Pharmaceutical Medicine (2018) 32:391–401
https://doi.org/10.1007/s40290-018-0251-9

ORIGINAL RESEARCH ARTICLE

Training Augmented Intelligent Capabilities for Pharmacovigilance: Applying Deep-learning Approaches to Individual Case Safety Report Processing

Danielle Abatemarco[1] · Sujan Perera[2] · Sheng Hua Bao[2] · Sameen Desai[1] · Bruno Assuncao[1] · Niki Tetarenko[1] · Karolina Danysz[1] · Ruta Mockute[1] · Mark Widdowson[1] · Nicole Fornarotto[1] · Sheryl Beauchamp[1] · Salvatore Cicirello[1] · Edward Mingle[1]

Published online: 13 October 2018

Abstract

Introduction Regulations are increasing the scope of activities that fall under the remit of drug safety. Currently, individual case safety report (ICSR) collection and collation is done manually, requiring pharmacovigilance professionals to perform many transactional activities before data are available for assessment and aggregated analyses. For a biopharmaceutical company to meet its responsibilities to patients and regulatory bodies regarding the safe use and distribution of its products, improved business processes must be implemented to drive the industry forward in the best interest of patients globally. Augmented intelligent capabilities have already demonstrated success in capturing adverse events from diverse data sources. It has potential to provide a scalable solution for handling the ever-increasing ICSR volumes experienced within the industry by supporting pharmacovigilance professionals' decision-making.

Objective The aim of this study was to train and evaluate a consortium of cognitive services to identify key characteristics of spontaneous ICSRs satisfying an acceptable level of accuracy determined by considering business requirements and effective use in a real-world setting. The results of this study will serve as supporting evidence for or against implementing augmented intelligence in case processing to increase operational efficiency and data quality consistency.

Methods A consortium of ten cognitive services to augment aspects of ICSR processing were identified and trained through deep-learning approaches. The input data for model training were 20,000 ICSRs received by Celgene drug safety over a 2-year period. The data were manually made machine-readable through the process of transcription, which converts images into text. The machine-readable documents were manually annotated for pharmacovigilance data elements to facilitate the training and testing of the cognitive services. Once trained by cognitive developers, the cognitive services' output was reviewed by pharmacovigilance subject-matter experts against the accepted ground-truth for correctness and completeness. To be considered adequately trained and functional, each cognitive service was required to reach a threshold of F_1 or accuracy score $\geq 75\%$.

Results All ten cognitive services under development have reached an evaluative score $\geq 75\%$ for spontaneous ICSRs.

Conclusion All cognitive services under development have achieved the minimum evaluative threshold to be considered adequately trained, demonstrating how machine-learning and natural language processing techniques together provide accurate outputs that may augment pharmacovigilance professionals' processing of spontaneous ICSRs quickly and accurately. The intention of augmented intelligence is not to replace the pharmacovigilance professional, but rather support them in their consistent decision-making so that they may better handle the overwhelming amount of data otherwise manually curated and monitored for ongoing drug surveillance requirements. Through this supported decision-making, pharmacovigilance professionals may have more time to apply their knowledge in assessing the case rather than spending it performing transactional tasks to simply capture the pertinent data within a safety database. By capturing data consistently and efficiently, we begin to build a corpus of data upon which analyses may be conducted and insights gleaned. Cognitive services may be key to an organization's transformation to more proactive decision-making needed to meet regulatory requirements and enhance patient safety.

Extended author information available on the last page of the article

Key Points

Augmented intelligence may be the key to lightening the overwhelming workload and cognitive burden placed on pharmacovigilance professionals today.

Transcription and annotation of source documents is a scalable process to create an annotated corpus from which cognitive modules may be trained.

Feedback-driven training of deep-learning algorithms with a pharmacovigilance subject-matter expert's guidance has proven successful in training a consortium of cognitive services for individual case safety report processing.

1 Introduction

The thalidomide tragedy of the 1950s and 1960s is often cited as the key mobilizing factor toward greater health legislation and heightened drug surveillance to ensure and improve patient safety. Before its recognition as a highly teratogenic drug, thalidomide was hailed as a 'wonder drug' and was widely prescribed as a sedative to pregnant women suffering from headaches, nausea, and insomnia; in the drug's first year of production, sales reached 90,000 units per month in 20 countries. Although thalidomide was never approved for sale in the USA because of the recommendation of Frances Kelsey, a researcher at the US Food and Drug Administration (FDA), it has been estimated that approximately 10,000 children globally were born with serious abnormalities, including phocomelia [1]. The sale of thalidomide was banned in most countries by 1961 [1, 2].

From that tragedy were borne two principles that guide the pharmaceutical and research industries today. The first is the regulation that all products undergo developmental toxicology tests in two species, one of which is not a rodent [2]. The second is the science of pharmacovigilance. Pharmacovigilance is defined by the World Health Organization (WHO) as the science and activities relating to the ongoing detection, assessment, and understanding of adverse events (AEs) or adverse drug reactions (ADRs) to assess a product's risk profile. After the thalidomide disaster, the WHO established its Programme for International Drug Monitoring, in which 134 countries participate by supplying country-level data to assess and determine the risk–benefit profile of drugs [3]. Thalidomide is just one example of a product found unsafe in its original use; from 1964 to 2002, 75 products have been withdrawn from the market because of safety concerns [4].

Individual case safety reports (ICSRs), which contain ADRs, are shared between stakeholder groups, namely pharmaceutical companies and regulatory authorities, to promote public health. Despite appropriate mitigation strategies in place, ADRs account for 3–7% of all medical hospital admissions and rank among the top ten leading causes of mortality [4, 5]. Estimates put the cost of drug-related morbidity and mortality at US$136 billion annually [6]. Regulations set forth by both the US FDA and European Medicines Agency are increasing the scope of activities that fall under the remit of drug safety, including ADR collection, risk management, signal detection, medical assessment, and risk communication. The WHO and the Council for International Organizations of Medical Sciences (CIOMS) have increasingly pushed for standardized use of electronic formats and secure submissions of regulatory documents [3].

The emergence of new methods and sources for collecting data, such as from social digital media and other electronic sources, has led to the expanding scope of responsibility faced by global drug safety and risk management divisions within industry [6]. In fact, the combined industry ICSR volume growth from 2003 to 2014 increased 33%, much of which is attributable to cases from new data sources and regulatory requirements changes [7]. Today pharmacovigilance professionals typically perform many transactional activities to enter pertinent data before it can be assessed for aggregated analyses and evaluation. For a biopharmaceutical company to meet its responsibilities to patients and regulatory bodies regarding the safe use and distribution of its products, automation may serve as a potential scalable solution for ICSR intake and processing [8]. Machine-learning and natural language processing (NLP) have shown success in detecting AEs from safety databases [9], social media [8], and medical discharge summaries [10]. In this study, we investigate the creation of cognitive services using state of the art machine-learning and NLP techniques for use within the pharmacovigilance domain with variously structured spontaneous data sources.

2 Objectives

The aim of this study was to train and evaluate a consortium of cognitive services to identify key characteristics of spontaneous ICSRs satisfying the business-defined threshold of F_1 or accuracy $\geq 75\%$. As a precursory task to training the cognitive services, the creation of an annotated corpus in a scalable manner was required. The results of this study will serve as supporting evidence for or against the implementation of augmented intelligence in case processing to increase operational efficiency and data quality consistency when processing ICSRs.

3 Methods

3.1 Background

Machine-learning is a subfield of computer science that learns patterns from data without providing explicit programming instructions to create algorithms intended to perform a specific task. NLP is the subfield in computer science that intends to teach computers to understand, interpret, and manipulate the human language. Most of the NLP tasks leverage the capabilities of machine-learning to achieve its objective. Deep-learning is the latest advancement in the machine-learning domain which focuses on learning data representation. It aims to develop algorithms that are more generalizable as opposed to being task-specific. A cognitive service is a mixture of machine-learning and NLP algorithms to solve a given problem that requires human cognition (e.g., discriminate between health conditions and AEs in a spontaneous report). Cognitive services are trained using input data that has been appropriately curated; data must be transcribed into a machine-readable format and contain relevant annotation labels, or tagged metadata, explaining each data entity's relevance to the learning task. The entire set of annotated, machine-readable text forms the 'annotated corpus'.

The processes of transcription and annotation are completed manually by staff with domain knowledge relevant to the task at hand. The training of the cognitive services using manually curated pharmacovigilance input data is carried out by cognitive developers who have background in deep-learning and NLP. Once trained, the algorithms are tested on new data to produce outcomes. These outcomes, or predictions, are automatically generated via the software. To ensure that the algorithms are producing predictions at the acceptable level, a human subject-matter expert (SME) reviews them and provides feedback to the cognitive developers so that they may refine the models and calculate their final evaluative scores.

3.2 Scope

In pharmacovigilance, the intake of a case requires a pharmacovigilance professional to first assess case validity. A valid ICSR must include at minimum four criteria: an identifiable reporter, an individual identifiable patient, one suspect medicinal product, and one suspect adverse reaction. Therefore, five cognitive services that are needed to ensure case validity were identified by the pharmacovigilance SMEs for investigation in this study: AE detection, product detection, reporter detection [recognizing the presence of a reporter and then classifying the reporter as a healthcare professional (HCP) or non-HCP], patient detection, and validity classification. To characterize the detected data elements, a set of additional services were identified: seriousness classification, reporter causality classification, and expectedness classification. Once AEs are detected, each is coded using the *Medical Dictionary for Regulatory Activities* (MedDRA); all drugs detected are coded according to the WHO Drug Dictionary (WHO-DD).

Entity detection cognitive services were assessed using the F_1 measure, and classification cognitive services were assessed using accuracy. The business threshold for a service to be considered adequately trained is when its corresponding evaluation measure exceeds 75%. This threshold was decided by considering: (1) business requirement; (2) data and resource availability; and (3) technical feasibility. Business requirement stated that humans in the loop should be able to simply confirm the results generated by the cognitive services for the majority of outputs and they should not be required to reread the entire document to find the relevant datapoints. However, this requirement was constrained by the data and resource availability as manual annotation to train cognitive services is a laborious task, and it is necessary to invest a significant amount of time from both the cognitive service developer and the domain expert to understand the errors in the cognitive output and tune the models to rectify those errors. Limited availability of the data leads to the technical feasibility as cognitive services get better with greater volume and diversity of the data used to train them. Considering these factors and the derived insights and experience from the proof-of-concept projects conducted before this study, we decided that 75% is the minimum threshold for each cognitive service to be effective in a real-world setting.

3.3 Annotated Corpus Sampling Data

The annotated corpus comprises 20,000 ICSRs sampled from the total dataset of 168,000 cases received by Celgene's drug safety department from January 2015 through December 2016. This dataset served as the input data for training the ten cognitive services under development. The sample was chosen by the cognitive developers considering the diversity and representativeness of the dataset from both the pharmacovigilance and machine-learning perspective. The factors considered for sampling were: (1) report type (spontaneous, clinical/market study, medical literature); (2) source country; (3) number of unique preferred terms; (4) number of unique reported terms; (5) length of the reported term; (6) seriousness of the ICSR; (7) seriousness of the AE; (8) seriousness category of the AE; (9) number of unique suspect products; and (10) expectedness value for investigator brochure, company core data sheet, summary of product characteristics, and product insert.

3.4 Scalable Transcription and Annotation Process

The scalable process of creating transcribed and annotated documents for the annotated corpus was designed as follows (Fig. 1). All cases were stored and worked on within a restricted site; documents were organized into folders corresponding to unique ICSRs. An external vendor with staff specializing in pharmacovigilance case processing performed the transcription of source documents and manually transcribed the case data into a blank Microsoft Word® (Microsoft Corp., Redmond, WA, USA) template to match the original source document in content and format. A subset

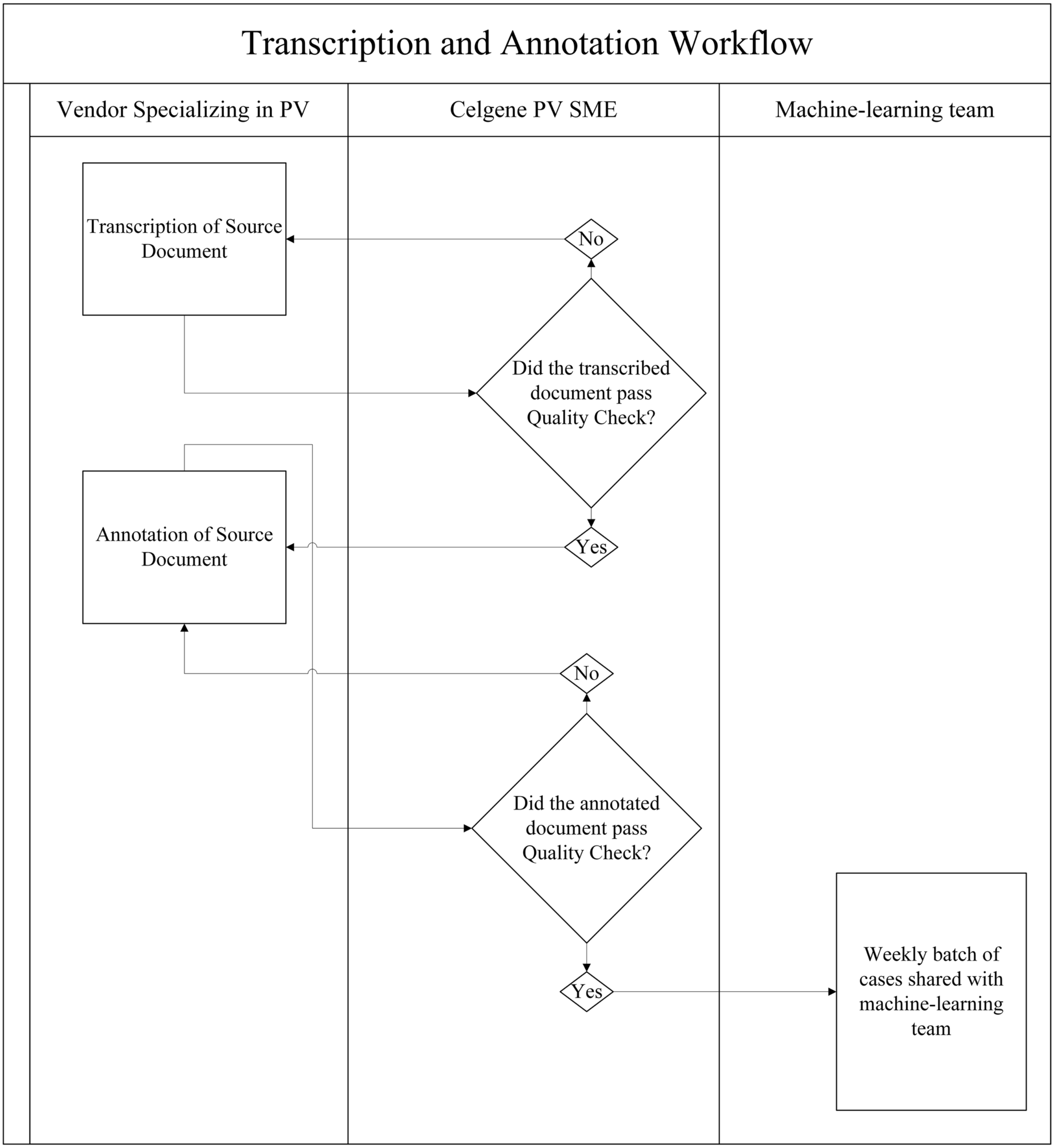

Transcription and Annotation Workflow
Vendor Specializing in PV
Celgene PV SME
Machine-learning team
Transcription of Source Document
No
Did the transcribed document pass Quality Check?
Yes
Annotation of Source Document
No
Did the annotated document pass Quality Check?
Yes
Weekly batch of cases shared with machine-learning team

Fig. 1 The scalable transcription and annotation workflow to produce an annotated corpus from which cognitive services can be trained in the pharmacovigilance domain. *PV* pharmacovigilance, *SME* subject-matter expert

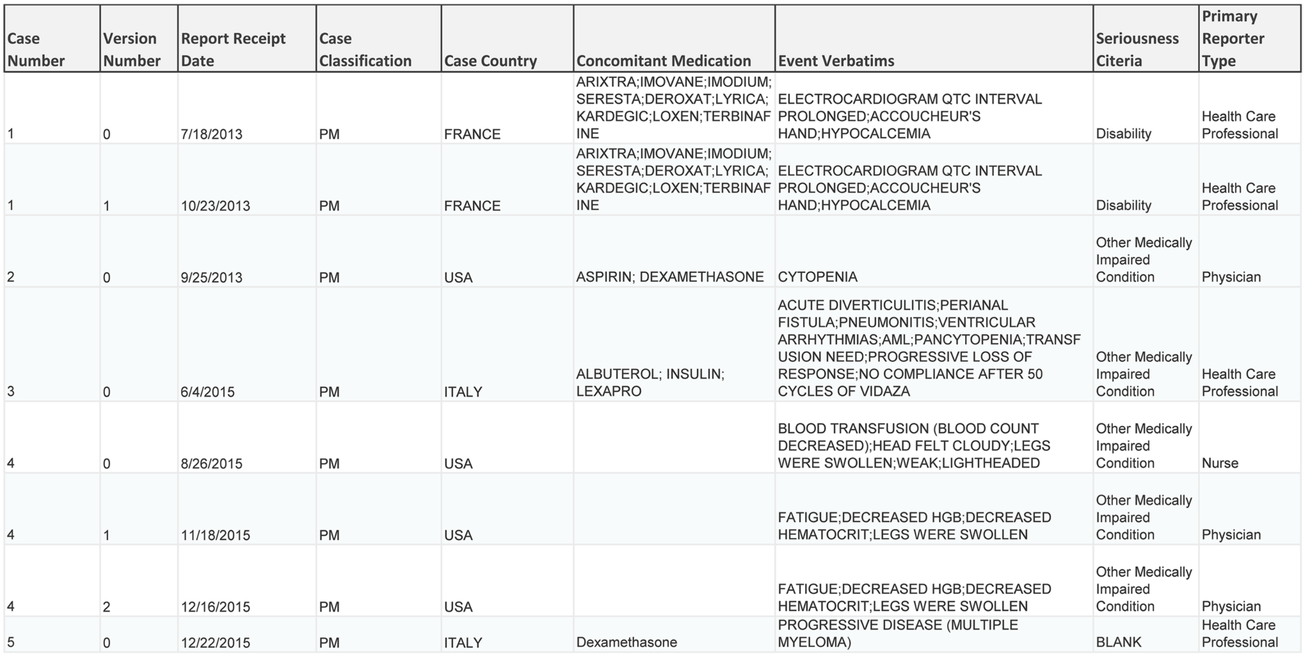
of these transcribed cases was quality checked in-house by pharmacovigilance SMEs. Once cases passed quality check, they were moved into the restricted site designated for the annotation, or tagging, of the transcribed documents. To guide their annotations, team members were provided a metadata sheet containing data for the 20,000 cases that comprise the annotated corpus. The metadata sheet is an output of the current safety database and contains the data pertaining to the way in which each case was originally processed and coded upon receipt. This document served as the ground truth against which the annotators applied the annotation labels. A section of the metadata sheet is shown in Fig. 2. The task of the human annotator is to find the datapoints present in the metadata within the corresponding case and annotate them in the transcribed document.

Criteria for a quality-check failure were as follows: omission or incorrect transcription or annotation of any major data element pertaining to case validity (patient, reporter, AE, or suspect product) or four or more minor errors. A minor error was defined as missing or incorrectly transcribing or annotating any data element that was not one of the four criteria for case validity (e.g., misspelling of a concomitant product, missing medical history, omission of patient's sex). Documents that failed quality check were reinserted into the vendor's queue for correction with a comment highlighting the error(s) found. All annotated documents that passed quality check were shared with the cognitive service development team on a weekly basis.

3.5 Standardization of Annotation

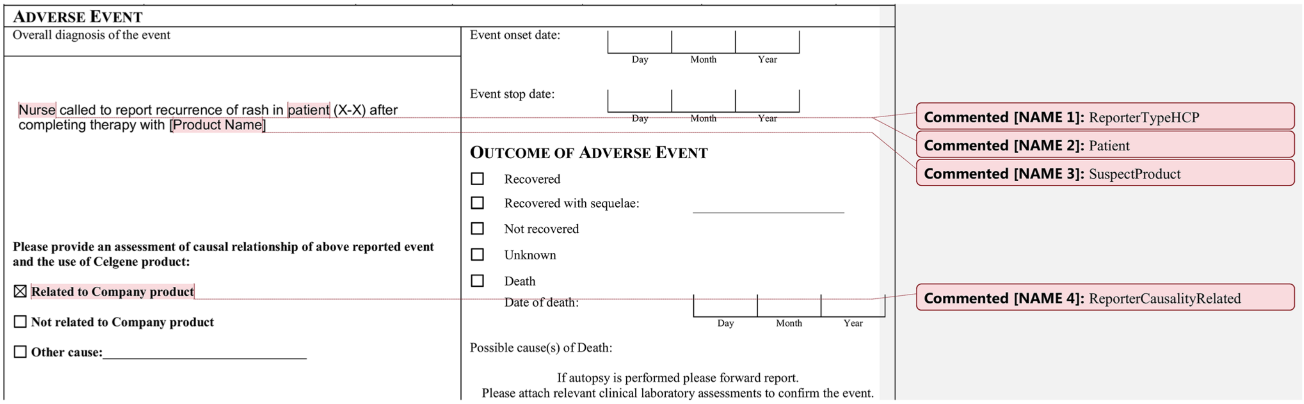
Annotation labels served as generic tags to represent the relevance of a data element to pharmacovigilance within the context of a given ICSR. Annotations were applied using the comment function within Microsoft Word® (Fig. 3). Some examples of annotation labels relating to the services within scope of this paper include ReporterTypeHCP, ReporterCausality, and SuspectProduct. Microsoft Word® macros were developed to standardize and expedite the annotation process; macros were installed by each annotator in Microsoft Word® and contained every annotation label in its proper format. Guidelines were provided to indicate proper label use, which contained all labels and their associated definitions, including an example of the context in which each label should be used.

A benefit of standardizing the approach to annotation allowed for quality control measures to be implemented. In addition to the manual quality check process for a statistically significant portion of annotated documents, 100% of the annotated documents went through an automated validator. This validation script ran on all completed documents and cross-checked the manual annotations with the metadata for completeness. For instance, the validation script checks the number of manual AE annotations against the number of AEs present in the metadata for the corresponding case to ensure completeness of the annotation task.



Case Number	Version Number	Report Receipt Date	Case Classification	Case Country	Concomitant Medication	Event Verbatims	Seriousness Citeria	Primary Reporter Type
1	0	7/18/2013	PM	FRANCE	ARIXTRA;IMOVANE;IMODIUM;SERESTA;DEROXAT;LYRICA;KARDEGIC;LOXEN;TERBINAFINE	ELECTROCARDIOGRAM QTC INTERVAL PROLONGED;ACCOUCHEUR'S HAND;HYPOCALCEMIA	Disability	Health Care Professional
1	1	10/23/2013	PM	FRANCE	ARIXTRA;IMOVANE;IMODIUM;SERESTA;DEROXAT;LYRICA;KARDEGIC;LOXEN;TERBINAFINE	ELECTROCARDIOGRAM QTC INTERVAL PROLONGED;ACCOUCHEUR'S HAND;HYPOCALCEMIA	Disability	Health Care Professional
2	0	9/25/2013	PM	USA	ASPIRIN; DEXAMETHASONE	CYTOPENIA	Other Medically Impaired Condition	Physician
3	0	6/4/2015	PM	ITALY	ALBUTEROL; INSULIN; LEXAPRO	ACUTE DIVERTICULITIS;PERIANAL FISTULA;PNEUMONITIS;VENTRICULAR ARRHYTHMIAS;AML;PANCYTOPENIA;TRANSFUSION NEED;PROGRESSIVE LOSS OF RESPONSE;NO COMPLIANCE AFTER 50 CYCLES OF VIDAZA	Other Medically Impaired Condition	Health Care Professional
4	0	8/26/2015	PM	USA		BLOOD TRANSFUSION (BLOOD COUNT DECREASED);HEAD FELT CLOUDY;LEGS WERE SWOLLEN;WEAK;LIGHTHEADED	Other Medically Impaired Condition	Nurse
4	1	11/18/2015	PM	USA		FATIGUE;DECREASED HGB;DECREASED HEMATOCRIT;LEGS WERE SWOLLEN	Other Medically Impaired Condition	Physician
4	2	12/16/2015	PM	USA		FATIGUE;DECREASED HGB;DECREASED HEMATOCRIT;LEGS WERE SWOLLEN	Other Medically Impaired Condition	Physician
5	0	12/22/2015	PM	ITALY	Dexamethasone	PROGRESSIVE DISEASE (MULTIPLE MYELOMA)	BLANK	Health Care Professional

Fig. 2 A section of the metadata sheet, an output of the current safety database, which contains the data elements to be annotated within each version of the cases comprising the annotated corpus



ADVERSE EVENT

Overall diagnosis of the event

Nurse called to report recurrence of rash in patient (X-X) after completing therapy with [Product Name]

Please provide an assessment of causal relationship of above reported event and the use of Celgene product:

☒ Related to Company product

☐ Not related to Company product

☐ Other cause:________________________

Event onset date: Day Month Year

Event stop date: Day Month Year

OUTCOME OF ADVERSE EVENT

☐ Recovered

☐ Recovered with sequelae: ____________________

☐ Not recovered

☐ Unknown

☐ Death

Date of death: Day Month Year

Possible cause(s) of Death:

If autopsy is performed please forward report.

Please attach relevant clinical laboratory assessments to confirm the event.

Commented [NAME 1]: ReporterTypeHCP

Commented [NAME 2]: Patient

Commented [NAME 3]: SuspectProduct

Commented [NAME 4]: ReporterCausalityRelated

Fig. 3 An example of annotation labels applied within a portion of a transcribed source document

Table 1 The ten cognitive services for spontaneous individual case safety report processing under development in this study with their corresponding service type

Cognitive service	Service type
Adverse event detection	Entity extraction
Drug detection	Entity extraction
Reporter detection	Entity extraction
Patient detection	Entity extraction
Validity classifier	Classifier
Seriousness classifier	Classifier
Reporter causality classifier	Classifier
Expectedness classifier	Classifier
MedDRA coding	Classifier
WHO-DD coding	Classifier

MedDRA Medical Dictionary for Regulatory Activities, *WHO-DD* World Health Organization Drug Dictionary

3.6 Designing and Training Cognitive Services

The ten cognitive services were categorized into two mainstream machine-learning and NLP tasks: (1) entity annotation; and (2) text classification (Table 1). We designed and developed deep-learning-based solutions to perform these cognitive tasks. To perform entity annotation, we implemented a recurrent neural network with bidirectional long–short-term memory layer to encode the input text and conditional random field-based decoder to identify the words that indicate interested entities in the encoded input text [11]. To perform classification tasks, we implemented a convolutional neural network with one convolutional layer and a max pooling layer [12]. The max-pooled outputs were concatenated and the softmax function was applied to determine the classification result. Interested readers may refer to Ma and Hovy [11] and Kim [12] for more details on these deep-learning models. The training process consisted of tuning the parameters of these networks to generate expected output for each service, and this was accomplished in an iterative manner with the feedback from the SMEs.

3.7 Guideline Iterations

Annotation guidelines were iterated upon and refined based on feedback by the pharmacovigilance SMEs during model training and approval. This cycle is referred to as ‘MATTER’, representing the process of modeling, annotating, training, testing, evaluating, and revising [13]. The first version of the guidelines for the annotated corpus was developed while annotating 3000 cases in-house to understand the breadth of entity labels that would be required for all case types to train the cognitive services within scope. Through model development, model predictions were analyzed for false negatives and false positives to pinpoint labels that were not impactful. Those found to be impactful remained; others were revised or removed. For example, the label ReporterCausality was revised into more specific labels: ReporterCausalityRelated, ReporterCausalityNotRelated, ReporterCausalityUnknown, and ReporterCausalityPossiblyRelated. Consideration was given to the tradeoff between creating very specific labels that might be used only rarely and creating generic labels that do not provide impactful knowledge to the cognitive services under development.

Version 1.0 of the annotation guidelines contained 75 annotation labels that were organized into nine categories. Version 2.0 of the guidelines contained 108 labels structured into the same nine categories. After model training and feedback sessions, it was determined that the version 2.0 labels would need to provide more detail and specificity. The most current version of the guidelines, version 3.0, contains 121 labels organized into 11 categories. This version was created to refine the previous labels and include new labels for annotating unstructured documents. The final 11 categories, which were created to group all annotation labels in that category by concept, are: Administrative, Event, Literature, Medical History, Patient, Product, Reporter, Reporter Causality/Seriousness, Tests, Study, and Questionnaire. Examples of annotation labels within the ‘product’ category include SuspectProduct, ConcomitantProduct, TreatmentProduct, PastProduct, ActionTaken, AdminRoute, DoseForm, DoseUnit, Frequency, and Indication.

It should be noted that in the development of the classification services, ICSR validity, WHO-DD, and MedDRA were not annotation dependent as these cognitive services do not require information about where in the original document the interested entities appear to make their decisions. The relevant labels for these services were present in the associated metadata file used for training. For instance, each AE in the metadata file is associated with its preferred and lowest-level term from its MedDRA hierarchy, and these data were directly used to train the MedDRA coding cognitive service without any extra manual annotation effort.

3.8 Model Approval

To date, approximately 14,000 cases of the 20,000 cases that compose the annotated corpus have completed the full workflow of transcription and annotation. Eighty percent of the cases shared with the machine-learning team were routed to training and tuning the cognitive services, 10% were used for testing the models with SME feedback, and 10% were reserved for final testing and validation. The models were assessed using the F_1 score, or accuracy. F_1 is an accepted measure of how well a cognitive model performs on entity detection scenarios; it represents the tradeoff between a model’s precision and its recall, as defined in Eqs. 1–3 (Table 2).

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive } + \text{False Negative}} \tag{3}$$

Testing and approval of services required collaboration between pharmacovigilance SMEs and machine-learning system developers. The pharmacovigilance SMEs reviewed all false negative and false positive outcomes of each service, providing feedback to be incorporated into the service for improvement. The aim of this feedback was to identify the common errors made by the cognitive services from the pharmacovigilance perspective. Once the model reached a score of at least 75%, the SMEs reviewed a sample of all true positives and, for binary models, true negatives to confirm that indeed they were true positives and true negatives. This process ensured occasional errors in manual annotations and labels were not counted positively or negatively to the final evaluation of the cognitive service. If the true positive and true negative review passed, the model was approved. If the model did not pass review, the teams worked to improve the model, typically through another round of false negative and false positive analysis.

Table 2 Table of confusion depicting the determination of true positive, false positive, true negative, and false negative for use in calculating the F_1 and accuracy measures

Ground-truth	Predicted	
	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

4 Results

To date, all ten cognitive services have reached their respective minimum evaluative score of 75%, and six of them reached 90% and above; thus, all models have been approved for future use (Table 3). These results demonstrate success in the creation and implementation of a scalable workflow to curate data comprising the annotated corpus used for training cognitive services in the pharmacovigilance domain. This repeatable framework may be applied to different datasets, and for the creation of novel cognitive services beyond the scope of this study to meet varying business needs. For example, a company with a specific product portfolio may see benefit in creating a cognitive service that identifies events of interest so that cases may be prioritized more efficiently. With this type of scalable, customizable technology, it is possible for a biotech company to implement augmented intelligence throughout the

Table 3 The ten cognitive services for spontaneous individual case safety report processing with their corresponding evaluative measure and score following model training and tuning

Cognitive service	Evaluation measure	Evaluation score
Adverse event detection	F_1 score	75.6
Drug detection	F_1 score	90
Reporter detection	F_1 score	94.99
Patient detection	F_1 score	79
Validity classifier	Binary accuracy	98.40%
Seriousness classifier	Binary accuracy	83%
Reporter causality classifier	Accuracy	78.43%
Expectedness classifier	Binary accuracy	92.50%
MedDRA coding	Top-5 accuracy	92%
WHO-DD coding	Top-5 accuracy	98%

MedDRA Medical Dictionary for Regulatory Activities, *WHO-DD* World Health Organization Drug Dictionary case-processing workflow to enable its employees to make more efficient and consistent decisions.

Through this undertaking, the researchers have learned the nuances of training cognitive services for specific tasks relating to ICSR intake and processing. The major takeaways are as follows:

1. When designing cognitive services, ensure the input data is readily available within the dataset.
2. When designing cognitive services, ensure the intended output is clearly defined and that it meets the business requirement. If possible, the cognitive service output should be normalized to values acceptable in ICSR processing. For example, if the case mentions that a patient was admitted to the hospital, the seriousness detection service should produce a standard output of "Serious—Hospitalization" so that the data are captured consistently regardless of the reporter verbatim.
3. When designing cognitive services, ensure the performance metrics are defined beforehand so that the services satisfy the business requirements.
4. The transcription of source documents, although time consuming, is important in that it preserves the ICSR's original formatting and most closely mimics cases that will be received in future scenarios, increasing the quality of training data fed into the machine-learning models.
5. The annotations must have very clear guidelines for their usage. For example, the annotation for AdverseEvent should not be used for disease states that are considered symptoms of the disease or medical history of the patient. This ensures high consistency and reliability between all users performing annotations.
6. The feedback provided by the pharmacovigilance SMEs during model approval and validation should be as generalized as possible. Incorporation of feedback that reflects very specific company conventions may hinder the performance of the cognitive services in future cases. This is known as overfit in machine-learning terminology, meaning the model is over-tuned to perform well on the current dataset. Such specific scenarios should instead be addressed as a post-processing step on the output generated by the machine-learning model. One possibility is to develop rules to post-process the cognitive services' output to meet the business expectation.

5 Discussion

5.1 Research Trends in Technology

Early use of technology and informatics in pharmacovigilance dates back to 2000; drivers for change included faster computer processing speeds, improved storage capabilities, and increased use of automation [3, 14]. Then, most studies leveraged homogeneous data sources (e.g., discharge summaries, clinical notes, or literature articles) to target isolated elements of a case report, such as AE detection or drug–drug interactions (DDIs) through keyword or trigger-phrase searches [14]. This approach yielded relatively limited success because it relied heavily on the presence or absence of defined keywords and phrases and did not dynamically incorporate growing vocabularies or changes in reporting norms [14]. Symbolic methods expanded on keyword searches to use semantic and syntactic patterns within one dataset to widen their results to include medical concepts and drugs. Symbolic methodologies proved more flexible and successful than simple keyword searches at capturing detailed information [14].

Beginning in 2010, researchers recognized that their methods could be improved by training with multiple data sources and implementing NLP techniques [14, 15]. Multiple data sources that combine electronic health records with spontaneous reports have been shown to outperform single sources in detecting ADRs by 30% [6]. Given the amount of free text and medical information found in literature articles and clinical summaries, these data sources make good candidates for training NLP models [16]. Furthermore, current estimates show the extent of these untapped data sources for use in drug surveillance: only 1% of AEs from electronic sources are reported to federal databases [14].

Many investigative studies using publicly available NLP systems have demonstrated limitations of these services, including insufficient dictionaries and ontological information that require additional tools, such as customer-customized dictionaries, to create large corpora of medical terms to classify data elements adequately [14, 17]. Established NLP systems also perform poorly in linking temporal relationships between entities, making them subpar models for determining whether AEs are related to pre-existing medical history and for determining causality between drugs and AEs [14].

Because of the overall limitations of using NLP alone in pharmacovigilance, recent trends seek to leverage statistical analysis, machine-learning, and NLP together to create a superior approach [4, 14]. The advancements of deep-learning algorithms are driving the latest momentum in cognitive service development. In this study, we used NLP in tandem with deep-learning approaches to build cognitive services that can support the pharmacovigilance professional in ICSR intake and assessment without the creation and maintenance of dictionaries. Unlike previous studies, this approach used the heterogeneity of source documents received globally to build ten cognitive services. The cognitive services, if implemented, will continue to improve over time on the basis of a controlled feedback loop to ensure processing consistency and quality. This can be achieved by performing incremental training of machine-learning models on samples of new datapoints selected by analyzing and identifying the error patterns of the outputs [18–20]. Incremental training approaches select the regions in the deep-learning networks to be adapted and train them to account for the new datapoints.

5.2 Resource Limitations

The major limitations faced in building an annotated corpus for use in pharmacovigilance is the time, effort, and cost of producing high-quality data. Future approaches may seek to leverage transcription technologies, such as optical character recognition, for automated transcription of documents into a machine-readable format necessary for machine-learning. An automated approach would eliminate the time spent creating blank templates for the variety of structured and unstructured forms found in the annotated corpus. Additionally, considering the trend toward fully electronic medical records, the popularity of wearable medical devices that transmit data directly from a patient, and the widespread preference for email over postal mail, it is likely that the number of handwritten ICSRs requiring transcription will decrease over time.

5.3 Data Limitations and Constraints

Because of the use of patient reports for this project, data privacy has been another constraint on quickly completing the annotated corpus. ICSRs from across the globe adhere to different countries' guidelines for protecting patient privacy. Many documents encountered in the transcription and annotation workflows contained redacted information, making it impossible to train models using those data (e.g., patient initials, name, address, date of birth).

Varying and limited data also posed a limitation when building the annotated corpus. Periodically updated legislation, upgrades to the WHO-DD and MedDRA dictionaries, and alterations to company conventions pose challenges to the creation and maintenance of the annotated corpus over time. Data elements that have been particularly difficult to obtain include rare AEs, uncommon seriousness criteria (e.g., congenital anomaly), and the identification and assessment of DDIs in these reports.

5.4 Effort and Benefit of Cognitive Service Development

Although effortful, the creation of an annotated corpus is a one-time task that will enable a company to build and leverage cognitive services to meet their regulatory obligations in an environment of ever-increasing ICSRs. Between 2004 and 2015, the US FDA has seen a 477.2% increase in number of total reported drug-related AEs [21]. Rather than continually increasing resources and headcount, a recent trend, beginning around 2009, is for companies to outsource their less complex pharmacovigilance functions. As of 2016, about 50% of pharmacovigilance activities across industry are externally outsourced [7]. Outsourcing is a finite solution, however, which has led many companies to look to the benefit of automation and other emerging technologies to handle big data.

The benefit of automation has been estimated in previous research, which has shown that an ICSR classifier, when used to classify valid ICSRs from a set of social digital media data, can outperform humans in processing time. The ICSR classifier demonstrated success in reviewing 311,189 unannotated social media posts in less than 48 h; the estimate for humans to process this volume is 44,000 work hours [8]. While it is difficult to estimate exactly how much time would be saved across a safety team with the support of the ten cognitive services in this study, it is reasonable to state that their time spent searching for datapoints and making decisions will lessen, and, as such, a company will need to allocate fewer resources, both internally and externally, to meet the demand of manual data collection as is done today. Furthermore, the cognitive services will eliminate the required maintenance of finite dictionaries for automatic coding that are often used in industry today. Semantic and syntactic understanding will replace synonym lists and will account for the variety of ways in which ICSRs may be reported. These models will allow for a solution that is volume-, style-, and vocabulary-agnostic in its ability to understand, extract, and classify all data elements important to pharmacovigilance. Future work will seek to quantify the hypothesized additive beneficial effect of implementing ten cognitive services once they have been fully integrated into the case processing workflow.

5.5 Future Work

Given the favorable results of this study in creating ten cognitive services for initial spontaneous ICSR intake, the pharmacovigilance SMEs have begun identifying additional cognitive services to support the processing of study and literature ICSRs. By designing cognitive services that may assist in more efficient data collection, pharmacovigilance professionals will have more time to apply their knowledge in assessing the case rather than spending it performing numerous transactional activities to simply capture the pertinent data within a safety database.

While the focus of this research is on the creation of foundational cognitive services that identify, extract, and classify data elements for optimized data collection, these technical applications may prove valuable across the drug safety workflow. Future direction relating to machine-learning implementations should include cognitive services to support the submission, assessment, and safety evaluation of ICSRs. Because it is estimated that 21% of prescriptions are given for off-label indications, continued training of AE detection from ICSRs in tandem with the product insert is another objective [14]. Deeper understanding of DDIs, linkages between *International Classification of Diseases* codes and events, and linking temporal relationships across multiple source documents are also under consideration for future work and enhancements.

The incorporation of social media into the annotated corpus will support mining for AEs that are not directly reported to the manufacturer. Twitter is a commonly used platform for data exchange, boasting over 500 million posts per day [22]. With this amount of untapped data, we begin to solve the problem of underreported events, as it is estimated that no more than 5% of serious ADRs are actually reported [23]. Some challenges with utilizing social media have been reported previously, such as the identification of medical keywords, mapping of drug–event relationships, classification of seriousness, data privacy considerations, determination of report validity, and overall quality of information provided [22, 23]. In the same vein, mobile technology such as health applications and wearable devices can transmit patient information directly from the source. Not only would use of mobile technology increase sources of information, it would also create a framework in which patients understand the significance of voluntarily reporting their health data [3]. Recent studies have demonstrated success in using machine-learning to mine social media for valid cases before SME review [8]. These emerging capabilities together would provide a comprehensive safety system that could handle the imminent influx of fresh case types and previously untapped data sources.

The intention of augmented intelligence is not to replace the pharmacovigilance professional, but rather support them in their consistent decision-making, so that in time a more consistent and cleaner dataset is available for more robust and timelier signal detection to better inform the actions taken based on those signals. The use of automation is the beginning of the transformation of drug safety, leading to better oversight of products, more proactive approaches to pharmacovigilance, and better understanding of the risk–benefit profiles of medicinal products for enhanced patient safety.

6 Conclusion

In this paper we demonstrate how we curated an annotated corpus in a scalable manner, while maintaining data quality, to train cognitive services such that they achieve a target accuracy level effective in a real-world setting. Ten cognitive services were identified based on their importance in ICSR intake and processing, and include five services relating to validity assessment, seriousness classification, reporter causality classification, expectedness classification, and WHO-DD and MedDRA coding. All cognitive services under development have achieved the minimum evaluative threshold to be considered adequately trained, demonstrating how machine-learning and NLP techniques together provide accurate outputs that may augment pharmacovigilance professionals' processing of spontaneous ICSRs quickly and accurately if implemented into the case processing workflow. Through this support, pharmacovigilance professionals may be better equipped to handle the increasing volume of case reports. More consistent and clean data collation may result in improved signal detection over time, ushering in a new era of pharmacovigilance as a patient-centered, value-based function for patients globally [24].

Compliance with Ethical Standards

Funding This study was financially supported by Celgene Corporation and IBM. Open access was funded by Celgene Corporation.

Conflict of Interest Danielle Abatemarco, Sameen Desai, Bruno Assuncao, Niki Tetarenko, Karolina Danysz, Ruta Mockute, Mark Widdowson, Nicole Fornarotto, Sheryl Beauchamp, Salvatore Cicirello, and Edward Mingle were employed by Celgene Corporation at the time this research was completed. Sujan Perera and Sheng Hua Bao were employed by IBM at the time this research was completed.

Ethics Approval All human subject data used in this research was stored and shared securely.

References

1. Moro A, Invernizzi N. The thalidomide tragedy: the struggle for victims' rights and improved pharmaceutical regulation. Hist Cienc Saude Manguinhos. 2007;24(3):603–22. https://doi.org/10.1590/s0104-59702017000300004.
2. Kim JH, Scialli AR. Thalidomide: the tragedy of birth defects and the effective treatment of disease. Toxicol Sci. 2011;122(1):1–6. https://doi.org/10.1093/toxsci/kfr088.
3. Beninger P, Ibara MA. Pharmacovigilance and biomedical informatics: a model for future development. Clin Ther. 2016;38(12):2514–25. https://doi.org/10.1016/j.clinthera.2016.11.006.
4. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from

the literature. J Biomed Inform. 2014;52:293–310. https://doi.org/10.1016/j.jbi.2014.07.011.
5. World Health Organization. Pharmacovigilance: ensuring the safe use of medicines. Geneva: World Health Organization; 2004.
6. US Food and Drug Administration. Preventable adverse drug reactions: a focus on drug interactions. https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/DrugInteractionsLabeling/ucm110632.htm. Accessed 14 Dec 2017.
7. Holm-Petersen. Digitally transformed pharmacovigilance. http://www.navitaslifesciences.com/collaterals/Whitepapers/WP-Digitally_transformed_pharmacovigilance.pdf. Accessed 5 Aug 2018.
8. Comfort S, Perera S, Hudson Z, Dorrell D, Meireis S, Nagarajan M, et al. Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media. Drug Saf. 2018;41(6):579–90. https://doi.org/10.1007/s40264-018-0641-7.
9. Polepalli Ramesh B, Belknap SM, Li Z, et al. Automatically recognizing medication and adverse event information from Food and Drug Administration's Adverse Event Reporting System narratives. JMIR Med Inform. 2014;2(1):e10. https://doi.org/10.2196/medinform.3022.
10. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform. 2010;160(Pt 1):739–43.
11. Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: 54th Annual Meeting of the Association for Computational Linguistics; 7–12 Aug 2016; Berlin. http://www.aclweb.org/anthology/P16-1101. Accessed 14 Aug 2018.
12. Kim Y. Convolutional neural networks for sentence classification. In: 19th Conference on Empirical Methods in Natural Language Processing (EMNLP); 25–29 Oct 2014; Doha. http://www.aclweb.org/anthology/D14-1181. Accessed 17 Aug 2018.
13. Pustejovsky J, Stubbs A. Natural language annotation for machine learning. Sebastopol: O'Reilly Media, Inc.; 2013.
14. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf. 2017;40(11):1075–89. https://doi.org/10.1007/s40264-017-0558-6.
15. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005;12(4):448–57. https://doi.org/10.1197/jamia.M1794.
16. Névéol A, Zweigenbaum P. Clinical natural language processing in 2015: leveraging the variety of texts of clinical interest. Yearb Med Inform. 2016;1:234–239. 10.15265/IY-2016-049.
17. Ho TB, Le L, Thai DT, Taewijit S. Data-driven approach to detect and predict adverse drug reactions. Curr Pharm Des. 2016;22(23):3498–526.
18. Alpaydin E. Introduction to machine learning. Cambridge: MIT Press; 2004.
19. Gepperth A, Hammer B. Incremental learning algorithms and applications. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN); 27–29 Apr 2016; Bruges. https://hal.archives-ouvertes.fr/hal-01418129/file/article.pdf. Accessed 17 Aug 2018.
20. Sarwar S, Ankit A, Kaushik R. Incremental learning in deep convolutional neural networks using partial network sharing. 2017. https://arxiv.org/pdf/1712.02719.pdf. Accessed 17 Aug 2018.
21. Wynn M, Fauber J. Analysis: reports of drug side effects increase fivefold in 12 years. 19 Mar 2017. https://www.jsonline.com/story/news/investigations/2017/03/17/analysis-reports-drug-side-effects-see-major-increase/99211376/. Accessed 5 Aug 2018.
22. Eshleman R, Singh R. Leveraging graph topology and semantic context for pharmacovigilance through twitter-streams. BMC Bioinform. 2016;17(Suppl 13):335. https://doi.org/10.1186/s12859-016-1220-5.
23. Bousquet C, Dahamna B, Guillemin-Lanne S, Darmoni SJ, Faviez C, Huot C, et al. The adverse drug reactions from patient reports in social media project: five major challenges to overcome to operationalize analysis and efficiently support pharmacovigilance process. JMIR Res Protoc. 2017;6(9):e179. https://doi.org/10.2196/resprot.6463.
24. Smith MY, Benattia I. The patient's voice in pharmacovigilance: pragmatic approaches to building a patient-centric drug safety organization. Drug Saf. 2016;39(9):779–85. https://doi.org/10.1007/s40264-016-0426-9.

Affiliations

Danielle Abatemarco[1] · Sujan Perera[2] · Sheng Hua Bao[2] · Sameen Desai[1] · Bruno Assuncao[1] · Niki Tetarenko[1] · Karolina Danysz[1] · Ruta Mockute[1] · Mark Widdowson[1] · Nicole Fornarotto[1] · Sheryl Beauchamp[1] · Salvatore Cicirello[1] · Edward Mingle[1]

✉ Danielle Abatemarco
Dabatemarco@celgene.com

[1] Celgene Corporation, 86 Morris Avenue, Summit, NJ 07901, USA

[2] IBM Watson Health, 75 Binney Street, Cambridge, MA 02142, USA