Pharmaceutical Medicine (2019) 33:109–120
https://doi.org/10.1007/s40290-019-00269-0

ORIGINAL RESEARCH ARTICLE

# Artificial Intelligence Within Pharmacovigilance: A Means to Identify Cognitive Services and the Framework for Their Validation

Ruta Mockute[1] · Sameen Desai[1] · Sujan Perera[2] · Bruno Assuncao[1] · Karolina Danysz[3] · Niki Tetarenko[1] · Darpan Gaddam[1] · Danielle Abatemarco[1] · Mark Widdowson[4] · Sheryl Beauchamp[1] · Salvatore Cicirello[3] · Edward Mingle[1]

Published online: 29 March 2019


## Abstract

**Introduction** Pharmacovigilance (PV) detects, assesses, and prevents adverse events (AEs) and other drug-related problems by collecting, evaluating, and acting upon AEs. The volume of individual case safety reports (ICSRs) increases yearly, but it is estimated that more than 90% of AEs go unreported. In this landscape, embracing assistive technologies at scale becomes necessary to obtain a higher yield of AEs, to maintain compliance, and transform the PV professional work life.

**Aim** The aim of this study was to identify areas across the PV value chain that can be augmented by cognitive service solutions using the methodologies of contextual analysis and cognitive load theory. It will also provide a framework of how to validate these PV cognitive services leveraging the acceptable quality limit approach.

**Methods** The data used to train the cognitive service were an annotated corpus consisting of 20,000 ICSRS from which we developed a framework to identify and validate 40 cognitive services ranging from information extraction to complex decision making. This framework addresses the following shortcomings: (1) needing subject-matter expertise (SME) to match the artificial intelligence (AI) model predictions to the gold standard, commonly referred to as 'ground truth' in the AI space, (2) ground truth inconsistencies, (3) automated validation of prediction missing context, and (4) auto-labeling causing inaccurate test accuracy. The method consists of (1) conducting contextual analysis, (2) assessing human cognitive workload, (3) determining decision points for applying artificial intelligence (AI), (4) defining the scope of the data, or annotated corpus required for training and validation of the cognitive services, (5) identifying and standardizing PV knowledge elements, (6) developing cognitive services, and (7) reviewing and validating cognitive services.

**Results** By applying the framework, we (1) identified 51 decision points as candidates for AI use, (2) standardized the process to make PV knowledge explicit, (3) embedded SMEs in the process to preserve PV knowledge and context, (4) standardized acceptability by using established quality inspection principles, and (5) validated a total of 126 cognitive services.

**Conclusion** The value of using AI methodologies in PV is compelling; however, as PV is highly regulated, acceptability will require assurances of quality, consistency, and standardization. We are proposing a foundational framework that the industry can use to identify and validate services to better support the gathering of quality data and to better serve the PV professional.

**Key Points for Decision Makers**

As individual case safety report volumes increase, artificial intelligence can be a means to help mitigate complex decision making for pharmacovigilance professionals.

At various decision points in the PV process, cognitive services were identified and developed to assist pharmacovigilance users. These services were validated using a framework leveraging the Acceptance Quality Limit method, to ensure appropriate performance and quality control.

✉ Ruta Mockute
Rmockute@celgene.com

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

[2] IBM Watson Health, Almaden Research Center, San Jose, CA, USA

[3] Celgene Corporation, Boudry, Switzerland

[4] Celgene Corporation, Stockley Park, UK

## 1 Introduction

The World Health Organization defines pharmacovigilance (PV) as "the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem" [1]. PV is governed by legislation and must not only collect, collate, and evaluate adverse events (AEs) that are reported, but also has the regulatory expectation to evaluate these reports and understand drug-event causality both at the patient and population level.

The pharmaceutical industry has continued to work through growing volumes of AE data. In an 8-year period (1998–2005), the number of serious event reports that the FDA received increased 2.6-fold, and reports of deaths increased 2.7-fold [2]. The increase in volumes is faced globally, and despite these growing numbers there is still extensive underreporting of spontaneous AEs. A strategy to overcome underreporting is to explore and mine new data sources. One such potential area could be social media and health-related social networks, which many adults are using to discuss health information. Applications such as Twitter, have several hundred million users; within this application and other health-related social networks, users often discuss their health-related experiences, such as the use of prescription drugs, side effects, and treatments, which makes social networks unique sources of patient health information. Although unique, AEs through these channels are often missing information, are brief, lack structure, and use informal language [3]. Further research from Dr. Ola Caster et al. suggests that using Social Digital Media doesn't appear to add anything to what is already known from spontaneous and clinical trial AEs [4].

A more promising external data source could be electronic health records (EHRs), which generally provide more holistic and thorough representations of patient health that can include clinical narratives. This source of data can help improve AE detection through additional information within the narrative texts, such as symptoms, disease status, severity, and confounding factors [5]. Although EHRs are considered a robust source of health information, it is estimated that only 1% of AEs within EHRs are reported to federal databases [6]. These data sources have differing value for PV, nevertheless, they represent additional, and expanding, stores of knowledge that could house the presence of impactful AEs.

This growing body of available digitized data is coupled with regulatory initiatives that are expanding the range of activities that fall within the remit of PV, for example, the trend to analyze more real-world data that is available in the public domain [7]. Despite the environment and the shifting regulatory initiatives, integration of these data sources would likely disrupt the traditional methods of spontaneous or clinical reporting [8].

To manage the increase in AE data thus far, the pharmaceutical industry has been scaling operations by leveraging a combination of increasing human resources and outsourcing; however, the transcription and data entry tasks required to process these data remain largely manual in nature.

There is a need to identify assistive technologies that provide the automation of repetitive tasks involved with the collection and collation of AEs, as well as providing support and evidence to enhance complex decision making within PV. New technology options should be able to automate mundane activities, harness and provide a synthesized view of the growing amount of data, and provide evidence of recommendations to a PV professional.

We propose that using artificial intelligence (AI) can reduce the manual effort associated with transcription and data entry to allow greater focus on scientific and medical evaluation of AEs, work that ultimately brings greater value to the patient.

## 2 Objectives

The aim of this study was to identify areas across the PV value chain that can be augmented by cognitive service solutions using the methodologies of contextual analysis and cognitive load theory. It will also provide a framework of how to validate these PV cognitive services leveraging the acceptable quality limit (AQL) approach.

## 3 Methods

### 3.1 Background

AI is a subfield of computer science in which a computer system is taught to perform tasks that normally require human intelligence. Natural language processing (NLP) is the ability of a computer system to understand and interpret human language. Machine learning is an area of AI that gives computer systems the ability to learn without explicitly being programed. Cognitive services, are the combination of both NLP and machine learning algorithms that aim to solve specific tasks that would otherwise require human intelligence. In order to develop cognitive services, an annotated corpus, or data used to teach the cognitive service, must be prepared and created. These terms can be referenced within the glossary.

## 3.2 Glossary

Annotated corpus: the data used to teach a cognitive service the syntactic and semantic patterns of a language.

Artificial intelligence (AI): is a subfield of computer science in which a computer system is taught to perform tasks that normally require human intelligence.

Cognitive services: the combination of both natural language processing and machine learning algorithms that aim to solve specific tasks that would otherwise require human intelligence.

Machine learning: a subfield of AI that learns patterns from data without explicitly being programmed

Natural language processing (NLP): the ability of a computer system to understand and interpret human language.

## 3.3 Identification of Cognitive Services

The research began with identifying cognitive services; the cognitive services focused on were related to the intake, collection, and collation of safety information. The identification process can also extend to services outside of this scope of work. A critical decision was how to best leverage machine learning algorithms to develop cognitive services that would be beneficial for the end user within receipt, triage, data entry, and assessment steps of individual case safety report (ICSR) processing. To achieve this, a user-centered design method called a contextual analysis was used to research the end users. A contextual analysis consists of an interviewer observing and interviewing a user about their role, breaking down their functions into the tasks being performed, and the PV decisions that were being made. Some of the benefits of contextual analysis were insights into the behavioral aspect of an end user, an understanding of the issues an end user faced, and why those issues existed. This knowledge helped narrow the scope of the cognitive services that were being developed by designing technologies to diagnose the underlying issues rather than the effects of those issues.

Following this, the tasks that the PV professionals were performing were analyzed utilizing cognitive load theory. Cognitive load theory illustrates and categorizes cognitive load, or the effort that it takes to commit something to working memory. There are three types of cognitive load: (1) intrinsic, the cognitive load imposed by the characteristics of information; (2) extraneous, the cognitive load imposed on how the information is presented to the user; and (3) germane, the cognitive load effort that contributes to the construction of schemas, or patterns of organizational thought [9]. Using both contextual analysis and cognitive load theory the cognitive services for PV use were identified.

## 3.4 Development of the Cognitive Services

### 3.4.1 Specification of the Annotated Corpus

After identification of the cognitive services, the corpus of training data was planned. An annotated corpus is the data used to teach a cognitive service the syntactic and semantic patterns of a language to help identify and extract the data points of interest to the PV process. Volume is necessary to train a successful cognitive service, so 20,000 ICSRs, (approximately 50,000 source documents), consisting of initial and follow-up cases, were selected from Celgene Corporation's Global Drug Safety database records from the years 2015–2016. The ICSRs selected needed to be both diverse and representative of the data and incorporate factors ranging from (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 the Investigator's Brochure (IB), Company Core Data Sheet (CCDS), Summary of Product Characteristics (SmPC), and Prescribing Information (PI) into consideration. The sampling strategy ensured appropriate diversification and representation of possible values for each factor. This approach resulted in a corpus whose report type broke down into 63% spontaneous, 27% clinical/market study, and 10% medical literature ICSRs; 105,397 total unique reported terms; and whose ICSR seriousness broke down into 50% serious and 50% non-serious ICSRs.

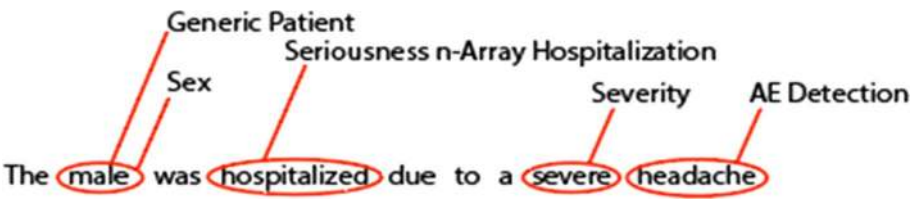


**Fig. 1** Annotated sentence. An example of an annotated sentence, in which PV concepts are labeled by their relevant annotations. Concepts can have multiple annotations as long as they fall within the annotation definitions

### 3.4.2 Building and Allocating an Annotated Corpus

Once the corpus is specified, it must then be prepared into electronic format, and its relevant data labeled. Documents were made electronic, or machine readable, by manual transcription, and then all appropriate metadata were tagged in a manual annotation process. An annotation is labeled metadata, to ensure consistent annotations across documents and users a standardized PV annotation dictionary was created. This dictionary consisted of a breakdown of 122 PV concepts and information ranging from regulatory clock start date to reporter causality and served as a way to make PV knowledge explicit to the cognitive services. To see an example of a sentence with PV specific annotations see Fig. 1.

Once the annotated corpus was developed, the training set was allocated into different groups for training (80%), tuning, or model refining (10%), and testing (10%) the cognitive services. The purpose of the training data is to teach the cognitive services, the tuning data to optimize the parameters of those services, the testing data to create a feedback loop for errors and evaluate the services in a real-world setting [10]. Without subdividing the annotated corpus into independent groups, the services would be measured on data that they had been previously exposed to, thus rendering the predictive performance overly positive. As it would in practice, a true performance understanding would measure how the service would perform on new data [10].

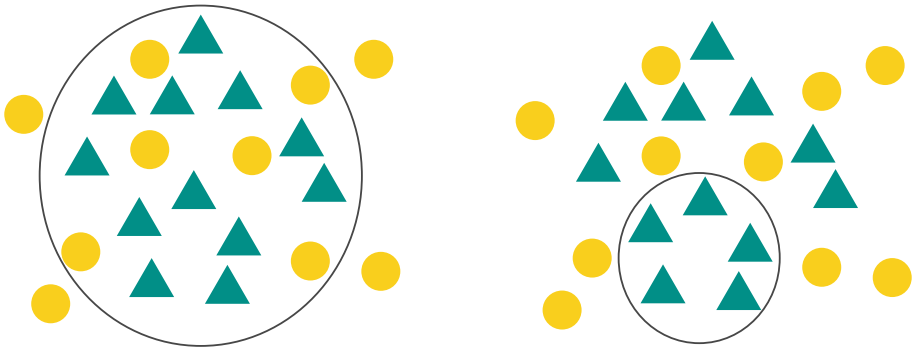

**Fig. 2** Recall vs precision. For this example the cognitive service's purpose is to identify triangles. The figure on the left would indicate a service with high recall, because it is identifying all of the triangles; however, it identifies some circles as triangles as well. The figure on the right would indicate a service with high precision, in that it is not identifying all of the triangles that exist, but the elements that it is identifying as triangles, are correct

**Fig. 3** Calculating $F_1$ score. Delineation of how the $F_1$ score is measured and a visual representation of the parameters are for true positives, false positives, false negatives, and true negatives

$$F_1 = 2\frac{Precision\ x\ Recall}{Precision + Recall}$$

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

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

| Actual | Cognitive Prediction: Positive | Cognitive Prediction: Negative |
|---|---|---|
| Positive | True Positive | False Negative |
| Negative | False Positive | True Negative |

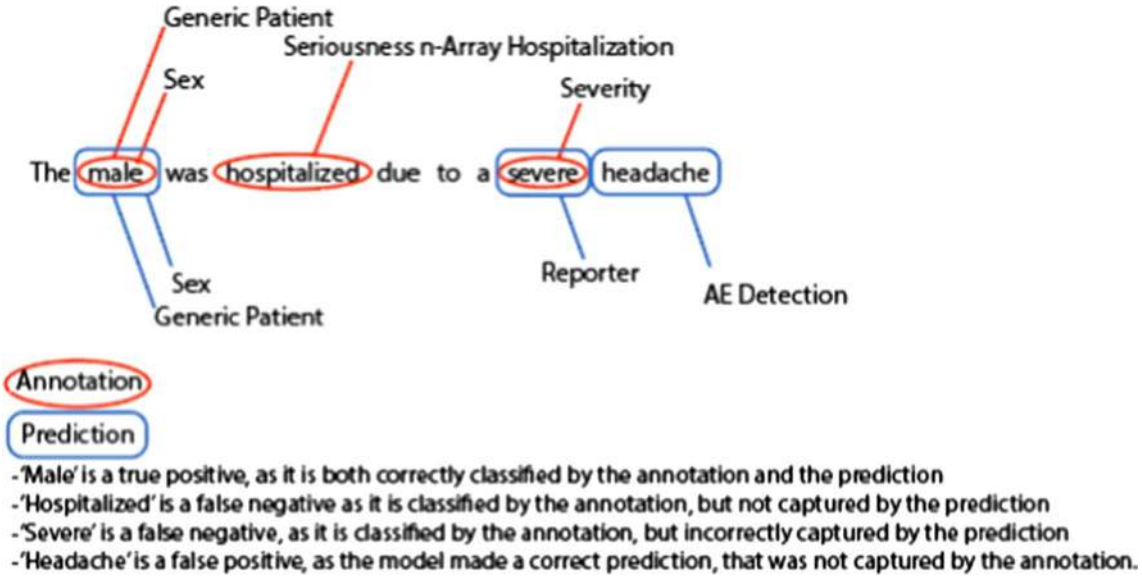


**Fig. 4** Comparison of annotations and cognitive service predictions. A holistic view of how annotations and predictions are compared during the review process and classified on the basis of their accuracy

#### 3.4.3 Measuring Performance

A cognitive service was considered successfully trained when it reached an $F_1$ score of 75% or higher. This evaluative score was the pre-determined minimum threshold for each cognitive service to be effective in a real-world setting. An $F_1$ score is the combined measure of both precision and recall and is a common measure for evaluating machine learning algorithms [11, 12]. Precision, also referred to as positive predictive value (PPV), is the ability of a service to correctly identify elements. The risk of having a very high precision is that the service may not capture all of the correct elements, but those elements it does capture will be captured correctly. This translates to the service as having many false negatives (FNs), or elements that should have been identified but were not predicted. On the other hand, recall, also referred to as sensitivity, is ensuring that the totality of results is identified correctly. The shortcoming of having a very high recall rate is that although the service may classify all of the instances of identifying an element, it may classify some incorrectly. A high recall will run the risk of many false positives (FPs), or elements that were predicted by the service that should not have been [13], see Fig. 2 for a visual representation of how high precision and high recall differ in practice. A cognitive service must therefore have a balance of both precision and recall to be truly effective. True positives (TPs) are entities that are predicted correctly or elements that are predicted positive and are actually positive, and true negatives (TNs) are elements that are labeled as negative and are actually negative; refer to Fig. 3 for how to calculate $F_1$, TP, TN, FP, and FNs; and refer to Fig. 4 for an annotated example indicating a TP, FP, and FN.

During training development, the PV SME would review 100% of the FP and FN results from the testing data, and for binary classification services (e.g. ICSR detection), all of the TN results as well. The review process would entail the SME referencing the annotations, metadata, or original source documents used to train the cognitive service and providing feedback to the developer as to whether the predicted FP, FN, or TN results were correctly classified. Often, the data showed trends that the SME could link back to industry rules or company-specific guidance, which the developers would use to help further train the service. This was done to accurately reflect what the measurement of the service was, and to decrease the chances of the services failing during the cognitive service approval process.

### 3.5 Validation of Cognitive Services

#### 3.5.1 What is the Framework?

As multiple cognitive services were developed to the desired threshold, there needed to be a process to validate and ensure the services' quality within a regulated environment. To achieve this, a framework was created to have a structured and consistent approach to quality assurance. This framework utilized the Acceptable Quality Limit (AQL) method, a sampling method that is commonly used in manufacturing. Our framework is a derivative of the AQL method and it is customized to address the nuances of a PV environment.

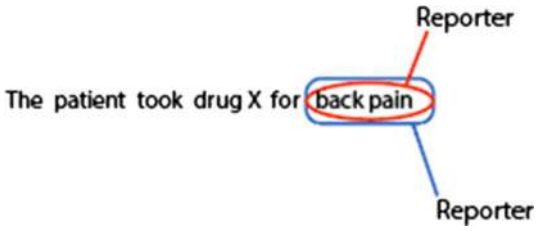


**Fig. 5** Review of an incorrect true positive. An example of an incorrect true positive wherein both the prediction and annotation classified back-pain incorrectly

### 3.5.2 Why Did We Use Acceptable Quality Limit?

Validation has been the means by which stakeholders ensure that processes or products are performing at the quality level claimed for them [14]. Currently, there are no regulatory guidelines that delineate how AI should be validated before being introduced into a PV environment. There were high volumes of cognitive service outputs, so to be practical and scalable, a sampling approach to quality review was taken. GMP regulations provide a number of requirements for sample sets of validation: samples must represent the batch under analysis, the sampling plan must result in statistical confidence, and the batch must meet its predetermined specifications [15]. One of the sampling methods that meets these requirements and is recognized by the US Food and Drug Administration, was the American National Standards Institute/American Society for Quality Z1.4 standard, or the AQL method. AQL is used when a buyer is inspecting batches of goods delivered by a supplier. In its entirety, AQL defines the maximum number of components that can be rejected for the buyer to accept the whole batch [16]. It was adopted for our study because the method is scalable, reproducible, and customizable, and can be done without complicated statistics. In our study, we have used this concept to confirm, through inspection based on AQL sampling, that the cognitive service predictions met the acceptability criteria after they achieved the threshold $F_1$ score of 75% or higher.

### 3.5.3 Determination of AQL Parameters

To apply AQL, one of the first decisions to make is to define what needs to be measured. TPs were the measurement used to determine if the cognitive services had high quality outputs. TPs were the combination of both the cognitive service prediction and the ground truth, or gold standard annotation correctly identifying an output. A TP result was incorrect when both the prediction and ground truth did not correctly classify the PV concept, for an example, see Fig. 5.

The next decision to make is to define the AQL, or the tolerance for nonconforming parts. This was the worst quality of cognitive services that would be considered acceptable.


**Fig. 6** Defining acceptable quality limit (AQL) parameters. The step by step process of defining AQL parameters, in order to perform quality review

Current guidance outlines the defect attributes, their categories, and their associated tolerance percentages as (1) critical defects (0%), defects that can pose a threat to user safety or cause a product to be unusable; (2) major defects (2.5%), defects that would result in products being returned because they adversely affect product performance but do not pose a safety risk; and (3) minor defects (4%), defects that are unacceptable at high levels but are generally small or insignificant issues that can be fixed [16]. Our cognitive services were developed in tandem with PV professionals, and the services were only reviewed for quality once they achieved a minimum $F_1$ score of 75% within the training phase. Because of the integration of the SME within the process and the reviews occurring once the service was performing at a certain accuracy, the decision was to classify defective outputs as minor, with a 4% acceptance quality threshold.

After defining the tolerance, AQL requires an appropriate inspection level to be chosen. The options for general levels of inspection are levels I, II, and III. For most instances, level II is the recommended use, but level I may be used when less discrimination is needed, such as in the case of a vendor with a positive history, and level III may be used when quality needs to be more stringent. There are four special levels (S1 through S4) that can also be used when sample sizes are required to be small and sampling risks can

be tolerated [16, 17]. Because we had no previous history with these services, general inspection level II was deemed best for our purposes.

The last parameter to consider is lot size. For each cognitive service, the lot size was determined by the number of TPs that existed within the annotated corpus. The lot size varied for each cognitive service on the basis of the prevalence of annotations and commonness of a concept. For review, the entirety of the TPs for a single cognitive service were randomized to remove any potential bias. After randomization, an appropriate sample size was selected on the basis of the lot size, inspection level, and AQL percentage using the AQL graph. In the instances in which the amount of data was small and the total number of TPs was less than 150, a 100% review of the TPs was conducted instead of using the AQL method, using the same 4% acceptance threshold. This was based on the need to ensure quality of the cognitive services and was also deemed within the work capacity of the team. To view a visual representation in how to determine AQL parameters refer to Fig. 6.

#### 3.5.4 Execution of the Framework

During quality review, the predicted TPs were assessed and marked as either correct or incorrect. Quality review of TPs was performed by PV professionals who received an excel output of the cognitive services' TP, FP, FN, TN classification results. To ensure high quality review and to support decision making, the PV professionals also had access to the annotated corpus replete with annotations, and the metadata of the annotated corpus pulled from the legacy database. To expedite the review process, the surrounding verbiage from which the prediction was made was provided to help give context. If the cognitive service TP errors did not exceed the rejection limit, the cognitive service was approved. If the service failed, it was returned to the developer for retuning, after which an additional AQL was performed. Before defining and using this framework for our quality review purposes, a 100% review of TPs was conducted. This was determined to be too time consuming, so a statistical sampling approach was adopted.

## 4 Results

As a result of this research, 51 decision points were identified as candidates for AI use that could help PV professionals in their decision making. A framework for validation has been developed resulting in the validation of 43 cognitive services for spontaneous reports, 45 cognitive services for clinical trial cases, and 38 cognitive services for medical literature cases, for a total of 126 validated cognitive services.

The 51 decision points that were identified were determined through contextual analysis and classifying the tasks identified at contextual analysis into their respective cognitive loads. Within the ingestion and data collation steps, intrinsic cognitive load was identified at several points when (1) ICSRs had to be routed and classified on the basis of their inherent details, and identified as to whether they needed follow-up by referring to their contents; (2) at the data collection step, case information had to be manually collected from multiple sections of source documents and entered into an external database; (3) at coding, PV professionals required extensive training and a thorough understanding of taxonomy; (4) strict regulatory timelines had to be adhered to and understood; (5) at triage, safety information needed to be classified, and demanded a medical background to evaluate key features of a case.

This process also identified extraneous cognitive load, specifically during (1) the intake of ICSRs, where information arrived in a multitude of formats and needed to be assessed and processed for quality; (2) ingestion when information had to be identified as duplicate or not, and any information that was entered into the PV database had to be either unique or correctly linked to its associated data; (3) prioritizing and processing the perpetual influx of cases that needed to be both processed and continuously prioritized to ensure that timelines were met and workload was balanced.

After evaluation of the cognitive burden, decision points where AI could assist the user were identified based on the need or opportunity to increase efficiency or to decrease the cognitive load. The first high-level cognitive services identified were: ICSR validity service, suspect product detection, reporter detection, patient detection, seriousness classification, World Health Organization Drug Dictionary (WHO-DD) coding, Medical Dictionary for Regulatory Activities (MedDRA) coding, expectedness classification, reporter causality, and drug mention detection. As the study progressed, additional cognitive services were identified (listed in Fig. 7) and are grouped into the PV concepts of adverse event, reporter, patient, case, and product.

In adopting the AQL methodology into our validation process to maintain quality control, the result has been the development of a framework that addresses specific PV needs (Fig. 8). This framework can be replicated and used to validate new PV cognitive services that have yet to be developed and is advantageous in that it is: (1) a consistent approach to quality assurance; (2) a scalable, sampling approach; (3) reproducible in that it was used for a variety of cognitive services, report types, users, and data; (4) customizable based on desired inspection and tolerance; (5) meets good manufacturing practice (GMP) regulatory requirements for sample sets of validation; and (6) integrates

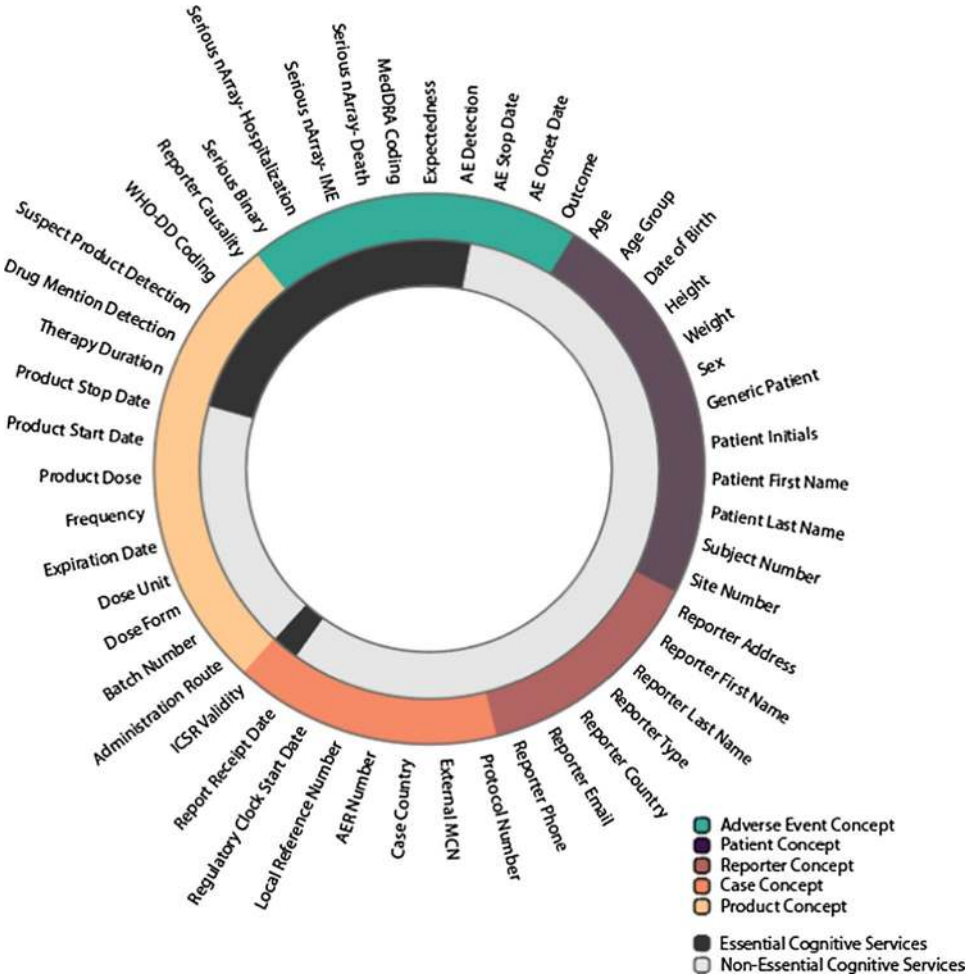


**Fig. 7** These were the 51 decision points that were identified and then developed into cognitive services, and are grouped into their associated concepts marked by the outer ring. The decision points were also categorized into essential vs non-essential by assessing the tasks the end user was performing, and identifying which tasks imposed higher intrinsic and extrinsic cognitive load. The 'essential' decision points that became the 'essential' cognitive services are indicated by the inner ring in black. Adverse event (AE); Manufacturer control number (MCN); Adverse event report (AER); Individual case safety report (ICSR)

the PV professional throughout the identification and validation process.

As an overarching view of our research, key milestones and accomplishments that have been reached include (1) an establishment of a process that identified 51 PV decision points that are candidates for AI; (2) a standardized identification and labeling dictionary of PV information, serving as a means to make PV knowledge explicit; (3) the incorporation of PV expertise throughout the development and validation process by the inclusion of the PV SME at every step; (4) a standardized validation process of cognitive services using established quality principles originating from manufacturing; and (5) validated a total of 126 cognitive services.

## 5 Discussion

AI is becoming increasingly used throughout the healthcare industry, and it has been seen in the increasing uses of NLP and machine learning to automatically detect AEs and drug-drug interactions. As there has been limited success in these endeavors due to reliance on keywords, or the limitations of medical dictionaries, many opportunities still exist to discover the full extent to which AI can be introduced as a support structure to augment and empower the PV professional [6, 18].

### 5.1 Ground Truth Inconsistencies and Quality Considerations

As this research was conducted, it became apparent that for the cognitive services to have high quality predictions, the data within the corpus had to be of high quality as well. However, the data within the corpus, or the ground truth, could vary because of the innate way PV information was ingested and received. The annotated corpus was reflective of the real-world data and had representation from a variety of sources and channels, thus, the ground truth could vary because of (1) spacing (whether the annotation included the space before or after the correct ground truth); (2) misspellings; (3) line breaks (e.g. when information was dispersed throughout a document); (4) limitations in the annotation dictionary; and (5) annotator and reviewer bias. Because of these inconsistencies, the

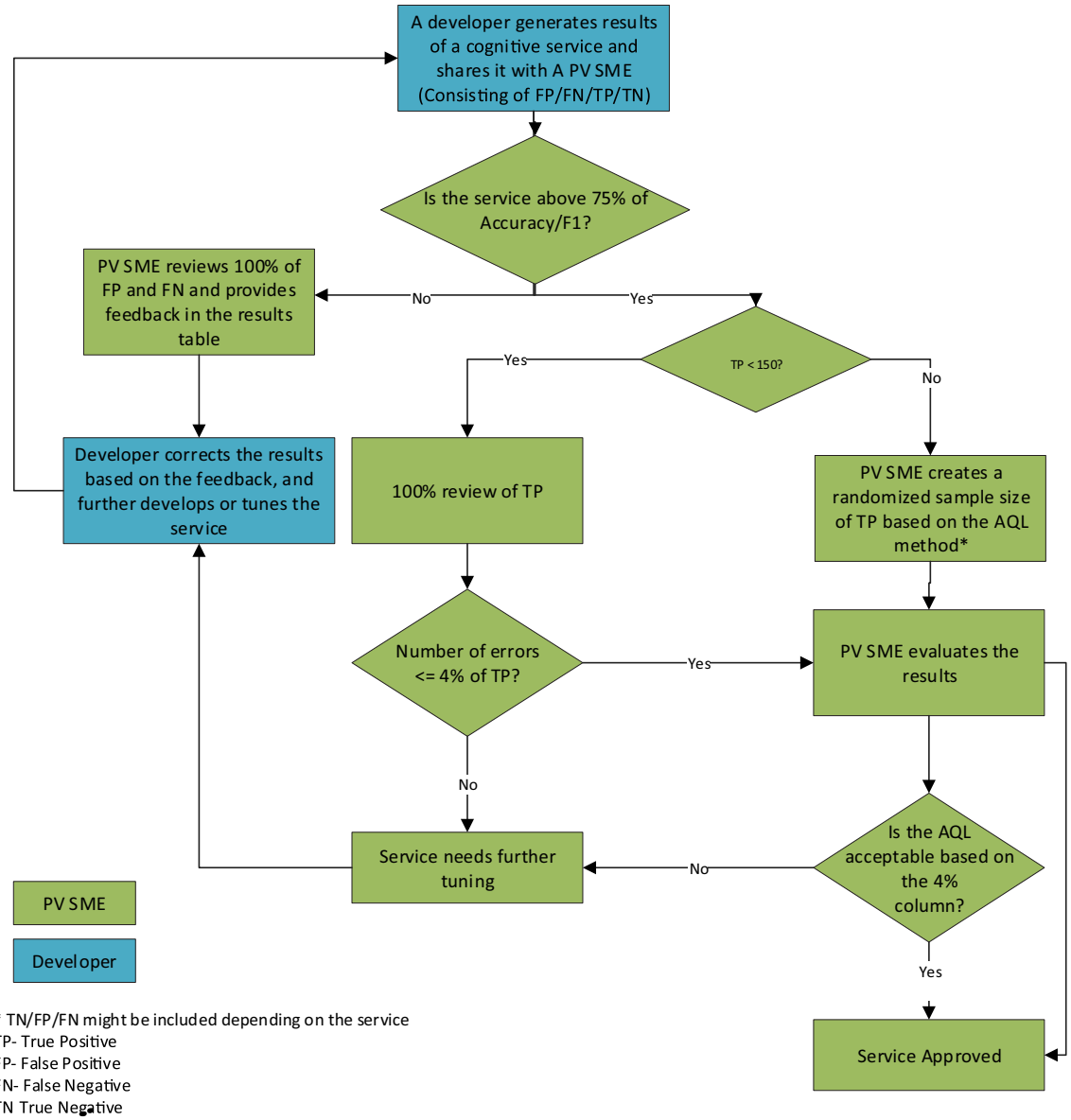


**Fig. 8** Acceptable quality limit (AQL) process for pharmacovigilance cognitive services. This process depicts the framework for the validation of cognitive services leveraging the AQL method. It was customized in a way to accommodate for the inherent needs of pharmacovigilance (PV). The validation process begins once the developer generates the results of a cognitive service and creates an excel output of all of the TP, FP, FN, and TN's. If the $F_1$ score is below the 75% threshold, the PV subject matter expert (SME) reviews 100% of the FP and FN results, and reports any trends in errors and results of the review to the developer for further training. If the $F_1$ score is above 75%, the PV SME reviews the TP results to ensure the service is performing at the accuracy claimed. For our purposes, if the number of TPs was less than 150, the PV SME would perform a 100% review of TPs to ensure a high-quality service, as it was within the work capacity of the team. If there were more than 150 TPs, the PV SME would randomize the TPs, select the appropriate AQL sample of TPs, and then review results. For both instances, it the TP error rate was $\leq 4\%$, then the service was deemed passed, and if not, it was sent back to the developer for further training

PV SMEs manual review was fundamental to provide a consistent input of PV knowledge and provide the context and reasoning for the annotations.

Other quality considerations were due to the fact that documents were manually transcribed, and there were limitations to recreating documents based on sizes, tables, and formatting. The annotation process, which was also manual, could have varied and affected the data quality as well. The annotations were periodically redefined, some annotations were retired, while others were added to ensure correct specificity for practical use. Because of this, the dictionary was updated as the study developed to accommodate for these needs and correctly highlight PV concepts. Because of this process, the version of the dictionary used during annotation

could have limited the accuracy, quality, and existence of the annotations and thus the ground truth. Another factor that could have affected data quality is annotator and reviewer bias, which was addressed by having frequent trainings, and tracking inconsistencies among reviewers and annotators to pinpoint trends.

If a cognitive service was unable to be trained successfully due to inconsistent ground truth, the process of either fixing the annotated corpus or the reannotation of new training data was an alternative route. This was generally considered only for high-impact cognitive services because the rework was not only time consuming but ran the risk of being insufficient after the re-annotation. This was especially true if a concept was inherently complex as in the case of cognitive services like regulatory clock start date, which involves many company-specific guidelines.

### 5.2 The Different Sources of Ground Truth

When approving the cognitive services, referencing the ground truth was crucial, but this differed by the type of service and its inherent complexities and needs. The most common way to develop and validate services was using the annotations as ground truth; however, this was not always the most advantageous route. WHO-DD coding was a service that was developed and validated using the metadata from our legacy database as its primary ground truth. This was because WHO-DD coding is an international standard, so the service's predictions came out in this desired format. When the annotations were used as the ground truth, there was a disparity between the predictions, which were all in the standard WHO-DD coding format, and the annotations, which would reflect the raw, initially ingested data that came in a variety of formats and spellings. To address this delta, the reviewer used the metadata, which reflected processed and coded products and therefore allowed the predictions to match the ground truth. By using the metadata as the ground truth, this service was unique in that it was developed independent of its report type and therefore required only one comprehensive round of validation.

Another approach for the source of ground truth was using a combination of both the metadata and the annotations. This approach was used specifically for the seriousness service because the initial output of the service would predict a seriousness independent of its associated event. Therefore, for instances with multiple seriousness criteria and events, the reviewer could not make an accurate judgment on the accuracy of the service. Thus, the metadata were used as a supplemental ground truth when there were multiple seriousness criteria, because this use helped delineate each seriousness with its associated AE.

Conversely, the ground truth could be missing and had to be generated by auto-annotation. One reason for this situation was missing annotations, so "simple rule" scripts were created to generate ground truth. For example, in the patient mention service, the developers created an annotation rule whereby the word "patient" was highlighted. However, this posed a risk of incorrect ground truth; a way to overcome this limitation was to auto-annotate solely in the AE verbatim section, or unstructured text area, instead of the entirety of the document. Another challenge was that the ground truth was sometimes heavily skewed in one direction. We discovered that in developing the reporter country cognitive services that the majority of the training data used were for US-based cases. Because of this finding, we discovered that the service could not predict global countries, and retraining was done to increase the service's exposure to non-US cases during development before proceeding to final approval.

### 5.3 Applications of the Identification and Validation Framework

If the cognitive services are introduced to a production environment, any outputs of the services would be subject to a PV SME's review and feedback. This user feedback would be logged, reviewed, quality checked, and fed back again to the developer to further train and improve the models. This improvement could be achieved by performing incremental training of machine learning services on samples of the new data points selected by analyzing and identifying the error patterns of the outputs [19, 20]. Validating the new versions of the cognitive services would be necessary before the release of every feedback loop. Therefore, this validation framework could be used as the foundation to validate not only the service upgrades in the future, but new cognitive services as well. With respect to the validation framework, adjustments could be made to the inspection level as the services mature and have a positive AQL history. With repeated positive performance, the thresholds of inspection could become less stringent, and fewer resources and less time would need to be allocated for inspection. On the other hand, it there are repeated failures of validation of a service in the future, it would be prudent to use tighter AQL criteria such as an increased sample size or tightened general inspection levels. The tier of retesting should be limited and defined because repeated test failure may result in a service's rejection [21].

An additional consideration with this framework is the inclusion of protected health information and personal identifying information. Additional effort was necessary on our part to accommodate for this because in order to share results, it was essential to have a secure file transfer to an appropriate environment. Initially, there were delays in creating this environment, so sharing results and providing

feedback demanded much collaboration without direct access to results.

In a similar vein to the validation framework, the means for identifying cognitive services can be replicated in other groups or areas that wish to seek to identify areas or tasks that can be supported by AI. These two methodologies could become standards for future identification and validation to occur across industry members to ensure both quality and consistency in various groups. As automation becomes more prevalent within PV, the focus will decrease on repetitive tasks and data collection, and create more opportunities to concentrate on evolving regulatory requirements and complex healthcare cases. Patient safety is central to PV, and in the space in which AE reports are becoming more frequent and more convoluted, it is imperative that innovation and technology are intertwined to create quality data that will promote patient benefit and health. The PV industry has needed a long-term solution to unsustainable volumes of ICSRs, and by embracing AI it is possible to improve the way we approach PV and strive for excellence in our processes for the benefit of our patients.

## 6 Conclusion

This paper demonstrates the way in which we identified points across the PV value chain that can be augmented by artificial intelligence with the aim of decreasing cognitive burden and supporting efficiencies in various PV processes. There were 51 decision points that were identified across the data ingestion and data collection and collation steps of ICSR case management that covered common PV concepts including patient, reporter, adverse event, case, and product. We also outlined a framework for validating cognitive services through the AQL process such that services were validated in a consistent and reproducible fashion within a regulated environment and could be used as a future standard to validate technologies yet to be developed; to our knowledge, this is the first instance of a validation process of AI within PV. The drive for innovative technologies must continue as PV professionals continue to face challenges of growing case volumes and data consumption. And as we adopt new approaches aimed at enhancing the future of PV, we require not only better data quality and consistency, but ultimately to improve the safety of patients.

**Acknowledgements** A thank you to Sheng Hua Bao for his knowledge and expertise in the creation of this manuscript.

### Compliance with Ethical Standards

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

**Conflict of Interest** Ruta Mockute, Sameen Desai, Bruno Assuncao, Karolina Danysz, Niki Tetarenko, Darpan Gaddam, Danielle Abatemarco, Mark Widdowson, Sheryl Beauchamp, Salvatore Cicirello, and Edward Mingle were all employed by Celgene Corporation at the time this research was completed. Sujan Perera was employed by IBM Watson Health at the time this research was completed.

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



## References

1. WHO Policy Perspectives on Medicines. Looking at the Pharmacovigilance: ensuring the safe use of medicines. Geneva: World Health Organization. http://apps.who.int/medicinedocs/pdf/s6164e/s6164e.pdf. Published October 2004. Accessed 15 Dec 2009.
2. Moore TJ, Cohen MR, Furberg CD. Serious adverse drug events reported to the Food and Drug Administration, 1998–2005. Arch Intern Med. 2007;167(16):1752–9.
3. Sarker A, Ginn R, Nikfarjam A, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform. 2015;54:202–12.
4. Caster O, Dietrich J, Kürzinger ML, et al. Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR Project. Drug Saf. 2018;41:1355.
5. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. JAMIA. 2009;16(3):328–37.
6. Luo Y, Thompson WK, Herr TM. Natural Language Processing for EHR-based pharmacovigilance: a structured review. Drug Saf. 2017;40(11):1075–89.
7. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018. https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf.
8. Beninger P, Ibara MA. Pharmacovigilance and biomedical informatics: a model for future development. Clin Ther. 2016;38(12):2514–25.
9. Sweller J, van Merriënboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educ Psychol Rev. 1998;10(3):251–96.
10. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, CA: Morgan Kaufmann Publishers; 1999.
11. Powers DMW. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37–63.
12. Freitag D. Machine learning for information extraction in informal domains. Mach Learn. 2000;39(2–3):169–202.
13. Olson DL, Delen D. Advanced data mining techniques. Berlin, Germany: Springer; 2008.

14. Dashora K, Singh D, Saraf S, Saraf S. Validation—the essential quality assurance tool for pharma industries. http://www.pharminfo.net. 2005. vol. 3, pp. 45–47.
15. US Food and Drug Administration. Guidance for Industry: Process Validation: General Principles and Practices. https://www.fda.gov/downloads/drugs/guidances/ucm070336.pdf. Published January 2011.
16. ASQ. ANSI/ASQ Z1.4-2003 (R2013): Sampling procedures and tables for inspection by attributes.
17. ISO TC69/SC5–Acceptance sampling: ISO 2859-1:1999—Sampling procedures for inspection by attributes-Part 1: Sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection. 2nd ed. Published November 1999.
18. Abatemarco D, Perera S, Bao SH, et al. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. Pharmaceut Med. 2018;32(6):391–401.
19. Gepperth A, Hammer B. Incremental learning algorithms and applications. In: European symposium on artificial neural networks (ESANN). 2016.
20. Sarwar SS, Aayush A, Kaushik R. Incremental learning in deep convolutional neural networks using partial network sharing. arXiv preprint arXiv:1712.02719. 2017.
21. Mathonet S, Mahler HC, Esswein ST, et al. A biopharmaceutical industry perspective on the control of visible particles in biotechnology-derived injectable drug products. PDA J Pharm Sci Technol. 2016;70(4):392–408.