimized query, 59 cases of drug-induced anaphylaxis were identified among the 253 patient records extracted from the test set as possible anaphylaxis cases. Specifically, the optimal query identified 41 drug-induced anaphylaxis cases that were not detected by searching the French pharmacovigilance database, but missed seven cases detected only by spontaneous reporting.

**Discussion:** We proposed an information retrieval-based method for detecting drug-induced anaphylaxis, by querying structured and unstructured data in a CDW. CDW queries are less specific than spontaneous reporting and DRG queries, although their sensitivity is much higher. CDW queries can facilitate monitoring by pharmacovigilance experts. Our method could be easily incorporated in the routine practice.

## **1. Introduction**

Pharmacovigilance is an essential component of drug safety. The main goals of pharmacovigilance are the post-marketing drug surveillance and reducing the risk of adverse

drug reactions (ADR) associated with drug use. The French pharmacovigilance system is based on a network of 31 regional pharmacovigilance centers. Spontaneous reporting (i.e., the unrequested notification of ADR events) by healthcare professionals or consumers to pharmacovigilance centers is the cornerstone of this surveillance system. Moreover, pharmaco-epidemiological studies assess the benefit/risk ratio when a drug is prescribed in “real life” to a large population of patients. Safety data are then analyzed by the competent authorities at different levels (national, European, global) that can then call attention to pharmacovigilance signals. Validated signals allow the authorities to develop regulatory measures to better control drug-related risks.

Spontaneous reporting has the advantage of covering a large number of patients (ideally, the entire population) and a wide range of drugs. It is a cost-effective method for monitoring drug safety <sup>1</sup>, but has some limitations, mainly the under-reporting of ADR events, particularly those occurring in hospitals or leading to hospitalization. Indeed, it is estimated that more than 90% of ADR events are currently not reported to the authorities <sup>1</sup>.

Several studies have assessed whether medico-economic databases could be used to improve ADR detection <sup>2-4</sup>. These authors detected more ADRs using the billing codes of the international classification of diseases, 10<sup>th</sup> edition (ICD-10), compared to the number of ADRs identified via spontaneous reporting for the same period. Moreover, they observed little or no overlap between the numbers of ADR events detected using the two methods.

Similarly, the French Medical Information System Program (PMSI) database generates Diagnosis-related Groups (DRGs) that contain administrative and medical data, including diagnoses that are classified according to the ICD-10 codes. By querying this database using selected ICD-10 billing codes, it is possible to identify serious ADR events. This system is now routinely used by the Regional Pharmacovigilance Center of Rennes<sup>5</sup> to detect serious ADRs, such as generalized rash, anaphylaxis, nephropathy, liver damage, polyneuropathy, neuroleptic malignant syndrome and interstitial lung disease (see Appendix 1 for a list of ICD-10 billing codes used by the Regional Pharmacovigilance Center of Rennes). All hospital stays involving a hospitalization summary that includes at least one of the selected ICD-10 billing codes are investigated to validate whether the extracted ADR is relevant to pharmacovigilance.

Big data analysis also could contribute to improving the relevance of the detected signals. In France, this could be done with the support of institutions that run large databases and/or manage several health professionals' databases, such as the French National Agency for Medicines and Health Products Safety (ANSM), the French National Health Insurance Agency (CNAM), the French Public Regional Health Agencies (ARS), the Hospital Information Technology Agency (ATIH), or the French national inter-scheme information system of health insurance (SNIIRAM, Système national d'information inter-régimes de l'Assurance maladie), which combines the French healthcare reimbursement system and PMSI databases. Nevertheless, they should be considered as extra sources of information rather than as replacements for spontaneous notifications that remain the most effective tool to date <sup>6</sup>. As these methods are not exhaustive, the use of other data sources and methods becomes necessary to effectively identify ADR events in hospitals <sup>7</sup>. Due to the widespread use of electronic health records (EHRs), biomedical data collected during the clinical care process could be reused and analyzed to improve ADR detection<sup>8-10</sup>. Technologies, such as Clinical Data Warehouses (CDW), have these abilities. They are currently deployed in many hospitals, thus making possible the efficient exploitation of clinical data <sup>11-13</sup>. We believe that such technologies can also be used to address some of the issues currently encountered in pharmacovigilance.

We are interested in improving the detection of anaphylactic shock events. We previously carried out a study to evaluate the performance of a data-gathering method using a CDW for cases of anaphylactic shocks (IgE-mediated) that specifically occurred during anesthesia at the Academic Hospital of Rennes (CHU-RENNES) <sup>14</sup>. In that study, anaphylactic shock during anesthesia proved to be an easily identifiable clinical entity. However, we think that the identification of anaphylactic shock events due to other causes also could be improved by using the same kind of information retrieval method. Therefore, the objective of the current study was to assess whether clinical data from narrative EHRs can help pharmacovigilance centers to detect drug-induced anaphylaxis. More specifically, we propose an optimal method, based on a CDW, to detect drug-induced anaphylaxis events.

## 2. Material and methods

### 2.1. Data sources and studied population

At the CHU-RENNES, we developed our own CDW solution, called eHOP (formerly Roogle) <sup>15,16</sup>. Briefly, this CDW integrates all types of documents produced by the hospital information system and connected with healthcare:

- structured data using reference terminologies (e.g., ICD-10 diagnoses from DRGs, local terminology codes for laboratory tests, Association for the Development of Informatics in Cytology and Pathology codes for pathology diagnoses, Anatomical Therapeutic Chemical terminology corresponding to drug prescriptions and administration);
- unstructured data, such as clinical narrative notes, surgical protocols, X-rays or pathology reports.

Hence, a unique attribute of eHOP is that it allows users to search for information from both structured and unstructured data. Additionally, two different ways of querying data can be combined. Users can build queries based on reference terminologies (e.g., ICD-10), or simply submit keywords to retrieve both structured (e.g., terminology labels) or unstructured documents that contain these terms or keywords. Users can then access documents via a dedicated interface that incorporates functionalities to allow navigating through the entire patient EHR. eHOP is routinely used at the CHU-RENNES to support clinical research in feasibility studies, or for screening patients for eligibility criteria.

The eHOP CDW currently provides the possibility to search among 25 million unstructured data and 170 million structured elements. Some unstructured data, such as laboratory results or diagnoses, are also recorded in a structured form thanks to the corresponding terminological codes. All these data are collected from EHRs and cover more than 1.2 million patients.

For this study, we defined two datasets from eHOP: i) a training set that contained information on patients who stayed at the hospital between January 1 and December 31, 2011; and ii) a test set that contained data on patients who stayed at the hospital between January 1 and December 31, 2012.

The inclusion period of one year for the two sets was extended by three months to ensure that the data obtained would be complete. This was done to take into account possible delays in the production of documents for patients with a suspected anaphylaxis event who, at the end of the studied period, had allergy-related consultations. All documents related to these patients and produced during the studied period (including the three additional months) were used as part of the information retrieval process.

## **2.2. Definition of anaphylaxis**

This study focused on patients who had drug-induced anaphylaxis. Anaphylaxis is an acute, potentially life-threatening hypersensitivity reaction that involves the release of mediators from mast cells, basophils and recruited inflammatory cells. Anaphylaxis is defined by a number of signs and symptoms, alone or in combination, that occur within minutes, or up to a few hours after exposure to a causative agent<sup>17</sup>. Concerning the underlying mechanisms, the anaphylactic reaction can be IgE-mediated or non-IgE-mediated. Based on the reaction severity, anaphylaxis can be classified in four levels:<sup>15</sup>

- level I: presence of cutaneous signs;
- level II: presence of measurable, but not life-threatening symptoms, including cutaneous effects, arterial hypotension, cough or ventilation difficulties;
- level III: presence of life-threatening symptoms (e.g., cardiovascular collapse, tachycardia or bradycardia, arrhythmias, severe bronchospasms);
- level IV: circulatory failure, cardiac and/or respiratory arrest.

All anaphylaxis episodes that occurred at CHU-RENNES or led to hospitalization or required allergy investigations at the hospital were considered relevant for our study. As in France, there are 31 regional pharmacovigilance centers, we included only the anaphylaxis cases that occurred in the geographic area covered by the Regional Pharmacovigilance Center of Rennes.

Moreover, a given patient was included in the study and considered as having had a new anaphylaxis episode each time that she/he was hospitalized or required allergy investigations at the hospital for drug-induced anaphylaxis during the study period.

### **2.3. Extraction of the known drug-related anaphylaxis cases**

The known cases of drug-related anaphylaxis (reference drug-related anaphylaxis cases throughout the text) that occurred within or required care at CHU-RENNES during the training period (2011) were retrieved from the French pharmacovigilance database, which contains all reported ADRs, spontaneously reported and identified using DRGs (based on two ICD-10 billing codes: T88.2 and T88.6) (see Figure 1 for the study design). The *Medical Dictionary for Regulatory Activities* (MedDRA®) term “Anaphylactic responses” and the high-level term belonging to the System Organ Classes “Immune system disorders” were used to query the database. Criteria on the active substance responsible for the anaphylaxis were not included in the query.

The same procedure was used to extract the drug-related anaphylaxis cases from the French pharmacovigilance database for 2012 for the comparison with the results obtained with the test set (2012).

### **2.4. Identification of keywords related to the reference drug-related anaphylaxis cases**

Two different methods were used to identify potentially relevant keywords:

- a) Experts from the Regional Pharmacovigilance Center of Rennes reviewed the discharge summaries of the reference drug-related anaphylaxis cases extracted from the French pharmacovigilance database, in order to identify terms used by clinicians in connection with such episodes.
- b) Analysis of the keywords used in the previous study to detect cases of anaphylactic shock occurring during anesthesia<sup>14</sup>. These keywords were "anaphylaxis" and "anesthesia". Only the cases in which "anaphylaxis" OR "anaphylactic" OR "anaphylactoid" was not mentioned in the discharge summaries could not be identified in this study. Moreover, these previous results suggested that keywords connected with symptoms of anaphylaxis were not sufficiently specific.

### **2.5. Definition of the optimized query**



To define the optimized query, the set of keywords identified in 2.4 were used to query structured and particularly unstructured data to find, at least, the reference drug-related anaphylaxis cases in the training set (2011) from the eHOP CDW. Moreover, the performance of each keyword was independently tested in terms of retrieval of the reference drug-related anaphylaxis cases, by using the following standard evaluation measures:

- Precision: fraction of drug-related anaphylaxis cases among all retrieved patients
- Recall: fraction of retrieved drug-related anaphylaxis cases among all anaphylaxis cases
- Synthetic F-measure:  $2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$

A given keyword was considered to be relevant if its recall was at least 0.5. The recall measure was prioritized because it is essential for pharmacovigilance experts not to miss any relevant case. The optimized query was built by combining the retained keywords (recall  $\geq 0.5$ ) with the help of the inclusive disjunction (“OR”), so that the query would retrieve all documents matching at least one of the relevant keywords.

## **2.6. Evaluation of the optimized query in the test set**

The real performance (precision, recall and F-measure) of the optimized query and of each keyword forming this query (i.e., all keywords with a recall  $\geq 0.5$  in the training set) was assessed in the test set.

Two pharmacovigilance experts from the Regional Pharmacovigilance Center of Rennes reviewed all potential cases returned by the query in an independent and blind validation process to confirm that they were real drug-related anaphylaxis cases. Any disagreements between experts required a consensus session before reaching the final decisions on the patient’s status and relative to the definition of anaphylactic shock (and to the inclusion criteria) mentioned above. Validated cases from the test set were then compared with the drug-related anaphylaxis cases (spontaneous and DRG-based reporting) identified in the French pharmacovigilance database for 2012. Duplicated cases present in different data sources were identified using administrative data (initials, date of birth, etc.), the characteristics of the side effects, suspected drugs and date of occurrence. This step was necessary because a given patient could present with several episodes of anaphylaxis during the study period.

## **2.7. Final tuning of the query**

During the final step, exclusion criteria were added to the optimized query, consisting of keywords connected with the causes of anaphylaxis, identified by experts, in patients who had not suffered drug-induced anaphylaxis. This pruning phase was validated by testing the optimized, pruned query using the training set to check whether all reference drug-induced anaphylaxis cases were still detected.

The study design is summarized in Figure 1.

## **2.8. Ethics statement and funding sources**

This study was approved by the Ethics Committee of the academic hospital of Rennes and performed in accordance with the Declaration of Helsinki guidelines.

The funding source, the French National Agency for Medicines and Health Products Safety (ANSM), had no role in the study.

## **3. Results**

### **3.1. Extraction of reference drug-induced anaphylaxis cases from the French pharmacovigilance database**

In total, 30 anaphylaxis cases were identified in the French pharmacovigilance database for the area covered by the Regional Pharmacovigilance Center of Rennes for the year 2011. Among them, 19 cases were selected as confirmed cases that occurred or required care at the CHU-RENNES. For the remaining 11 cases, 9 did not occur at the CHU-RENNES, one had a final diagnosis of mastocytosis and one included an anaphylactoid reaction.

### **3.2. Keyword identification and evaluation in the training set**

On the basis of the review of the 19 reference drug-related anaphylaxis cases by the experts at the Regional Pharmacovigilance Center of Rennes and the previous study on anaphylaxis during anesthesia, the following keywords were selected:

- Anaphylaxis query: words prefixed by “anaphy.”
- Tryptase query: “level” AND “tryptase” OR “increase” AND “tryptase,” with a maximum of two words between them.
- Allergo-anesthesia query: words prefixed by “allergo-anesth.”
- Allergen-specific IgE query: “allergen-specific IgE.”

- Prick testing query: “prick-test.”
- Intradermal allergy testing query: “intradermal” OR “intra-dermal.”
- Histamine query: words “histamine” AND “increased” with a maximum of two words between them.
- Contraindication query: “strictly contraindicated.”
- Immunoallergic query: words prefixed by “immunoaller.”
- RAST inhibition query: “RAST inhibition.”

The next step was to evaluate the capacity of different queries in which the term “shock” or “collapse” was combined with one of these keywords (all Oracle Text SQL queries are in Appendix 2) to retrieve the reference drug-induced anaphylaxis cases from the training set (n=178,676 patients with at least one hospital visit in 2011). Among the tested queries, “anaphylaxis” showed the highest recall, because it retrieved 17 of the 19 reference drug-induced anaphylaxis (0.89). It was also the keyword with the lowest precision (0.08) (Table 1). The queries “allergo-anesthesia,” “tryptase,” and “allergen-specific IgE” retrieved at least half of the reference cases (i.e., with a recall higher than 0.5).

On the basis of these results, an optimized query that included all individual queries with recall  $\geq 0.5$  was built by using the inclusive disjunction (“OR”), so that any fulfilled condition was sufficient to retrieve a relevant case. This optimized query had a recall value of 1 and a precision value of 0.07 (all 19 reference cases among the 270 patients retrieved by the query) (Table 1).

### **3.3. Optimized query performance in the test set**

The test set included 182,127 patients with at least one hospital visit in 2012. The optimized query identified 253 patients and 452 matching documents: 159 outpatient discharge summaries (35% of all documents), 110 biology results (24%), 73 inpatient discharge summaries (16%), 42 DRGs (9%), 25 discharge summaries from the emergency department (6%) and miscellaneous narrative documents (10%). The “anaphylaxis” and “allergo-anesthesia” queries (which correspond to signs and symptoms) mainly matched outpatient discharge summaries, whereas the “tryptase” and “allergen-specific IgE” queries (which correspond to biological results) matched laboratory results (Table 2).

### **3.4. Expert evaluation**

The two experts identified 59 cases of drug-induced anaphylaxis among the 253 patients retrieved from the test set using the optimized query: 38 cases of drug-induced anaphylaxis that occurred or required care at the CHU-RENNES, and 21 cases that required allergy investigations at the hospital (Table 3). The detailed results on each keyword performance can be found in Table 4.

Specifically, among the 253 patients, one expert identified 60 and the other one 56 cases of drug-related anaphylaxis. Ten disagreements between experts were resolved during consensus sessions. These disagreements included three cases that were validated by one of the experts, although the date of occurrence of the shock was not included in the study period; one case that was wrongly excluded because the date of occurrence was inaccurate, but within the study period; three cases that were included by one expert, although the name of the drug was not reported (only its pharmacological class); and three other cases that were not identified by one of the two experts when reading the records.

### **3.5. Comparison of data sources for the identification of anaphylaxis cases**

Analysis of the in French pharmacovigilance database for the year 2012 highlighted the presence of 17 drug-induced anaphylaxis cases at the CHU-RENNES and 10 cases that required care at the hospital (Table 3). Comparison of the data concerning the drug-induced anaphylaxis cases identified in the test set (eHOP) and in the French pharmacovigilance database showed that among the 41 cases that occurred at the hospital (test set + French pharmacovigilance database cases, see Table 3), 17 were already recorded in the French pharmacovigilance database (10 cases identified from DRGs and 7 spontaneously reported by physicians). Only two cases were shared by these two last data sources. Among these 17 cases, only three cases, which were spontaneously reported by physicians, were not found by querying the test set. One had a missing discharge summary, one was described as an “allergic reaction”, and the last one was only described using symptoms related to anaphylaxis. Twenty six cases were only found by querying the test set from eHOP.

Among the 25 cases that required allergy investigations at the hospital (Table 3), 10 were spontaneously reported by physicians. Four of these spontaneously reported cases were not detected via eHOP. These four cases occurred in the area covered by the Regional

Pharmacovigilance Center of Rennes, but allergy investigations were not conducted at the CHU-RENNES during the studied period. Fifteen cases were detected only in the test set.

### **3.6. Final tuning**

Besides drug-induced anaphylaxis, the main causes of anaphylaxis identified by experts were food and insect stings. Therefore, the following keywords were added to the optimized query as exclusion criteria: “flour,” “rye,” “wheat,” “venom,” “wasp,” “hymenopteran,” “stings,” “bee,” “bumble bee” and “insect”. The other source of false positive cases was the mention of “shock” or “collapse” in the documents describing the patients’ medical history. Consequently, another exclusion criterion was added to exclude patients for whom only the keywords “shock” or “collapse” appeared in their medical history (see Appendix 3, containing the complete Oracle text SQL query). The new optimized query that included also the exclusion criteria retrieved 200 potential cases from the test set (2012), compared to the 253 cases obtained by using the first version of the optimized query. This reduced group still contained all 59 drug-related anaphylaxis cases, thus resulting in a final precision of 0.29. The same process using the training set (2011) still retrieved the 19 reference drug-related anaphylaxis cases among 202 potential cases, compared with 270 cases retrieved by the original query.

Finally, based on the dates on the retrieved documents, it was calculated that the median number of potential cases per week was three (IQR: [2-5]), with a median number of validated drug-induce anaphylaxis cases of one every two weeks (IQR: [0-3]).

## **4. Discussion**

### **4.1. Comparison of data sources for the identification of anaphylaxis cases**

Our study demonstrates the added value of using information technologies, such as CDWs to improve the current practice, specifically in terms of identifying ADRs. The methods currently used, such as spontaneous reporting and DRG queries, do not detect all relevant ADRs and thus lead to underestimating drug safety issues. As suggested by some authors, it is extremely valuable to use additional data sources that allow better ADR detection and contribute to improving drug safety for patients <sup>7</sup>.

We do not intend to offer a method that exhaustively detects every case of drug-induced anaphylaxis. Producing reliable estimates for missing cases is undoubtedly a complex and time-consuming task because it would require reviewing all hospital visits during a given period to confirm both the occurrence of the anaphylaxis episode and the causal link with the administration of a drug. Yet, we do believe that spontaneous reporting, DRG queries and CDW (eHOP in our study) queries are complementary and can be integrated into a global strategy for the systematic detection of ADRs.

#### **4.2. Evaluation of the query on the test set**

If we compare the performance of three ADR detection methods on the test set, all cases spontaneously reported by physicians were valid drug-related anaphylaxis cases. In addition, DRG reporting allowed the validation of 10 of the 11 (90.9%) potential cases of anaphylactic shock (ICD-10 billing codes T88.2 and T88.6). In contrast, only 59 cases were validated among the 200 anaphylaxis cases (29.5%) identified in the test set (from eHOP) using the improved query with exclusion criteria (38 cases that occurred or required care at the CHU-RENNES [19%], and 21 cases that required allergy investigations at the hospital [10.5%]). The eHOP query is less specific than spontaneous reporting and DRG queries, although its sensitivity is far higher, thus allowing the detection of a larger number of relevant cases (41 of the 59 cases identified via the eHOP query were new). Concerning the drug-related anaphylaxis cases that occurred or required care at the CHU-RENNES, the proposed eHOP query detected all the cases identified through the DRG query, four of the seven spontaneously reported cases, and 26 not previously known cases. The eHOP query also detected several cases that did not occur at the CHU-RENNES, but that only required allergy investigations (including outpatients). This kind of case cannot be identified via DRG queries because DRGs are only produced for inpatients. In addition, our method retrieved patients mainly from outpatient discharge summaries or laboratory results that are inaccessible to the DRG query.

Comparison of the keyword recall values in the training (2011) and test (2012) sets showed that the keywords “anaphylaxis” and “allergen-specific IgE” allowed the detection of more relevant cases. Conversely, “allergo-anesthesia” and “tryptase” displayed a lower recall value in the test set than in the training set. This could lead to false positive results when using these two keywords, which can be explained by the fact that consultations for anesthesia-related

allergy problems and tryptase tests are only performed after the identification and investigation of drug-induced anaphylaxis (i.e., the type of cases that were identified in the training set). Thus, these keywords are not as pertinent for the identification of unknown cases that have not been spontaneously declared or identified via DRG. These observations highlight the complementarity of these three methods to improve the use of the available data sources for ADR detection.

### **4.3. Added value of the eHOP CDW**

One of the strengths of the eHOP technology is the possibility to query both unstructured and structured data. For instance, 56.9% of all retrieved documents were unstructured: discharge summaries from inpatients, outpatients or emergency departments (Table 2). Structured data (DRG codes and laboratory results) constituted only 33.6% of all retrieved documents, and most information found in EHRs is recorded as free text. More specifically, the search terms “tryptase” and “allergen-specific IgE” may improve the thoroughness of anaphylaxis detection. Most institutions currently use the standard i2b2 platform as their CDW technology, although its main purpose is the integration of structured data. Therefore, when only structured data are obtainable (or full-text information retrieval systems are not available), querying tryptase laboratory tests or allergen-specific IgE assays from i2b2, or even directly from laboratory information systems, can help to improve the detection of drug-induced anaphylaxis cases.

We consider that it is crucial to have the appropriate tools to adequately leverage the richness of the different data sources. Therefore, the eHOP CDW is currently deployed in the six main academic hospitals of western France. This could lead to the implementation of our query method by other pharmacovigilance centers, thus allowing the assessment of its potential and also an easier ADR detection at a larger scale.

Another advantage of eHOP over other CDW technologies is its easier access to EHR data that could facilitate the investigation process by pharmacovigilance experts. We found that if the eHOP query was used in routine practice by pharmacovigilance experts, the average number of potential cases to be investigated would be three per week (IQR: [2-5]). This method would result in one valid case detected every two weeks (IQR: [0-3]). This additional workload would be viable and could be handled by pharmacovigilance experts as part of their

practice. In addition, eHOP provides a user interface for navigating through EHRs, thus making the investigation process even easier.

#### **4.4. Perspectives**

Several options can be further explored to improve the proposed method. For instance, natural language processing (NLP) could increase the detection accuracy, as the presence of the keyword “anaphylaxis” in the patients’ medical history yielded several false positive cases. Most importantly, the available amount of health data and their heterogeneity require using machine learning and text mining approaches. Our study demonstrates that simple information retrieval methods are very efficient when the concepts to be retrieved can be described with relatively specific keywords. This is particularly true in pharmacovigilance, where there is still space for significantly improving ADR detection rate and accuracy. Our method could be applied to other ADR type. However, many diseases have complex characteristics and several etiologies, besides ADRs. In such cases, machine learning approaches could help to detect hidden or latent characteristics that are specific to complex ADRs.

#### **5. Conclusion**

Pharmacovigilance is crucial for the efficient long-term management of drug safety. This requires the development of suitable tools. Here, we described an information retrieval-based method for the detection of drug-induced anaphylaxis, based on querying both structured and unstructured data from a CDW. Besides the 25 cases already known from spontaneous and DRG reporting for 2012, with this method we could identify 41 additional cases. Our method can be easily implemented in the routine practice and could be proposed to other regional pharmacovigilance centers to better identify well-defined ADRs. Additional improvements may be necessary for the detection of more complex ADRs, possibly by using NLP processing, as well as machine learning and text mining methods.

#### **Acknowledgements**

We would like to thank the French National Agency for Medicines and Health Products Safety (ANSM), for funding this work inside the Breizh project (Evaluation of under-



reporting of adverse drug reactions in public hospitals in Brittany: the contribution of PMSI and a biomedical data warehouse) (grant no. AAP-2014-014).

## References

1. Hazell L, Shakir SAW. Under-Reporting of Adverse Drug Reactions. *Drug Saf.* 2012;29(5):385-396.
2. Jorup-Rönström C, Keisu M, Wiholm B-E. Could Swedish 'Yellow Cards' Be Substituted by E-Coded Summaries? *Drug Saf.* 2012;5(1):72-77.
3. Wodtke JM, Generali JA. Use of medical record codes to identify adverse drug reactions. *Am J Hosp Pharm.* 1993;50(9):1915-1916.
4. Cox AR, Anton C, Goh CHF, Easter M, Langford NJ, Ferner RE. Adverse drug reactions in patients admitted to hospital identified by discharge ICD-10 codes and by spontaneous reports. *Br J Clin Pharmacol.* 2001;52(3):337-339.
5. Osmont M-N, Cuggia M, Polard E, Riou C, Balusson F, Oger E. Use of the PMSI for the detection of adverse drug reactions. *Thérapie.* 2013;68(4):285-295.
6. Vial T. French pharmacovigilance: Missions, organization and perspectives. *Thérapie.* 2016;71(2):143-150.
7. Asfari H, Bousquet C, Trombert Paviot B, et al. Drug-related anaphylactic shocks: under-reporting and PMSI. *Thérapie.* 2014;69(6):483-490.
8. Coloma PM, Trifirò G, Schuemie MJ, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiol Drug Saf.* 2012;21(6):611-621.
9. Trifirò G, Sultana J, Bate A. From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources. *Drug Saf.* August 2017.
10. Black C, Tagiyeva-Milne N, Helms P, Moir D. Pharmacovigilance in children: detecting adverse drug reactions in routine electronic healthcare records. A systematic review. *Br J Clin Pharmacol.* 2015;80(4):844-854.
11. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. *AMIA Annu Symp Proc.* 2009;2009:391-395.
12. Cuggia M, Garcelon N, Campillo-Gimenez B, et al. Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform.* 2011;169:584-588.
13. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc.* 2010;17(2):124-130.
14. Osmont M-N, Campillo-Gimenez B, Metayer L, et al. Perianesthetic Anaphylactic Shocks: Contribution of a Clinical Data Warehouse. *Thérapie.* October 2015.
15. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. *Stud Health Technol Inform.* 2015;210:702-706.
16. Cuggia M, Garcelon N, Campillo-Gimenez B, et al. Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform.* 2011;169:584-588.

17. Montañez MI, Mayorga C, Bogas G, et al. Epidemiology, Mechanisms, and Diagnosis of Drug-Induced Anaphylaxis. *Front Immunol.* 2017;8:614.
18. Dong SW, Mertes PM, Petitpain N, Hasdenteufel F, Malinovsky JM, GERAP. Hypersensitivity reactions during anesthesia. Results from the ninth French survey (2005-2007). *Minerva Anesthesiol.* 2012;78(8):868-878.

## Figures legends

**Figure 1. Study design to build the optimized query used to identify drug-related anaphylaxis cases**

## Tables

**Table 1. Query results using the training set (2011)**

Id	Query	No. of retrieved cases out of 19	No. of returned patients out of 178,676	Precision	Recall	F-measure
1	anaphylaxis	17	226	0.08	<b>0.89</b>	0.15
2	allergo-anesthesia	13	44	0.30	<b>0.68</b>	0.42
3	tryptase	10	38	0.26	<b>0.53</b>	0.22
4	allergen-specific IgE	10	87	0.11	<b>0.53</b>	0.18
5	prick-testing	7	28	0.25	0.37	0.30
6	intradermal allergy testing	6	100	0.06	0.32	0.10
7	histamine	4	9	0.44	0.21	0.28
8	contraindication	4	6	0.67	0.21	0.32
9	immunoallergic	2	13	0.15	0.11	0.13
10	RAST-inhibition	1	2	0.50	0.05	0.09
	Optimized query (1 OR 2 OR 3 OR 4)	19	270	0.07	1.00	0.13

**Table 2. Summary of the types of retrieved documents (test set, 2012)**

Query	Overall		Outpatient discharge summary		Inpatient discharge summary		DRGs		Discharge summary from emergency department		Laboratory results		Other	
	N	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
anaphylaxis	343	100	<b>149</b>	<b>43.4</b>	73	21.3	42	12.3	25	7.3	11	3.2	42	12.5
allergo-anesthesia	38	100	<b>36</b>	<b>94.8</b>	1	2.6	0	0	0	0	0	0	1	2.6
tryptase	65	100	19	29.2	2	3.1	0	0	0	0	<b>42</b>	<b>64.6</b>	2	3.1

allergen - specific IgE	12 7	100	33	25.9	3	2.4	0	0	2	1.6	<b>88</b>	<b>69.3</b>	1	0.8
Optimi zed query	45 2	100	<b>159</b>	<b>35.2</b>	73	16.2	42	9.3	25	5.5	110	24.3	43	9.5

Table 3. Comparison of data sources for the identification of drug-induced anaphylaxis cases (2012)

Source	Number of cases that occurred at CHU-RENNES	Number of cases that required allergy investigation at CHU-RENNES	Total number of cases
<b>eHOP CDW (test set)</b>			
<i>total</i>	38	21	59
<i>only detected in eHOP</i>	<b>26</b>	<b>15</b>	<b>41</b>
<i>not detected in eHOP</i>	3	4	7
<b>DRGs</b>			
<i>total</i>	10	n/a	10
<i>only detected with DRGs</i>	0	n/a	0
<i>not detected with DRGs</i>	31	n/a	31
<b>Spontaneous reports</b>			
<i>total</i>	7	10	17
<i>only detected by spontaneous reports</i>	3	4	7
<i>not detected by spontaneous reports</i>	34	15	59
<b>Total*</b>	41	25	66

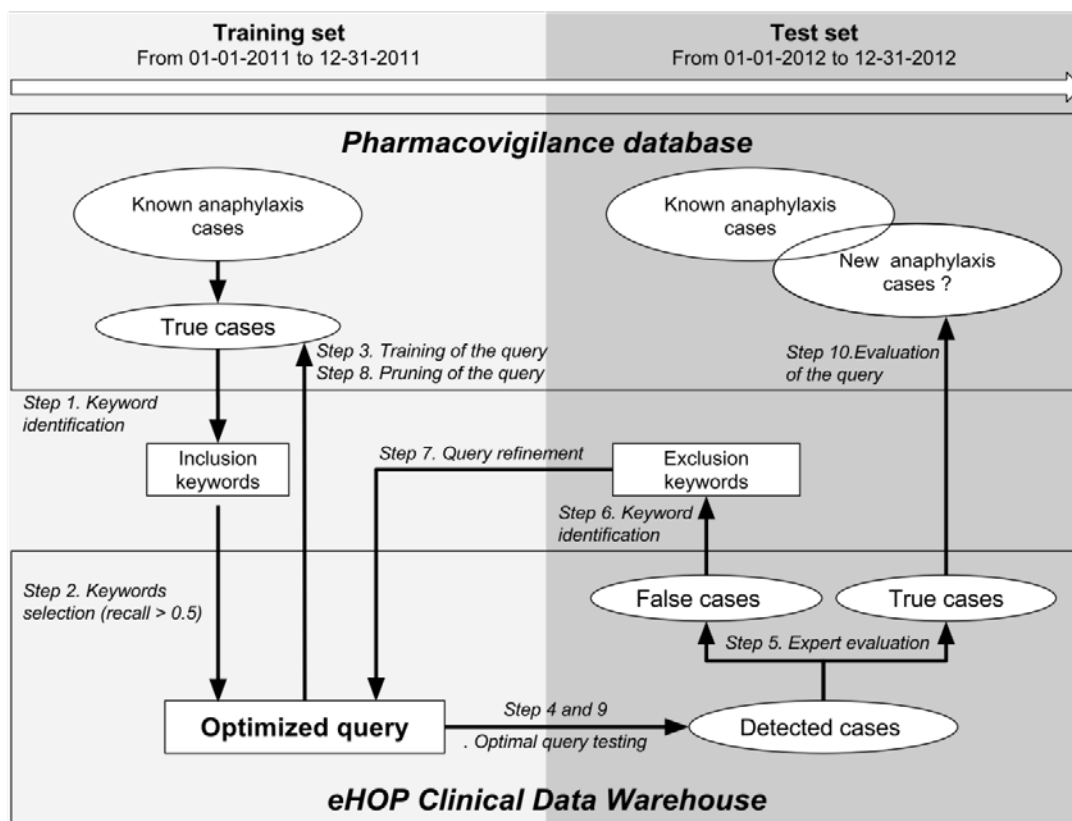
\*Total number of individual cases of drug-related anaphylaxis identified using the three sources.

Table 4. Optimized query evaluation using the test set (2012)

Id	Query	No. of retrieved cases out of 59	No. of returned patients out of 182,127	Precision	Recall	F-measure
1	anaphylaxis	44	190	0.23	0.74	0.15
2	allergo-anesthesia	11	35	0.31	0.19	0.42
3	tryptase	22	56	0.39	0.37	0.22
4	allergen-specific IgE	42	114	0.37	0.71	0.18
	Optimized query (1 OR 2 OR 3 OR 4)	59	253	0.23	1	0.13

## Figures

Figure 1



## Appendices

### Appendix 1: List of ICD-10 codes used by the pharmacovigilance center of Rennes

ICD-10 billing code	Description
G21.0	Malignant neuroleptic syndrome
G62.0	Drug-induced polyneuropathy
J70.2	Acute drug-induced interstitial lung disorders
K71.0	Toxic liver disease with cholestasis
K71.1	Toxic liver disease with hepatic necrosis
K71.2	Toxic liver disease with acute hepatitis
K71.6	Toxic liver disease with hepatitis, not elsewhere classified
K71.8	Toxic liver disease with other disorders of liver
K71.9	Toxic liver disease, unspecified
L27.0	Generalized skin eruption due to drugs and medicaments
L51.0	Nonbullous erythema multiforme
L51.1	Bullous erythema multiforme
L51.2	Toxic epidermal necrolysis [Lyell]
N14.1	Nephropathy induced by other drugs, medicaments and biological

	substances
N14.2	Nephropathy induced by unspecified drug, medicament or biological substance
N17.0	Acute renal failure with tubular necrosis
N17.1	Acute renal failure with acute cortical necrosis
N17.8	Other acute renal failure
N17.9	Acute renal failure, unspecified
T88.2	Shock due to anesthesia
T88.6	Anaphylactic shock due to adverse effect of correct drug or medicament properly administered

#### Appendix 2: Oracle text queries based on keywords identified by experts

Id	Query name	Oracle text query expression
1	anaphylaxis	(shock or collapse) and anaphy%
2	allergo-anesthesia	(shock or collapse) and allergo-anesth%
3	tryptase	(shock or collapse) and (NEAR((level, tryptase), 2, TRUE) OR NEAR((increase%, tryptase), 2, FALSE))
4	allergen-specific IgE	(shock or collapse) and allergen-specific IgE%
5	prick-testing	(shock or collapse) and prick-test
6	Intradermal allergy testing	(shock or collapse) and intradermo or intra-dermo
7	histamine	(shock or collapse) and NEAR ((increase%, histamine), 2, FALSE)
8	contraindication	(shock or collapse) and contraindication for life
9	immunoallergic	(shock or collapse) and immuno-aller%
10	RAST inhibition	(shock or collapse) and RAST inhibition
	Optimized query	(shock or collapse) and (anaphy% OR NEAR((level, tryptase), 2, TRUE) OR NEAR((increase%, tryptase), 2, FALSE) OR allergo-anesth% OR allergen-specific IgE%)

#### Appendix 3: Optimized query with exclusion criteria: complete Oracle text SQL query

(shock or collapse) and (anaphy% OR NEAR((level, tryptase), 2, TRUE) OR NEAR((increase%, tryptase), 2, FALSE) OR allergo-anesth% OR allergen-specific IgE%) not (flour or rye or wheat or venom% or wasp% or hymenopter% or sting% or bee% or bumble bee% or insect% or near((medical history%, shock% or collapse), 50, true))