An Automated Detection System of Drug-Drug Interactions from Electronic Patient Records Using Big Data Analytics

Guillaume Bouzillé, Camille Morival, Richard Westerlynck, Pierre Lemordant, Emmanuel Chazard, Pascal Lecorre, Yann Busnel, Marc Cuggia

MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190180

An Automated Detection System of Drug-Drug Interactions from Electronic Patient Records Using Big Data Analytics

Guillaume Bouzillé[a], Camille Morival[b], Richard Westerlynck[c], Pierre Lemordant[a], Emmanuel Chazard[d], Pascal Lecorre[b], Yann Busnel[c], Marc Cuggia[a]

[a] Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France,
[b] Laboratoire de Pharmacie Galénique, Biopharmacie et Pharmacie Clinique, IRSET U1085, Faculté de Pharmacie, Université de Rennes 1, F-35043, Rennes Cedex, France,
[c] IMT Atlantique, F-35576, Cesson-Sévigné, France,
[d] Univ Lille, CHU Lille, CERIM EA2694, F-59000 Lille, France

Abstract

The aim of the study was to build a proof-of-concept demonstratrating that big data technology could improve drug safety monitoring in a hospital and could help pharmacovigilance professionals to make data-driven targeted hypotheses on adverse drug events (ADEs) due to drug-drug interactions (DDI). We developed a DDI automatic detection system based on treatment data and laboratory tests from the electronic health records stored in the clinical data warehouse of Rennes academic hospital. We also used OrientDb, a graph database to store informations from five drug knowledge databases and Spark to perform analysis of potential interactions betweens drugs taken by hospitalized patients. Then, we developed a machine learning model to identify the patients in whom an ADE might have occurred because of a DDI. The DDI detection system worked efficiently and computation time was manageable. The system could be routinely employed for monitoring.

Keywords:

Computing Methodologies, Drug Interaction, Machine Learning.

Introduction

Drug-drug interactions (DDIs) are a critical issue in patient care because they can lead to adverse events and ultimately increase care costs and patient mortality. Therefore, these events must be identified and prevented as early as possible [1]. However, many new drugs are released each year, and therefore, it is very difficult for healthcare professionals to be informed and to consider all DDIs. Moreover, the alarm functionalities of drug computerized physician order entry (CPOE) systems are frequently not used because they do not focus on clinically relevant DDIs and lead users to alarm fatigue. Although focused on specific interactions, studies on DDI prevalence show the existence of risks for polymedicated patients and highlight the importance of pharmacovigilance programmes [2,3].

With the unprecedented development of digital health and hospital clinical data warehouses (CDW), data produced during the healthcare process are now easily reusable [4]. Electronic health records (EHR) contain real-time information on drug prescription/regimens during hospitalization as well as all clinical information. Such data could be analysed to estimate DDI prevalence, to facilitate health professionals' practice assessment and to detect the occurrence of DDI-linked adverse drug events (ADE). In France, pharmacovigilance currently relies mainly on the spontaneous reporting by physicians or/and detection of diagnoses that could be related to ADE from the hospital billing system (diagnosis related group, DRG, database). New data sources, such as national claim databases, are also leveraged to improve DDI and ADE detection [5,6]. EHR data-mining also could help pharmacovigilance professionals to improve drug safety assessment.

All these health-related databases fit perfectly with the big data paradigm because they contain voluminous, highly complex and heterogeneous information that is produced in real time [7]. In the last few years, many big data technologies have been developed. However, their implementation in a hospital information system for processing healthcare big data in real-world condition of use is still largely uncharted.

Here, we describe a method, which propose to use big data technology to improve drug safety monitoring in a hospital and could help pharmacovigilance professionals to make data-driven targeted hypothesis on ADEs.

Methods

Figure 1 presents the overall approach of the study and the big data technologies used in each step.

Patient Data

We used the Rennes academic hospital EHRs that are stored in a CDW called eHOP (entrepot HOPital). This CDW includes both structured data (e.g., laboratory results, drug prescriptions and regimens) and unstructured data (e.g., operative reports, discharge summaries), and is dedicated to data reuse for clinical research [8]. The eHOP's star schema architecture and graphic user interface allows researchers, even without any database language knowledge, to quickly access and efficiently search information within millions of patient records.

For this study, we used information about drug administrations (used drug(s) and regimens) and laboratory results (date, nature of the test and results: normal, abnormally high, or abnormally low).

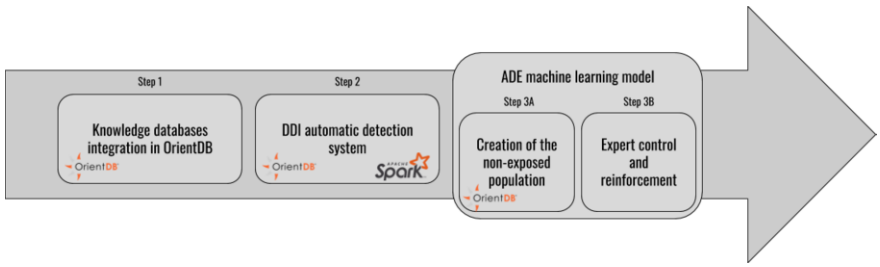

Figure 1 - Overall approach of the study

Knowledge Databases Integration (step 1)

To identify and collect information on potential DDIs, and also to compare information from different sources, we selected five drug knowledge databases: Thesaurus, Vidal, Theriaque, Micromedex and Drugs.com [9–13]. These databases are commonly used by health professionals, but are not specifically targeted to DDI detection. They are available via a web application programming interface (API) that requires a specific procedure because each database stores data with its own structure. To avoid this, we extracted the relevant information from these databases and stored it in OrientDB, a graph-oriented model database [14] that fits well with our objective because a DDI can be modelled as an edge between two drugs. Thus, once the information is stored in a single OrientDB database, no more computation is required to access such information.

DDI Automatic Detection System from Patient Records (Step 2)

For DDI identification, we collected drug data from the patient EHRs stored in eHOP and computed the active interval (i.e., the period during which a drug was effective) for all drugs taken by a patient during the hospital stay. If two active intervals overlapped (fully or partially), then analysis of the data collected in the OrientDB database allowed determining whether the two drugs interacted. In this case, the potential DDI event was stored in eHOP. As these are independent processes (each drug pair is checked independently), the Spark cluster-computing framework was used to perform distributed computing [15,16]. As all the potential DDI events can be stored in eHOP, then we could compute the prevalence of a DDI for any specific drug, molecule, or population.

Creation of a Machine Learning Model (Step 3a)

The data stored in the CDW eHOP do not allow direct confirmation of whether a patient reported a DDI-linked ADE or not. Indeed, this needs to be validated by the pharmacovigilance experts who do not have the proper means to check all the patient records. Therefore, we wanted to create a system to report to drug safety professionals only the most interesting cases among all DDIs detected by the DDI automatic detection system (i.e., patients in whom an ADE might have occurred because of a DDI).

We assume that laboratory results will change if an ADE occurs. So, we can train a machine learning model with two populations: those who experienced an ADE and those who did not. Unfortunately, we cannot identify manually who experienced an ADE. For this reason, we performed one of the research design presented by Hennessy et al. [17]: we choose to compare the population exposed to a DDI with another population non-exposed to this DDI and who did not experience an ADE, by design. There are likely many patients who do not experience an ADE in the exposed population, but the model will present only the most suspected cases and this problem will be solved with the gradual feedback of drug safety professionals: the system will adjust weights of patients in the model, giving a greater weight to the well-predicted patients.

We developed an artificial neural network system that allows us to predict an output. This system has a single hidden layer and the number of perceptrons was decided during cross-validation. Our machine learning model works in two phases. First, it uses all data available for patients who experienced a specific DDI and those who did not (exposed and non-exposed populations) to classify them as having reported an ADE or not. Then, the model is reinforced with information coming from drug safety professionals who infirm or confirm the previous classification (Fig. 2).

We then had to form the non-exposed population. Within a DDI, we called "Object" the drug under study, and "Precipitant" the other drug. Moreover, we called "Control-precipitant" any drug that has the same therapeutic use as the Precipitant, but that does not interact with the Object. For a given Object, we compared the exposed population, found with the DDI automatic detection system, to the non-exposed population. The non-exposed population included all patients, who were not in the exposed population and who had an overlap (fully or partial) between the action interval of the Object and of the Control-precipitant. We created this non- exposed population using the same process as for the exposed population.

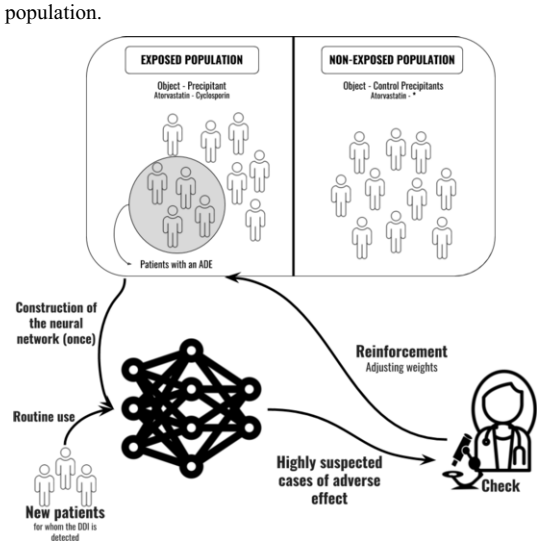

Figure 2 - Creation and use of the artifical neural network

Data processing was performed with Java 8, Spark 2.10 and OrientDB 2.2.4 on Intel(R) Xeon(R) CPU E5-2609 1,90GHz computer with 32,0 Go of RAM.

Big Data Technologies: Convenient Tools for Complex Data Processing

Here, we proposed a complete automated data treatment system, from the collection of heterogeneous data to their enhancement in a machine learning model. This system can monitor DDI prevalence and try to identify patients with a possible DDI-linked ADE, without the intervention of drug safety professionals. To achieve this, we used several convenient tools:

OrientDB is an easy-to-use tool to store pre-computed data. The OrientDB database model includes two main classes: vertices and edges that connect two vertices. In our study, the "vertex" interface represented the class "Drug" and included drug name, ID-code and half-life. The "edge" interface represented the class "Interactions" and included DDI severity level. We also specified from which drug database the information on the DDI came. Thus, via OrientDB, each drug knowledge database can be interrogated separately. The "edge" interface is also used to represent the class "Control-precipitant".

Figure 3 presents the database model through an example: Drug1 has an interaction with Drug3 according two different databases (two edges of class "Interaction"). Let consider the Object-Precipitant couple Drug1-Drug3, then Drug4 is a control-precipitant of Drug3 (one oriented edge of class "Control-precipitant"). An example of query would be: "give all the drugs that have an interaction with Drug1 according Micromedex and where the severity level is 1".

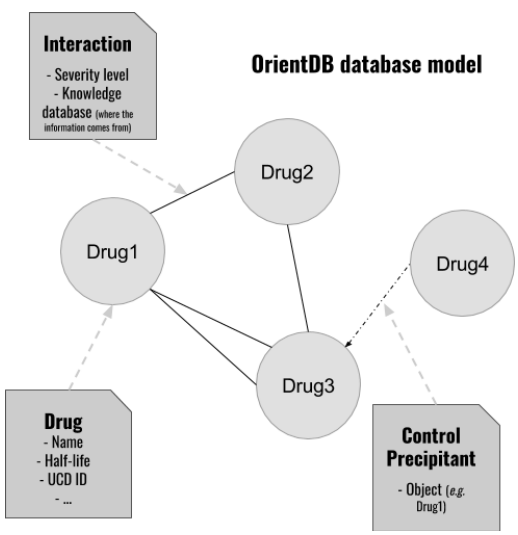

Figure 3 - OrientDb database model

The query language was very close to the structured query language (SQL) and allowed searching a vertex that walks along edges to another vertex, according to the chosen conditions. Data uploading is fast and based on a convenient Java Graph API. We manipulated a graph Java object that is automatically committed at the end of the process. Moreover, if access to a part of the graph is required (e.g., all the drugs that interact with pravastatin according to a severity level of 2), we used this object as a temporary store before processing.

Switching between different knowledge databases, stored in the same OrientDB database, involves only a variable on an edge. Ultimately, the little amount of time spent for the pre-calculation facilitates the storage and the access to multiple data sources. Only one kind of request is needed for all five databases. We could easily add information from other data sources (for example, composition of a drug and half-life of the active substances), or more precise information about DDI-linked ADEs (such as the relevant laboratory tests). Nevertheless, this task demands a manual work for each group of drugs [18].

Spark allows parallel processing easily. As many processes are independent from each other, their parallel treatment with Spark leads to a big time saving[19].

Evaluation (Step 3B)

To evaluate our DDI detection system (step 2 in Fig 1), we focused on a class of drugs called statins that are prescribed (long-term treatment) to patients with cardiovascular diseases, and particularly to elderly patients who are usually polymedicated and consequently prone to DDIs. We selected the study population (i.e., all patients taking statins) from all patients included in eHOP from January, 1 2015 to July, 8 2016. It included 10,506 hospitalized patients with a median hospitalization of 7 days, and a median age of 72 years (range: 19 to 98 years).

We defined statins as the "Object" and all the drugs that interact with them were considered as candidate "Precipitants". We selected as Control-precipitants (symbolized by * in fig 3) all the drugs that are in the same fifth level (i.e., chemical substance) as the Precipitant in the Anatomical Therapeutic Chemical (ATC) classification [20], but do not interact with the Object. Thus, Control-precipitants have the same (or a similar) therapeutic usage as the Precipitant. We stored all these data in OrientDB because each DDI is a link (i.e., edge) between drugs (i.e., two vertices).

Concerning the action intervals, we chose a period of seven half-lives for each statin molecule and arbitrarily selected one day for the Precipitant, because this information could not always be extracted automatically from the five drug knowledge databases.

To determine how well the machine learning model can identify patients who may have a DDI- linked ADE (step 3B in Fig 1), we evaluated the model prediction error using the out-of-bag (OOB) error method: several models are built with a bootstrapped dataset, the OOB error is the mean of the errors computed with non-used data in each model.

The neural network gives the probability to belong to a class. We used cross-validatin resampling to optimize the threshold separating the two class. We chose to study a specific DDI in which atorvastatin was the Object and cyclosporine the Precipitant (i.e., exposed population). The non-exposed population consisted of patients who took atorvastatin and a Control- precipitant (Fig 3). The used variables were: demographic data, pathologies (ICD-10 codes) and laboratory test results. We used all the laboratory test results included between the beginning of the event and 3 days later. If a laboratory test appeared more than once, we took the mode of the results.

The reinforcement phase was not evaluated because it is currently under construction in collaboration with drug safety specialists.

Results

DDI identification with the automatic detection system was very fast due to the use of a graph- oriented model. For instance, for the simple query "is there an interaction between these two drugs?", or the more complex query "select all drugs that interact with this specific drug", the OrientDB database was always faster (less than 20ms) than the Theriaque SQL database (several seconds). Moreover, switching to another drug knowledge database was very easy with OrientDB because it only needed to change a condition in the query (which database = 'Theriaque').

The time required to create these graph databases was reasonable: for instance, the information coming from the Theriaque database, which is equivalent to 18,800 vertices and 23 million edges, was integrated in one hour. Afterwards, data access was immediate.

Once the OrientDB database was ready, from the eHOP CDW, we checked the DDI occurrence for all drug couples in the study population. To this aim, we computed all the fully or partially overlapping action intervals for all drug couples involving a statin. For each patient, we visualized all the detected DDIs: between 22.5% and 52.2% (depending on the drug knowledge database) of the 10,506 patients who were taking statins presented at least one DDI involving a statin.

Computation time was reduced with the use of the Spark framework: the processing time of 800,000 rows of patient records decreased from 60 minutes initially to only 12 minutes with Spark.

To test the ADE prediction performance of the machine learning model, we then focused only on one specific DDI (atorvastatin-cyclosporine) to create the training sample. We could identify 102 patients with atorvastatin-cyclosporine DDIs (i.e., the exposed population) and 150 patients without this DDI (i.e., the non-exposed population) (Table 1).

Table 1- Demographic data of the exposed and non-exposed population samples

	Exposed population (n=102)	**Non-exposed population (n=150)**
Age (mean ± Sd)	72.1 ± 11.6	72.9 ± 10.9
Sex (% of men)	79.8	83.5
Cardiac pathology (%)	38.2	37.8

For the optimal threshold, the neural network out-of-bag error was 17.06%, sensitivity and specificity were 90.20% and 78% respectively, and the AUC was 0.757. The processing time was short (less than 30 seconds) and could be easily performed again during the reinforcement phase.

Discussion

DDI Automatic Detection System: A New Source of Refined Data for Drug Safety Professionals

With this DDI detection system and the CDW, we can compute the overall DDI prevalence for any drug pairs, and also according to a chosen interaction severity level, or for a specific population subset. These data are useful for drug safety monitoring/research and have been already used in a study on the use of statins [21-22]. Moreover, currently, pharmacovigilance studies use different case report databases [23]. We find DDIs directly in the patient EHRs. Therefore, after DDI detection, we can link this information to other data included in the EHR (e.g., demographic data, laboratory test, etc.) to contextualize the case.

However, our DDI detection system cannot identify all DDIs. This could be due to several reasons. First, the choice of the drug knowledge database is important, and we actually observed heterogeneity between these databases that might lead to variability in DDI detection [22]. Moreover, with more information concerning the changes in the blood concentration (and half-life) of the involved drugs, we could compute more precise action intervals, thus improving the identification of overlapping treatment periods. However, this would require extensive manual search of literature data. Finally, our system cannot detect a DDI caused by a drug prescribed/administered outside the hospital. For instance, the regular treatment is usually stopped when a patient is hospitalized in the emergency service and is recorded in the emergency report. Accessing this information requires a specific treatment of unstructured text. Another option could be to link data on the drugs prescribed in primary care settings (i.e., the national health insurance database) to the hospital data (e.g., eHOP). Despite the linkage problems and the issues due to the national health insurance database features (data only on refundable drugs and only on the drug purchase but not the regimen), the analysis of the entire patient path could bring useful information on treatment ruptures, which could suggest DDIs.

A machine learning model for search reinforcement

The automatic way used to create the non-exposed population works and selects a population similar to the exposed group in terms of demographics and pathology. If the sample is big enough, we can ask the system to select the most similar patients.

Although the study of the temporal correlations between laboratory test changes and drug administration is relevant for ADE detection [24,25], we chose a robust prediction-oriented machine learning model that can work without requiring too many adjustments. Indeed, we expect that clinical variables in the exposed population will change in the presence of a DDI. However, we do not know whether the detection of a DDI implies automatically an ADE, and accessing the information to confirm the ADE involves a considerable work for drug safety professionals that we want to avoid. Therefore, to automate the monitoring of DDI-linked ADEs, we took the data immediately available from eHOP.

As they have very similar demographic characteristics, comparing exposed and non-exposed populations seemed to be an effective way to initialize the system. An improvement would be to take into account also the information included, for example, in ADE report databases. However, this system can be easily improved even without more data. Indeed, the model predicts candidate ADE cases that are likely to have been caused by DDIs and proposes them to drug safety professionals. If these cases are confirmed by drug safety professionals, they are included in the training sample to automatically enhance the model.

On the other hand, and like for any automatic detection model, our neural network model does not allow understanding which anomaly led to the prediction of an ADE and for this the analysis of the patient record is required. A machine learning model requires a lot of work, especially the choice of the model and the features engineering. In particular, a larger sample could ollow other resampling strategies to be used, that do not require the out of bag error, which is prone to overestimation of

the true prediction error [26]. These questions need a suitable study including a better evaluation with drug safety specialists.

Conclusions

This study shows how to employ healthcare data for automated DDI monitoring and ADE prediction. It involves the complete data processing chain: data collection, processing and enrichment as well as the creation of a machine learning model. The developed statistical model is the first step for a simple and convenient use of data, and could be enriched with additional information from other databases that must be integrated (more specific drug knowledge databases, ADE report databases …).

Although no drug safety professional is required during the monitoring, their expertise is essential to properly understand the data and put them into context. Their recommendations were also important to build the monitoring system and to improve the model.

Acknowledgements

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

References

[1] F. Meier, R. Maas, A. Sonst, A. Patapovas, F. Müller, B. Plank-Kiegele, B. Pfistermeister, O. Schöffski, T. Bürkle, H. Dormann, Adverse drug events in patients admitted to an emergency department: an analysis of direct costs, Pharmacoepidemiol. Drug Saf. 24 (2015) 176–186.

[2] C. Bonnet, P. Boudou-Rouquette, E. Azoulay-Rutman, O. Huillard, J.-L. Golmard, E. Carton, G. Noé, M. Vidal, G. Orvoen, et al., Potential drug-drug interactions with abiraterone in metastatic castration-resistant prostate cancer patients: a prevalence study in France, Cancer Chemother. Pharmacol. 79 (2017) 1051–1055.

[3] E. Ramirez, A.J. Carcas, A.M. Borobia, S.H. Lei, E. Piñana, S. Fudio, J. Frias, A Pharmacovigilance Program From Laboratory Signals for the Detection and Reporting of Serious Adverse Drug Reactions in Hospitalized Patients, Clin. Pharmacol. Ther. 87 (2010) 74–86.

[4] J. Price, What Can Big Data Offer the Pharmacovigilance of Orphan Drugs?, Clin. Ther. 38 (2016) 2533–2545.

[5] K. Martin-Latry, B. Bégaud, Pharmacoepidemiological research using French reimbursement databases: yes we can!, Pharmacoepidemiol. Drug Saf. 19 (2010) 256–265.

[6] M.-L. Yeh, Y.-J. Chang, S.-J. Yeh, L.-J. Huang, Y.-T. Yen, P.-Y. Wang, Y.-C. Li, C.-Y. Hsu, Potential drug–drug interactions in pediatric outpatient prescriptions for newborns and infants, Comput. Methods Programs Biomed. 113 (2014) 15–22.

[7] G. Bouzillé, R. Westerlynck, G. Defossez, D. Bouslimi, S. Bayat, C. Riou, Y. Busnel, C. Le Guillou, J.-M. Cauvin, C. Jacquelinet, P. Pladys, E. Oger, E. Stindel, P. Ingrand, G. Coatrieux, M. Cuggia, Sharing health big data for research - A design by use cases: the INSHARE platform approach, in: 16th World Congr. Med. Health Inform. MedInfo2017, Hangzhou, China, 2017.

[8] G. Bouzillé, E. Sylvestre, B. Campillo-Gimenez, E. Renault, T. Ledieu, D. Delamarre, M. Cuggia, An Integrated Workflow For Secondary Use of Patient Data for Clinical Research., Stud Health Technol Inf. 216 (2015) 913.

[9] ANSM: Agence nationale de sécurité du médicament et des produits de santé. http://ansm.sante.fr (accessed August 1, 2017).

[10] VIDAL - La base de données en ligne des prescripteurs libéraux. https://www.vidal.fr (accessed August 1, 2017).

[11] Thériaque. http://www.theriaque.org (accessed August 1, 2017).

[12] Micromedex. https://www.micromedexsolutions.com (accessed August 1, 2017).

[13] Drugs.com | Prescription Drug Information, Interactions & Side Effects, Drugs.Com. https://www.drugs.com.

[14] OrientDB - Distributed Graph/Document Multi-Model Database. httpss//orientdb.com (accessed August 1, 2017).

[15] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica, Spark: Cluster computing with working sets., HotCloud. 10 (2010) 95.

[16] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica, Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing, in: USENIX Association, 2012: pp. 2–2.

[17] S. Hennessy, C. Leonard, J. Gagne, J. Flory, X. Han, C. Brensinger, W. Bilker, Pharmacoepidemiologic Methods for Studying the Health Effects of Drug-Drug Interactions (DDIs), Clin. Pharmacol. Ther. 99 (2016) 92–100.

[18] A. Neubert, H. Dormann, H.-U. Prokosch, T. Bürkle, W. Rascher, R. Sojer, K. Brune, M. Criegee-Rieck, E-pharmacovigilance: development and implementation of a computable knowledge base to identify adverse drug reactions, Br. J. Clin. Pharmacol. 76 (2013) 69–77.

[19] A.G. Shoro, T.R. Soomro, Big data analysis: Apache spark perspective, Glob. J. Comput. Sci. Technol. 15 (2015).

[20] WHOCC - Structure and principles. https://www.whocc.no/atc/structure_and_principles/ (accessed August 1, 2017).

[21] J. Stausberg, International prevalence of adverse drug events in hospitals: an analysis of routine data from England, Germany, and the USA, BMC Health Serv. Res. 14 (2014) 125.

[22] C. Morival, R. Westerlynck, G. Bouzillé, M. Cuggia, P.L. Corre, Prevalence and nature of statin drug-drug interactions in a university hospital by electronic health record mining, Eur. J. Clin. Pharmacol. (2017) 1–10.

[23] F. Kaguelidou, F. Beau-Salinas, A.P. Jonville-Bera, E. Jacqz-Aigrain, Neonatal adverse drug reactions: an analysis of reports to the French pharmacovigilance database, Br. J. Clin. Pharmacol. 82 (2016) 1058–1068.

[24] A. Newe, S. Wimmer, A. Neubert, L. Becker, H.-U. Prokosch, M.W. Beckmann, R. Fietkau, C. Forster, M.F. Neurath, G. Schett, T. Ganslandt, Towards a Computable Data Corpus of Temporal Correlations between Drug Administration and Lab Value Changes, PLOS ONE. 10 (2015) e0136131.

[25] A. Newe, Dramatyping: a generic algorithm for detecting reasonable temporal correlations between drug administration and lab value alterations, PeerJ. 4 (2016) e1851.

[26] S. Janitza, and R. Hornung, On the overestimation of random forest's out-of-bag error, *PLoS One*. **13** (2018).

Address for correspondence

Bouzillé Guillaume: guillaume.bouzille@univ-rennes1.fr