Artificial Intelligence for Unstructured Healthcare Data: Application to Coding of Patient Reporting of Adverse Drug Reactions

Louis Létinier[1,2,3,*], Julien Jouganous[3], Mehdi Benkebil[4], Alicia Bel-Létoile[3], Clément Goehrs[3], Allison Singier[1], Franck Rouby[5,6], Clémence Lacroix[5,6], Ghada Miremont[1,2], Joëlle Micallef[5,6], Francesco Salvo[1,2] and Antoine Pariente[1,2]

Adverse drug reaction (ADR) reporting is a major component of drug safety monitoring; its input will, however, only be optimized if systems can manage to deal with its tremendous flow of information, based primarily on unstructured text fields. The aim of this study was to develop an automated system allowing to code ADRs from patient reports. Our system was based on a knowledge base about drugs, enriched by supervised machine learning (ML) models trained on patients reporting data. To train our models, we selected all cases of ADRs reported by patients to a French Pharmacovigilance Centre through a national web-portal between March 2017 and March 2019 (*n* = 2,058 reports). We tested both conventional ML models and deep-learning models. We performed an external validation using a dataset constituted of a random sample of ADRs reported to the Marseille Pharmacovigilance Centre over the same period (*n* = 187). Here, we show that regarding area under the curve (AUC) and F-measure, the best model to identify ADRs was gradient boosting trees (LGBM), with an AUC of 0.93 (0.92–0.94) and F-measure of 0.72 (0.68–0.75). This model was run for external validation showing an AUC of 0.91 and a F-measure of 0.58. We evaluated an artificial intelligence pipeline that was found able to learn how to identify correctly ADRs from unstructured data. This result allowed us to start a new study using more data to further improve our performance and offer a tool that is useful in practice to efficiently manage drug safety information.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The use of artificial intelligence (AI) to facilitate the management of adverse drug reaction (ADR) reports is a research area that has become a priority in recent years.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Which machine learning methods are the most efficient to respond to this topic?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ An AI pipeline using a knowledge database and gradient boosting trees that was found able to learn to correctly identify ADRs from patient reports with unstructured data.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The system presented here will be deployed nationally in France to strengthen the coronavirus disease 2019 (COVID-19) vaccination campaign pharmacovigilance.

Over the last decades, the amount of healthcare data available has dramatically increased. This phenomenon started with the arrival of electronic health records and the vast amounts of related clinical data.[1,2] More recently, health data have been enriched by patients themselves using connected tools or digital platforms. These data should be more and more numerous and in increasingly varied and often unstructured formats.[3,4] The corresponding challenge is to find novel pathways to manage heterogeneous and often unstructured healthcare information, such as texts or images. Artificial intelligence (AI) methods appear promising in this perspective.

[1]INSERM, BPH, U1219, Team Pharmacoepidemiology, Univ. Bordeaux, Bordeaux, France; [2]CHU de Bordeaux, Pole de Santé Publique, Service de Pharmacologie Médicale, Centre de Pharmacovigilance de Bordeaux, Bordeaux, France; [3]Synapse Medicine, Bordeaux, France; [4]Surveillance Division, Agence nationale de sécurité du médicament et des produits de santé (ANSM), Saint Denis, France; [5]CRPV Marseille Provence Corse, Service Hospitalo-Universitaire de Pharmacologie Clinique et Pharmacovigilance, Assistance Publique Hôpitaux de Marseille, Marseille, France; [6]Institut des Neurosciences des Systèmes, INSERM 1106, Aix Marseille Université, Marseille, France. *Correspondence: Louis Létinier (louis.letinier@u-bordeaux.fr)

Received December 18, 2020; accepted March 22, 2021. doi:10.1002/cpt.2266

No consensual definition of AI actually exists; the University of Cambridge defines it as "the study of how to produce machines that have some of the qualities that the human mind has, such as the ability to understand language, recognize pictures, solve problems, and learn."[5] Contextualizing AI to healthcare unstructured data, it could be conceived as the combination of semantic information (SI) and machine learning (ML) to perform data analysis. SI allows users to add meaning to data through ontologies and knowledge databases,[2] therefore mimicking the theoretical learning patterns of a human being, whereas ML provides potentiality for new assumptions by aggregating information compounded from existing knowledge.[6] Taken together, SI and ML bring machines one step closer to the way human beings manage and learn from information. The combination of SI and ML has already shown interesting results in the field of drug identification.[7]

Beyond medical imaging, all medical specialties could benefit from these technologies. This is especially true for pharmacovigilance, which could evolve considerably in the coming years. Once marketed, medicinal products need to be continuously monitored in real-world settings for their use, effectiveness, and safety, the latter activity corresponding to pharmacovigilance. Pharmacovigilance activities cover the detection, assessment, understanding, and prevention of adverse drug reactions (ADRs),[8] which constitute one of the main causes of death in developed countries[9] and represent a considerable economic burden with preventable ADRs costing up to $3.5 billion yearly in the United States.[10,11] Moreover, according to the literature, 5–6.5% of all hospital admissions could be related to ADRs.[12]

Drug safety evaluation requires monitoring and analyzing a huge amount of information, which can be dispersed in various sources, such as scientific articles, structured or unstructured health data, ADR reporting, or social networks. In many countries, when a healthcare professional or a patient wishes to report an ADR, she/he can fill an online reporting form including structured and unstructured data. These forms are later processed by pharmacovigilance professionals who thus have to cope with the management of large amounts of unstructured data. Specifically, the identification of the reported ADRs and the evaluation of their seriousness can be very time-consuming from the free-text entered in these report forms. This raises two important concerns. First the importance of the resources it mobilizes could be detrimental to other pharmacovigilance activities. Second, even essential, the standardization of this process appears unlikely among different teams and individuals, with a risk of heterogeneously coding.[13] Both these aspects are deleterious in terms of public health by limiting the ability to detect new safety signals. The current context of pharmacovigilance because patients can report ADRs themselves is detailed in **Supplemental Information 1 Text file S1**. The development of tools allowing to automatically process ADRs is currently the object of intensive research and efforts, both from public and private entities.[14] When originated from industries, these developments appear to be mostly restricted to their sole products.[15]

The underlying assumption of the present work is that a global knowledge database on drugs, enriched by supervised ML models trained on reporting data, could allow setting up a universal tool for preprocessing free text reported ADRs. This tool would increase efficiency in dealing with such information and improve capacities in drug surveillance.

AIMS

We aimed to develop an automated system allowing to code ADRs and their seriousness from patients' reports free text through the identification and validation of the best AI pipeline. To do so, we defined a global pipeline from document parsing and features extraction to ML model selection.

METHODS

Data collection

We selected all cases of ADRs reported by patients to Bordeaux Pharmacovigilance Centre through the French national web-portal between March 1, 2017, and February 28, 2019, for a total of 2,058 report forms. For each of these, the report form containing the free text filled by the patient was obtained together with the expert annotation and coding realized within the pharmacovigilance center. The coding of ADR was realized using the MedDRA standardized medical terminology developed by the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use.[16] MedDRA has been translated and is maintained in several different languages that allow interoperability between data from different countries. There are five levels to the MedDRA hierarchy, arranged from the very general to the very specific. Level four "Preferred Terms (PTs)" is recommended for the coding; 24,313 PTs are available in the terminology MedDRA 23.0 that correspond to single concepts for symptoms, signs, disease diagnoses, therapeutic indications, investigations, surgical or medical procedures, and medical, social, or family history characteristics. The ADRs coded by experts in PTs were used as a gold standard to learn and to evaluate the performances of each AI pipeline.

Training and validation set

These coded ADRs were used to train and validate ML algorithms for the identification of ADRs and ADR seriousness from the patient report-forms: 90% of cases (n = 1,852) were used for the train set and 10% (n = 206) for the internal validation set (test set). We have kept 90% of the data for the train set to maximize our learning abilities. We tested the robustness of the best performing model on this internal dataset through an external validation on a 10% sample of ADRs reported by patients to the Marseille Pharmacovigilance Centre over the same period (n = 187). The used dataset included in total 10,675 reported potential ADRs for the learning set (train set + test set) and 407 for the external validation set.

Data extraction and preprocessing

The dataset we were provided with for this study is composed of tables in PDF forms containing several fields of interest. The very first step was to extract the text from these PDF files preserving their structure. To do so we used the Python library Camelot.[17] The most informative section of the report forms was the description of the adverse events written by the patients. We performed basic text processing on the raw data, such as accents and punctuation removal, case lowering, and stemming. These transformations are useful to reduce noise and to limit the size of the vocabulary, specifically when we deal with small datasets.

Some other fields can be turned into structured features represented as:

- Integers: age, body mass index as a summary feature of weight and size
- Booleans: sex and one-hot representation of drugs

These additional features make it possible to take into account the patient's specificities and treatments.

Model-specific data formatting

Once the data extraction and text preprocessing steps are performed, we have to arrange the data in a format specific to the ML model we use for the prediction task. Most conventional ML models need numerical feature vectors as inputs. Consequently, we have to vectorize the adverse event description text. For that, we used the term frequency—inverse document frequency (TF-IDF) method.[18] The final features vector is obtained expanding TF-IDF vector with the additional features (age, sex, body mass index, and drugs).

The recurrent long-short-term-memory (LSTM) neural network we included in this study needs to be fed with word index sequences so a slightly different data formatting is applied. This is also the case for FastText and its variants (ExtremeText), which is directly provided with the preprocessed text.[19]

Identification of ADR terms and ADR seriousness

Technically, the identification of ADRs in free text is a text classification problem that can be solved using Natural Language Processing in a global AI pipeline, including a knowledge base about drugs (**Figure 1**). Definition of ADR seriousness results from a worldwide consensus with defined criteria, the use of which is mandatory for regulators and pharmaceutical companies.[20] To be serious, an ADR needs resulting in death, life-threatening, hospitalization (or prolongation of existing hospitalization), persistent or significant disability or incapacity, congenital abnormalities/birth defect, or another significant medical event. The assessment of these criteria needs considering information about ADRs, patients' characteristics, and treatments. Ultimately, this task leads to a binary classification decision resulting from the integration of heterogeneous information.

Models used for the benchmark

We wanted to compare various ML models usually used for text classification from the simple logistic regression to the more complex LSTM recurrent neural net.

Conventional ML models. We have evaluated four classical ML models: Logistic regression[21,22] and Support Vector Machine (SVM),[22,23] Random Forests,[24] and Gradient Boosting.[25,26] None of these model implementations, except Random Forests, natively accepts multilabel classification. To work around this problem, we used the One Versus Rest strategy in which each modality (here MedDRA term) to predict is treated independently from the others and we train one model per modality, each model being a binary classifier. See for instance for an application to the SVM algorithm.[27]

Gradient boosting used in this study was a gradient-boosted trees method. Recent gradient-boosted trees methods, such as XGBOOST[28] or LGBM,[29] provide generally state-of-the-art performance on various classification and regression ML problems. We chose the widely used python lightGBM[29] library for its remarkable performance in terms of accuracy as well as computing speed. The main difference between lightGBM and the other tree-based gradient boosting methods consists in the fact that the binary decision trees grow at the leaf level instead, whereas in algorithms, such as XGBOOST, they grow at the depth level (each leaf of the same depth level splits at the same time). This leaf level split provides a better adaptability and thus a better fit to the dataset.

FastText. FastText is an open-source method and python library for words representation and text classification.[30] The classification algorithm built on top of this representation method is a multinomial logistic

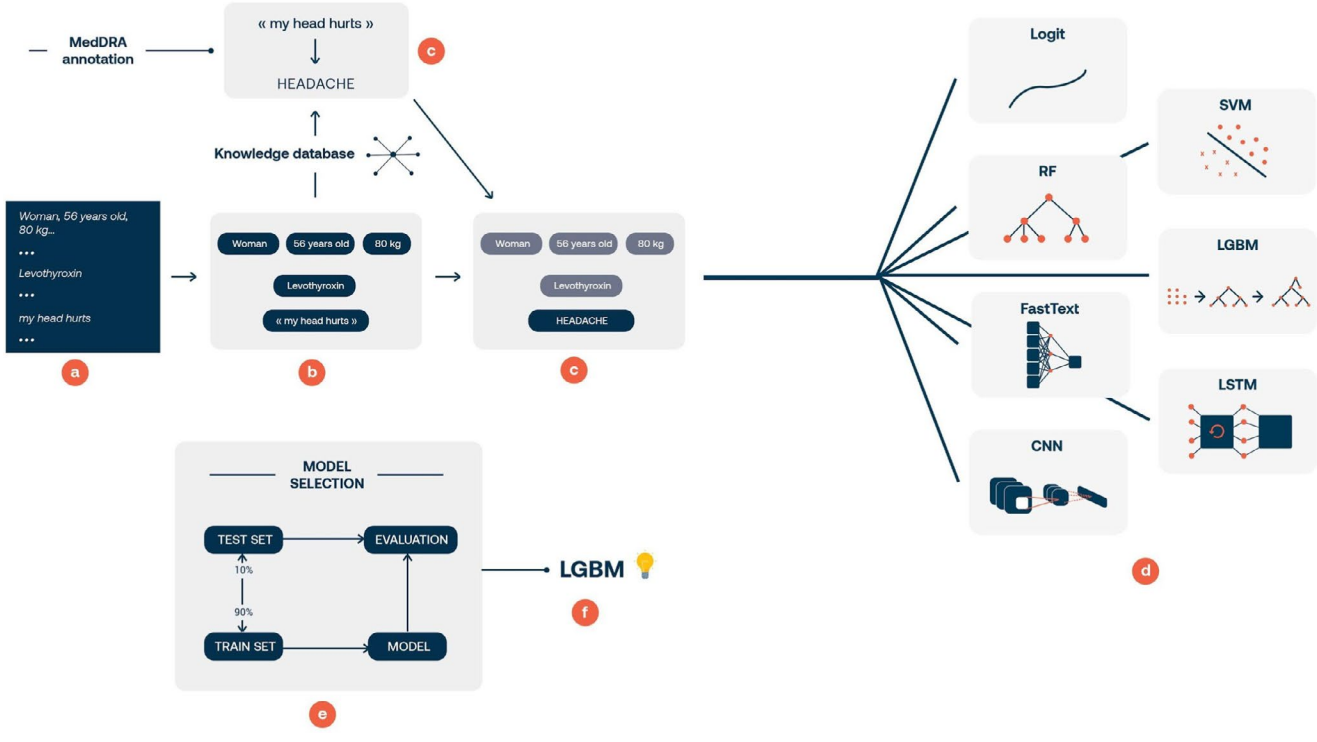

Figure 1 Artificial intelligence pipeline to identify and code an adverse drug reactions (ADRs) from free text using MedDRA terminology. (**a**) Patient sends an ADR report form with clinical information. (**b**) Text cleaning and extraction of relevant information. (**c**) Matching MedDRA terms into the case reports using knowledge graph. (**d**) Data formatting for our machine learning (ML) models. Conventional ML: logit, random forest (RF), support-vector machine (SVM) and light gradient boosting machine (LightGBM or LGBM). Neural networks and deep learning models: FastText, long short-term memory recurrent neural network (LSTM) and convolutional neural network (CNN). (**e**) Training ML models on 90% of our dataset (train set) and then computing evaluation metrics on the remaining 10% (test set). (**f**) Selection of the best ML model regarding area under the curve (AUC) and F-measure.

regression.[31] We trained the word embedding on the Synapse dataset composed of 260 MB of text dealing with medical topics extracted from French reference sources. This word embedding was used by the FastText classification model as well as the deep learning models as embedding layers. In this study, we used extremeText, which is an extension of FastText. It implements loss functions—like probabilistic label trees—well adapted for extreme classification tasks (i.e., classification tasks with a large amount of classes).

Deep learning models. Deep learning is a subset of ML methods gathering multilayer neural networks.[32] Deep learning methods are generally better than conventional methods to process complex data and large datasets. Two types of deep learning models were used for the study: Convolutional Neural Nets (CNN)[33] and LSTM[34] (**Figure S1**).

Regular expressions and hybrid models. A very simple and straightforward way to tag the case reports with MedDRA concepts is to look for their labels directly in the description text. A well-known limitation of this kind of approach is the need for a rich dictionary of synonyms as there are often several ways to express the same concept. Working with semantic resources like MedDRA is a great plus as the terminology provides us with—in addition to the PTs—several alternative labels for each concept (called Lower Level Terms (LLTs)).

We used regular expression (RegEx) after basic text preprocessing (accents removal, case lowering, and stemming) to match MedDRA terms into the case reports. This processing method was used on both the MedDRA labels and the narrative itself. On the one hand, we used this RegEx engine as a baseline for our ML models' benchmark; on the other hand, we built hybrid methods combining ML models and RegEx.

We do not expect patients to use terms that could exactly fit with a technical terminology, such as MedDRA. However, we can suppose that the RegEx approach allows the system to detect the rarer ADRs that are not present in the training dataset and for which ML techniques would not be efficient at all.

Hybrid models combine the RegEx approach and the ML models described above. There are several ways to aggregate predictions from several models. Here, we chose to average both vectors to build the final predictions, which is equivalent to methods used in bagging (for instance, random forests). Notice that due to the deterministic nature of regular expressions, the RegEx engine outputs binary vectors (0 or 1) whereas ML models provide scores between 0 and 1.

Assessment of model performances and tuning strategy

To identify ADRs and to determine ADR seriousness, each model was evaluated according to receiver operating characteristic (ROC) curve/area under the curve (AUC) and F-measure (harmonic mean of the precision and recall).[35]

We choose to select the decision threshold that maximizes the F-measure. Providing reliable evaluation metrics is essential to compare model performances. We used 1,000 bootstrap train-test samples to compute 95% confidence intervals for the evaluation metrics of each model. Some models, like random forests, gradient boosting, or neural networks, need hyper parameters tuning. For this, we used a classical grid search strategy on a dedicated train-test tuning split.

Overfitting and validation

Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. The best solution to control the overfitting of an ML model is to perform an external validation on independent data (external validation). In general, but even more for relatively small datasets, internal validation of prediction models by bootstrap techniques may not be sufficient and indicative for the model's performance in future data. Our internal validation was carried out on 10% of Bordeaux dataset: 206 cases. Our external validation was carried out on 10% of cases declared to the CRPV of Marseille during the same period: 187 cases. Moreover, the drug distribution of these 187 cases differed from the Bordeaux dataset and these cases were analyzed by different experts, which allows us to have a better estimate of the transposability of our system.

Programming

Concerning the model implementations, we used the python libraries scikit-learn and lgbm (gradient boosting) for the conventional ML models, keras (with Tensorflow backend) for the LSTM and CNN models and FastText + ExtremeText for the FastText model. The models were trained on an Intel Core i7 processor with 16 GB RAM.

Example code for LGBM (ML model with best performance) and for LGBM + regex are provided on the following git repository: https://github.com/louisletinier/MAITAI.git.

For more information about data management or our algorithms, you can contact the corresponding author at the email address indicated on the first page.

RESULTS

Dataset characteristics and representativeness

Patients who reported ADRs to the Bordeaux Centre were mostly women (88%) of median age 52 years (interquartile range (IQR): 40–62); 28% of reports mentioned serious ADRs. Altogether, 11,591 potential ADRs were reported in the 2,058 forms (median: 5 ADRs per report; IQR: 3–7). After coding by experts, they were found to correspond to 593 distinct ADR terms (MedDRA PTs). Overall, 2,396 drugs were mentioned; Levothyrox (levothyroxin) was the main represented (n = 1,623, 68%). We stratified the 10% random sampling of the Marseille reports for the representation of levothyroxine (10%) to ensure a wider variety of cases would be considered for external validation. Without stratification, the Marseille reports included a proportion of levothyroxine cases comparable to Bordeaux due to the "Levothyrox crisis" detailed in **Supplemental Information S1**. After this, 187 reports from Marseille were selected, 44% of which were serious. They concerned mostly women (74%); median age was of 44 years (IQR: 28–60). The reports mentioned 549 potential ADRs (median: 2 ADRs per report, IQR: 1–4) corresponding to 199 distinct PTs, and 201 drugs. The most frequent PTs for each center are listed in **Table S1**.

Owing to the size of the dataset, the training and the internal validation were performed considering PTs with at least 10 occurrences in the Bordeaux dataset, which led to consider 10,675 potential ADRs, and 125 distinct PTs in this dataset. The external validation focused on PTs present in the learning set, which led to consider 407 potential ADRs and 85 distinct PTs from the Marseille Centre reports in the validation set. Practically, the 125 PTs included in our dataset corresponded to 64.6% of the 393,407 ADRs entered in the BNPV for the period, and the 85 PTs in our external validation dataset corresponded to 58.3% of those. This guaranteed a strong representativeness of the BNPV in our dataset (**Figure S2**).

ML models performance to identify ADRs

Regarding AUC and F-measure obtained during internal validation, the model presenting the highest performance for automated coding of potential ADRs from patients' reports free text was

LGBM (AUC = 0.93, 95% confidence interval (CI) 0.92–0.94 and F-measure 0.72, 95% CI 0.68–0.75; **Figure 2a,b**). Among other models, logit (AUC = 0.91, 95% CI 0.89–0.92); F-measure 0.58, 95% CI 0.55–0.62), FastText (AUC = 0.89, 95% CI 0.87–0.91); F-measure (0.55, 95% CI 0.51–0.59), and LSTM (AUC = 0.86, 95% CI 0.88–0.89); F-measure (0.35, 95% CI 0.43–0.47) showed the highest performances. The hybrid LGBM with semantic approach model showed performances close to those of LGBM alone (AUC 0.94, 95% CI 0.93–0.95); F-measure (0.71, 95% CI 0.68–0.74).

Contrary to ML models that need a minimal occurrences number to be able to detect correctly a MedDRA term, the RegEx engine performances do not depend on the frequency of the terms and it will be able to detect even rare ADRs that were not selected among the 125 PTs on which we focused. However, for the sake of simplicity and to be able to compare models that include or not

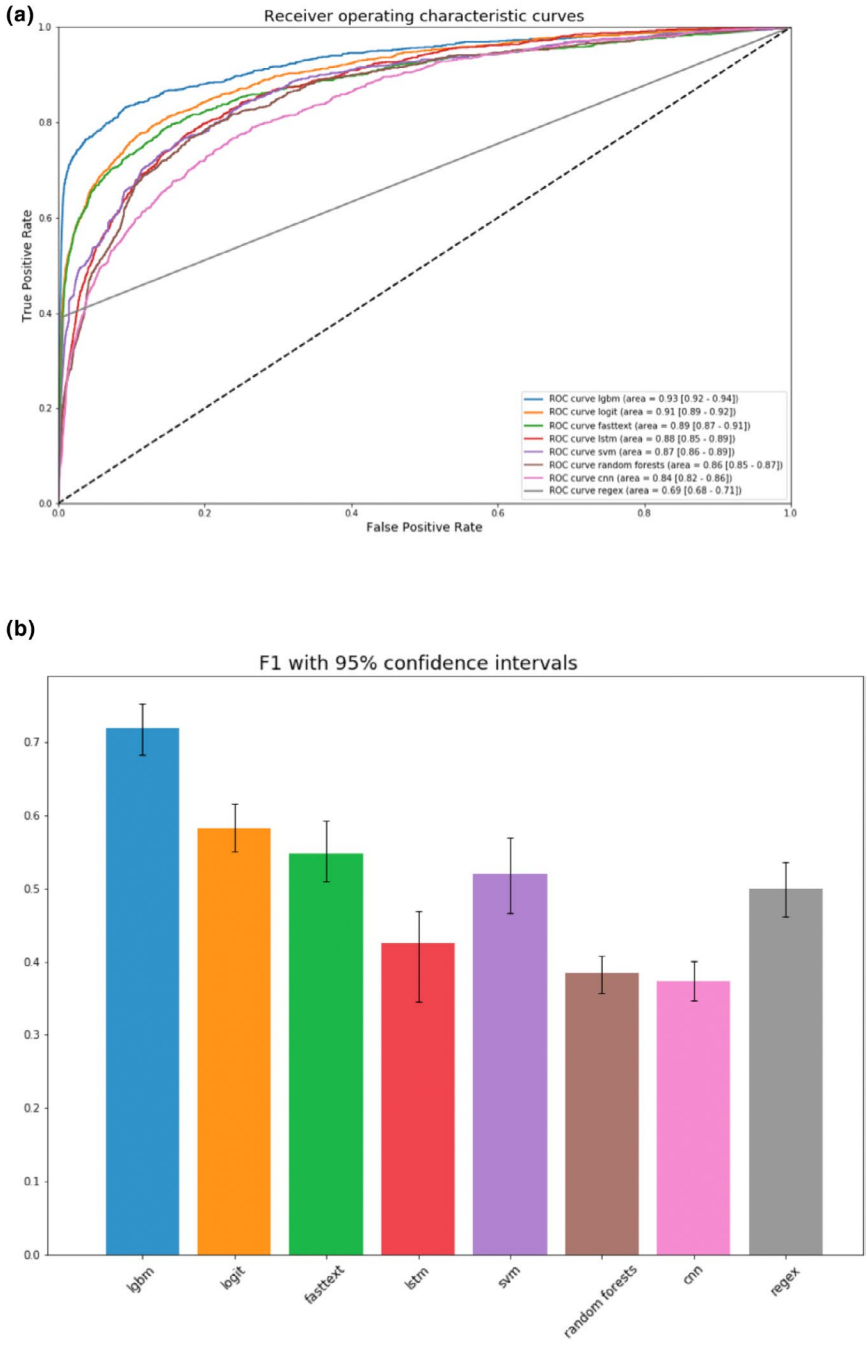

Figure 2 Performances of different machine learning models for the identification of adverse drug reaction from patient reports. (**a**) Performances of machine learning (ML) models in terms of receiver operating characteristic (ROC) curve and area under the curve (AUC) on the intern validation set. (**b**) Performances of ML models in terms of F-measure (F1) on the internal validation set). CNN, convolutional neural network; SVM, support-vector machine.

RegEx, we limited the evaluation of RegEx and hybrid models to the 125 most frequent PTs.

The best performing model, LGBM, was run for external validation showing an AUC of 0.91 and a F-measure of 0.58; the limited validation set size did not allow computing CIs.

Finally, we compared the results of the best performing model only trained on text features and the same model with text and structured features. The use of the structured features resulted in a slight performance improvement from 0.69 to 0.72 F-measure.

The overall performance of all models in internal validation and performance of the LGBM in external validation are detailed in **Table 1**. The parameters studied were ROC curve AUC, F-measure, true negatives (TNs), true positives (TPs), false negatives (FNs), and false positives (FPs).

ML models performance for ADR seriousness determination

Regarding AUC and F-measure, the 4 conventional ML models tested showed similar performance with AUC close to 0.75 (**Figure 3a**) and F-measure close to 0.60 (**Figure 3b**). Neural network models were unable to fit in the dataset effectively. As ADR seriousness determination is more complex than ADR identification, it is possible that this relates to the limited size of our learning set. Overall performance of conventional ML models to determine ADR seriousness are presented in **Table S2**.

DISCUSSION

Our study showed that an AI pipeline using a knowledge database to structure free text data and a ML model to learn ADR coding from human expertise could allow for automatic identification of a large proportion of the ADRs described in patients' report forms unstructured data. The best performing AI pipeline using LGBM (gradient boosting) showed interesting results with an AUC of 0.93 and an F-measure of 0.72. More concretely, from the used test set comprising 1,061 ADRs, our system succeeded in correctly identifying and coding 703 ADRs, and incorrectly spotted or coded 190 others.

The performances obtained for the automated determination of the ADR seriousness task were lower; enlarging the experiment to a larger set could allow improving these. Other teams, academic or industrial, are currently also working on this latter aspect using ML models disregarding the aspect of ADR identification from free text.[36,37]

For ADR identification, we found a relative discrepancy between AUC and F-measure, highlighting how these metrics differ in considering a classifier's performance. In the commonly used ROC curve/AUC, TP and the FP rates are considered of similar importance, which might present limitations for the evaluation of a classifier especially in the case of unbalanced data.[38] AUC can indeed present with very high overall values if the modality "absence of ADR" is dominant in the dataset and is related to a very high level of specificity, even the sensitivity of the model is low. The F-measure conversely provides a single score that balances both the aspects of positive predictive value (PPV) and sensitivity in just one score and is thus not affected by the predominance of a modality in the dataset. On the other hand, the main limitation of F-measure is that it gives equal weight to PPV and

Table 1 Overall performance of all models after internal validation and external validation for the best performing model (LGBM) to detect ADRs

Validation	Models	AUC (95% CI)	F-measure (95% CI)	TP (95% CI)	FP (95% CI)	TN[a] (95% CI)	FN (95% CI)
Internal (test set)	LGBM	0.93 (0.92–0.94)	0.72 (0.68–0.75)	703 (627–779)	190 (141–257)	24,496 (24,380–24,603)	358 (304–414)
	LGBM + RegEx	0.94 (0.93–0.95)	0.71 (0.68–0.74)	756 (680–838)	311 (259–369)	24 378 (24,261–24,485)	306 (253–359)
	Logit	0.91 (0.89–0.92)	0.58 (0.55–0.62)	539 (481–606)	248 (172–360)	24,442 (24,273–24,567)	520 (454–585)
	Fasttext	0.89 (0.87–0.91)	0.55 (0.51–0.59)	529 (454–602)	339 (225–483)	24,324 (24,057–24,507)	534 (468–595)
	LSTM	0.88 (0.85–0.89)	0.43 (0.35–0.47)	461 (370–545)	652 (452–1017)	24,033 (23,640–24,269)	597 (518–677)
	SVM	0.87 (0.86–0.89)	0.52 (0.47–0.57)	473 (400–543)	255 (174–516)	24,421 (24,156–24,559)	587 (509–677)
	Random forest	0.86 (0.85–0.87)	0.38 (0.36–0.41)	486 (351–540)	998 (472–1116)	23,690 (23,539–24,245)	583 (514–704)
	CNN	0.84 (0.82–0.86)	0.37 (0.35–0.40)	403 (331–481)	682 (446–1010)	24,008 (23,651–24,281)	656 (575–731)
	RegEx	0.69 (0.68–0.71)	0.50 (0.46–0.54)	418 (363–472)	196 (166–227)	24,491 (24,394–24,589)	644 (580–703)
External	LGBM[b]	0.91	0.58	240	185	22,783	167

ADRs, adverse drug reactions; AUC, area under the curve; CI, confidence interval; CNN, Convolutional Neural Nets; FN, false negative; FP, false positive; LSTM, long-short-term-memory; RegEx, regular expression; SVM, Support Vector Machine; TN, true negative; TP, true positive.

[a]TN correspond to all potential events correctly identified as non-present in the validation set amongst all possible. For instance, in the internal validation set, patient-reported ADRs were coded using 125 distinct MedDRA Preferred Terms (PTs) overall, the potential presence of which was assessed over the 206 cases included in the set corresponding to a potential for the identification of 25,750 events at maximum (125 * 206).

[b]As specified in the text, the limited validation set size did not allow computing confidence intervals.

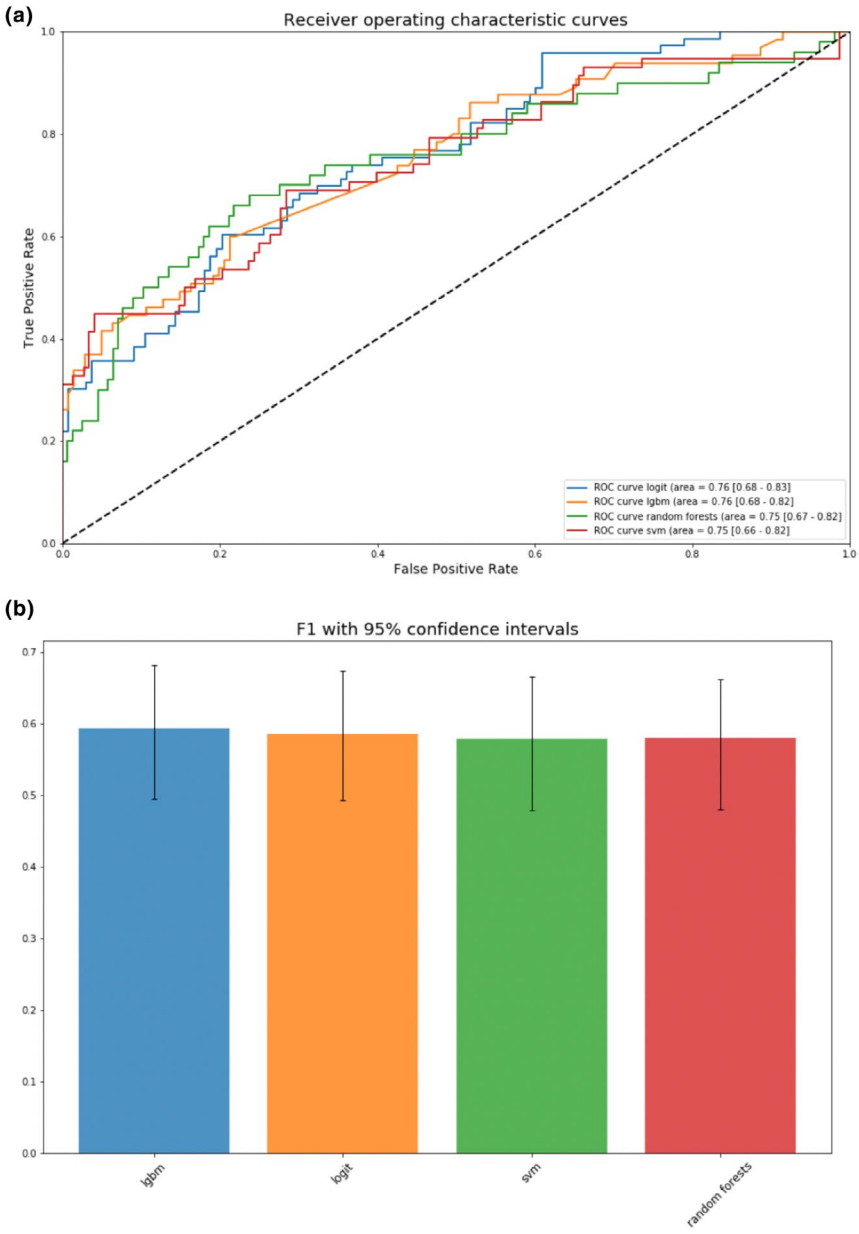

Figure 3 Performances of different machine learning models for the determination of adverse drug reaction seriousness from patient reports. (**a**) Performances of machine learning (ML) models in terms of receiver operating characteristic (ROC) curve and area under the curve (AUC) on the intern validation set. (**b**) Performances of ML models in terms of F-measure (F1) on the internal validation set). SVM, support-vector machine.

sensitivity, which in some circumstances might not correspond to the clinical needs of the tool being evaluated.[35] The results we present highlights the importance of using both measures when studying the performance of ML, especially when dealing with unbalanced data, such as health data.[39] Indeed, the comparison of our ML models shows that they all retain an AUC > 0.8 despite very average performance for some if we look at the TP, FP, or FN. This is explained by these models remaining very efficient for TN (**Table 1**). Moreover, being this a first study, it seems to us more appropriate global performance parameters, such as AUC and F-measure, than practical parameters, such as sensitivity or specificity.

The external validation strengthened the validity of our results, allowing us to determine how the model performances would transfer to new users and new patients.[40] In our results, the performances obtained were lowered during the external validation. Because the model was trained on data from a single center, this was, however, expected and should be improved by enlarging the learning to other centers. Similarly, the lower performances of the models in determining ADR seriousness, a more complex task involving numerous parameters, could be improved if training was performed on a larger dataset.

The use of real-life, expert-annotated data is the most important strength of our study. In contrast to simulation studies, it

demonstrates the concrete capacities of the AI pipeline to classify ADRs from unstructured textual data. Its performance could, however, consequently depend on peculiarities in the data and the transferability of the system to other settings (especially other languages) should be studied. As the overall pipeline we propose is entirely transposable to English and other languages with comparable structures, this should theoretically not constitute an issue. Medical concepts, drug names, and the MedDRA classification of ADRs indeed benefit from multilingual reference systems, and the semantic approach based on a knowledge base proposed here makes it possible to link identical concepts expressed in different languages.[41]

The main limitation of our study is the limited size of the datasets used. ML methods, especially deep learning ones, require a large amount of data.[42] One of the difficulties when dealing with textual data is the ability to differentiate between sentences almost similar in form but very different in meaning, for instance, because of a negation. Deep learning is theoretically the most adequate method to overcome this[32,43] but it requires having learned enough about these complex situations to identify them and determine new ones afterward.

Attention-based transformer architectures, such as BERT and its numerous variants, are showing very promising results and seem to be the best performing methods for this kind of task.[44] However, we did not find models pretrained on French convenient biomedical datasets and our dataset was too small to consider fine-tuning a BERT model so we did not include these models into the study. With more data available, it will be a good perspective to improve results obtained in the current work.

In larger studies, the model ranking regarding learning capacities could thus change.[45] However, even using a limited dataset, the performance obtained was already valuable. This could be due both to the quality of the expert-annotated data and to our semantic approach, which optimizes their understanding. However, the annotations we used to train the models and evaluate their performances were made manually. Moreover, due to the high complexity of the MedDRA terminology there is a high variability in the human ADR annotation process. This leads to a not insignificant noise in the dataset and hard to assess model errors. We did not perform any measurements of this inter-annotator variability in this study as the necessary qualified human resources were not available. However, we consider doing it as future work.

The results of this study are encouraging. On their basis, we will extend this learning to all pharmacovigilance French centers, which should enable to train our models on ~ 20,000 expert-annotated patient reports. We are also working to strengthen our semantic approach by taking into account textual data that correspond to a therapeutic indication and not to an adverse event. This will limit our FPs by excluding therapeutic indications.

The results of this preliminary study made it possible to launch the second phase of the study at the national level with 30 pharmacovigilance centers. The improvement in performance expected at the end of this second phase should make it possible to set up a new pharmacovigilance system allowing pharmacologists to save precious time. Indeed, this system will allow the ADR reports to be sorted by seriousness and by adverse events, which will allow pharmacologists to focus on the most serious cases or concerning unexpected adverse effects.

This new system is scheduled to be deployed in 2021 to facilitate pharmacovigilance of coronavirus disease 2019 (COVID-19) vaccines. The current French pharmacovigilance pipeline and the one that could be deployed after implementation are represented in **Figure S3**.

CONCLUSION

We elaborated an AI pipeline using a knowledge database and gradient boosting trees that was found able to learn to correctly identify ADRs from patient reports with unstructured data. If already interesting, the performance obtained can still be improved, especially for ADR seriousness determination, which will be intended using wider and more heterogeneous datasets. Such a system would allow answering the increasing needs of automated pharmacovigilance systems for the processing of ADR reporting.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors thank C. Faugas at Synapse Medicine for infographics. We also thank F. Tricot at Synapse Medicine for proofreading.

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

L.L., J.J., A.B., and C.G. were employed by Synapse Medicine at the time this research was conducted or hold stock/stock options therein. All other authors declared no competing interests.

AUTHOR CONTRIBUTIONS

L.L., J.J., M.B.K., A.B.L., C.G., A.S., F.R., C.L., G.M., J.M., F.S., and A.P. wrote the manuscript. L.L., J.J., F.S., and A.P. designed the research. L.L., J.J., M.B.K., A.B.L., G.M., J.M., F.S., and A.P. performed the research. L.L, J.J., C.G., A.S., F.R., C.L., F.S., and A.P. analyzed the data. L.L., J.J., M.B.K., A.B.L., C.G., A.S., F.R., C.L., G.M., J.M., F.S., and A.P. contributed new reagents/analytical tools.

ETHICS SECTION

This study and the use of French pharmacovigilance database have been validated by the French Agency for the Safety of Health Products: *Agence nationale de sécurité du médicament et des produits de santé*. These data remain the property of the *Agence nationale de sécurité du médicament et des produits de santé*. Furthermore, the storage of these data is managed by service providers certified as "Hébergeur de Données de Santé" (Health Data Host).

1. Sheikh, A., Jha, A., Cresswell, K., Greaves, F. & Bates, D.W. Adoption of electronic health records in UK hospitals: lessons from the USA. *Lancet* **384**, 8–9 (2014).

2. He, Z., Tao, C., Bian, J., Dumontier, M. & Hogan, W.R Semantics-powered healthcare engineering and data analytics. *J. Healthc. Eng.* **2017**, 1–3 (2017). http://doi.org/10.1155/2017/7983473
3. Vayena, E., Dzenowagis, J., Brownstein, J.S. & Sheikh, A. Policy implications of big data in the health sector. *Bull. World Health Organ.* **96**, 66–68 (2018).
4. Kish, L.J. & Topol, E.J. Unpatients—why patients should own their medical data. *Nat. Biotechnol.* **33**, 921–924 (2015).
5. Cambridge dictionary <https://dictionary.cambridge.org/fr/dictionnaire/anglais/artificial-intelligence>. Accessed April 9, 2021.
6. Wiens, J. & Shenoy, E.S. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin. Infect. Dis.* **66**, 149–153 (2018).
7. Liu, S., Tang, B., Chen, Q. & Wang, X. Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information* **6**, 848–865 (2015).
8. Laporte, J.-R. Fifty years of pharmacovigilance – medicines safety and public health. *Pharmacoepidemiol. Drug Saf.* **25**, 725–732 (2016).
9. Lazarou, J., Pomeranz, B.H. & Corey, P.N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279**, 1200–1205 (1998).
10. Aspden, P. Preventing medication errors: quality chasm series <http://psnet.ahrq.gov/issue/preventing-medication-errors-quality-chasm-series> (2006). Accessed June 24, 2020.
11. Formica, D. *et al*. The economic burden of preventable adverse drug reactions: a systematic review of observational studies. *Expert Opin. Drug Safety* **17**, 681–695 (2018).
12. Pirmohamed, M. *et al*. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ* **329**, 15 (2004).
13. Prescrire. MedDRA and pharmacovigilance a complex and little-evaluated tool. *Prescrire Int.* **25**, 247–250 (2016).
14. Basile, A.O., Yahi, A. & Tatonetti, N.P. Artificial intelligence for drug toxicity and safety. *Trends Pharmacol. Sci.* **40**, 624–635 (2019).
15. Schmider, J., Kumar, K., LaForest, C., Swankoski, B., Naim, K. & Caubel, P.M. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. *Clin. Pharmacol. Ther.* **105**, 954–961 (2019).
16. Harrison, J. & Mozzicato, P. MedDRA®: The tale of a terminology: side effects of drugs essay. *Side Effects of Drugs Annual* **31**, xxxiii–xli (2009).
17. Camelot: PDF Table Extraction for Humans — Camelot 0.7.3 documentation <https://camelot-py.readthedocs.io/en/master> (2020). Accessed March 6, 2020.
18. Ramos, J. Using TF-IDF to determine word relevance in document queries. <https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c> (2003). Accessed March 6, 2020.
19. Wydmuch, M., Jasinska, K., Kuznetsov, M., Busa-Fekete, R. & Dembczynski, K. A no-regret generalization of hierarchical softmax to extreme multi-label classification. NIPS 2018. <http://arXiv.org/abs/1810.11671> (2018). Accessed March 6, 2020.
20. Nebeker, J.R., Barach, P. & Samore, M.H. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Ann. Intern. Med.* **140**, 795–801 (2004).
21. Hilbe, J. *Logistic Regression Models* (CRC Press, Boca Raton, FL, 2009).
22. Pochet, N.I.M.M. & Suykens, J.A.K. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. *Ultrasound Obstet. Gynecol.* **27**, 607–608 (2006).
23. Mohamed, S.S. & Salama, M.M.A. Computer-aided diagnosis for prostate cancer using support vector machine. Medical Imaging 2005: Visualization, Image-Guided Procedures, and Display. International Society for Optics and Photonics <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5744/0000/Computer-aided-diagnosis-for-prostate-cancer-using-support-vector-machine/10.1117/12.598800.short> (2005). Accessed March 6, 2020.
24. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
25. Breiman, L. Bias, variance, and arcing classifiers <citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.7931&rep=rep1&type=pdf> (1996).
26. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Ann. Statist.* **29**, 1189–1232 (2001).
27. Hong, J.-H. & Cho, S.-B. A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification. *Neurocomputing* **71**, 3275–3281 (2008).
28. Chen, T., Guestrin, C. A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (ACM Press, San Francisco, CA, 2016) <http://dl.acm.org/citation.cfm?doid=2939672.2939785> Accessed March 12, 2020.
29. Ke, G. *et al*. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**, 1–9 (2017).
30. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword. Information. arXiv:160704606 <http://arxiv.org/abs/1607.04606> (2017). Accessed March 12, 2020.
31. Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. Bag of tricks for efficient text classification. arXiv:160701759. <http://arxiv.org/abs/1607.01759> (2016). Accessed March 12, 2020.
32. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
33. Krizhevsky, A., Sutskever, I. & Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
34. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
35. Hand, D. & Christen, P. A note on using the F-measure for evaluating record linkage algorithms. *Stat Comput.* **28**, 539–547 (2018).
36. Chauvet, R., Bousquet, C., Lillo-Lelouet, A., Zana, I. & Kimoun, I. Classification of the severity of adverse drugs reactions. *Studies Health Technol. Informat.* **270**, 1227–1228 (2020).
37. Routray, R. *et al*. Application of augmented intelligence for pharmacovigilance case seriousness determination. *Drug Saf.* **43**, 57–66 (2020).
38. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).
39. Zhao, Y., Wong, Z.-S.-Y. & Tsui, K.L. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *J. Healthcare Eng.* **2018**, 1–11 (2018).
40. Steyerberg, E.W. & Harrell, F.E. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
41. Letinier, L. *et al*. patent FR1661257 - Device and method for generating a database relating to drugs. <https://patents.google.com/patent/FR3059118A1/en> (2016). Accessed June 25, 2020.
42. Chen, X.-W. & Lin, X. Big data deep learning: challenges and perspectives. *IEEE Access.* **2**, 514–525 (2014).
43. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. Natural language processing (almost) from Scratch. arXiv:11030398 <http://arxiv.org/abs/1103.0398> (2011). Accessed June 25, 2020.
44. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, T. Pre-training of deep bidirectional transformers for language understanding. arXiv:181004805 <http://arxiv.org/abs/1810.04805> (2019). Accessed December 11, 2020.
45. Liu, Y., Chen, P.-H.C., Krause, J. & Peng, L. How to read articles that use machine learning. *JAMA* **322**, 1806–1816 (2019).