Innovation in Pharmacovigilance: Use of Artificial Intelligence in Adverse Event Case Processing

Juergen Schmider[1],*, Krishan Kumar[2], Chantal LaForest[3], Brian Swankoski[4], Karen Naim[5] and Patrick M. Caubel[6]

Automation of pharmaceutical safety case processing represents a significant opportunity to affect the strongest cost driver for a company's overall pharmacovigilance budget. A pilot was undertaken to test the feasibility of using artificial intelligence and robotic process automation to automate processing of adverse event reports. The pilot paradigm was used to simultaneously test proposed solutions of three commercial vendors. The result confirmed the feasibility of using artificial intelligence–based technology to support extraction from adverse event source documents and evaluation of case validity. In addition, the pilot demonstrated viability of the use of safety database data fields as a surrogate for otherwise time-consuming and costly direct annotation of source documents. Finally, the evaluation and scoring method used in the pilot was able to differentiate vendor capabilities and identify the best candidate to move into the discovery phase.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Case processing activities constitute a significant portion of a pharmaceutical company's internal pharmacovigilance (PV) resource use. Consequently, automation of adverse event (AE) case processing represents a significant opportunity to affect the strongest PV cost driver. Although automation incorporating artificial intelligence (AI) has been used in other industries, the nature of AE case processing is comparatively complex and no providers currently offer a comprehensive AE case processing solution.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Is it viable to use advanced AI tools in the application for AE case processing, specifically the extraction of case critical information from source documents to identify valid AE cases after training the machine-learning algorithms with source documents and database content rather than annotated source documents?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ It is feasible to use AI-based technology to support extraction from AE source documents and evaluation of case validity. In addition, it is viable to train the machine-learning algorithms using the safety database data fields as a surrogate for otherwise time-consuming and costly direct annotation of source documents. Finally, the evaluation and scoring method used in the pilot was able to differentiate vendor capabilities and identify a candidate to move into the discovery phase.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ With proof of concept and identification of a suitable vendor, progression into the discovery phase will explore the application of these machine-learning tools to additional business processes related to intake, processing, and reporting of individual safety cases.

Case processing activities constitute a significant portion of internal pharmacovigilance (PV) resource use, ranging up to two-thirds on the basis of PVNet benchmark data.[1] When additional costs related to outsourcing are taken into account, case processing spending, on average, consumes most of a pharmaceutical company's overall PV budget.

Automation of adverse event (AE) case processing through artificial intelligence (AI) represents an opportunity to affect the strongest PV cost driver. The past decade has witnessed increasing application of AI methods to the field of biomedicine. Some of the recent improvements in leveraging AI techniques against publicly available consumer data have created opportunities for assessing

[1]Pfizer R&D, Lake Forest, Illinois, USA; [2]Pfizer Business Technology, Artificial Intelligence Center of Excellence, La Jolla, California, USA; [3]Pfizer Global Product Development, Safety Solutions, Kirkland, Quebec, Ontario, Canada; [4]Pfizer Finance and Business Operations, Peapack, New Jersey, USA; [5]Pfizer R&D, Collegeville, Pennsylvania, USA; [6]Pfizer R&D, New York, New York, USA. *Correspondence: Juergen Schmider (juergen.schmider@pfizer.com)

the utility of AI techniques with the automation of PV processes.[2] With the emergence of electronic health records, a growing body of research has explored use of machine-learning techniques to develop disease models, probabilistic clinical risk stratification models, and practice-based clinical pathways.[3–6] A considerable number of studies have focused on information extraction, using natural language processing techniques and text mining to gather relevant facts and insights from available, largely unstructured sources, such as drug labels, scientific publications, and postings on social media.[7] Text mining techniques have also been combined with rule-based and certain machine-learning classifiers to demonstrate the possibility of developing effective medical text classifiers for spontaneous reporting systems, such as the US Vaccine Adverse Event Reporting System.[8] Some of the research in this area has been devoted to creation of the annotated source data that are used to develop and test new machine-learning natural language processing algorithms, including a recent study that explored use of machine learning with data crowd sourced from laymen annotators to identify prescription drug user reports on Twitter.[9] Other researchers have focused on developing and/or improving approaches using natural language processing to recognize and extract information from various medical text sources (e.g., detection of medication-related information[10] or patient safety events[11] from patient records; drug–drug interactions from the biomedical literature;[12,13] or disease information from emergency department free-text reports).[14] Natural language processing techniques have also been applied to extraction of information on adverse drug reactions from the growing amounts of unstructured data available from the discussion and exchange of health-related information between health consumers on social media.[6,7,15,16]

Automation has been in use within other sectors, such as the banking and financial industries, from as early as the 1950s (e.g., automated check handling),[17] and has incorporated AI for the past decade (e.g., automated underwriting within the insurance industry).[18] Despite this long history, there are currently no providers who offer a comprehensive AE case processing solution. A key differentiator to other industries is the highly complex nature of AE case processing involving significantly more decision points and adjudications within a highly regulated and audited environment compared with case processing workflows in other industries. In addition, most source documents are merely semistructured or are fully unstructured. At the highest level, AE case processing comprises four main activities, including intake, evaluation, follow-up, and distribution. Each of these four main activities is associated with multiple deliverables, and each of these deliverables is composed of multiple decision points. **Figure 1** provides a simplification to illustrate how the number of decisions increases with increasing depth of analytic scrutiny. Multiple, predominantly manual, business processes are traditionally in place to take in, process, and report out individual safety cases.

Text-based machine learning typically requires training data in the form of annotated source documentation (i.e., direct indication within the document to identify the appropriate text elements and provide the contextual relation within the text). This is a resource intensive process, particularly when hundreds of thousands of text pages have to be revisited and annotated in such a manner.

To simplify the manual preparation, Pfizer intended in this pilot to merely use the safety data extraction from the source documents as it is captured within the Pfizer safety database.

The current pilot was undertaken to prove the viability of commercially offered machine-learning solutions in the application for case processing. The pilot paradigm was used to simultaneously test proposed solutions of three commercial vendors for the ability to extract case critical information from source documents to identify valid AE cases after training the machine-learning algorithms with source documents and database content rather than annotated source documents. Validity was established by the presence of four elements (i.e., an AE (suspected adverse drug reaction), putative causal drug, patient, and reporter), which had to be extracted and specifically coded into the respective fields. In addition, the pilot was used to compare the performance of the three vendor proposals, allowing identification of the most suitable proposal to move into the discovery phase.

RESULTS

Overall accuracy of information extraction

The results for overall accuracy of information extraction were determined by the composite F1 scores of 0.72, 0.52, 0.74, and 0.69 for vendor 1, vendor 2, vendor 3, and Pfizer AI Center of Excellence, respectively, and the individual F1 scores for the nine entity types shown in **Figure 2**. The highest F1 scores were for reporter type and reporter occupation, and the lowest F1 scores were for AE verbatim. The overall F1 score for the machine-learning algorithms used by vendors 1 and 3 exceeded the established internal Pfizer AI Center of Excellence benchmark. On the basis of F1 scores, vendors 1 and 3 outperformed vendor 2, as well as the internal benchmark. Overall, vendor 3 demonstrated the highest scores.

Case-level accuracy

The results for case-level accuracy are summarized in **Figure 3**. The percentage of cases processed with ≥80–100% completeness (eight or nine of nine entity types correctly predicted) was notably higher for vendor 1 (34%) and vendor 3 (31%) than it was for vendor 2 (13%). In addition, learning was evident for vendors 1 and 3 on the basis of the fact that accuracy improved from cycle I to cycle II; however, this was not true for vendor 2. Overall, vendor 1 demonstrated the highest results.

Case-level validity

The results for case-level validity are summarized in **Table 1**. Vendor 1 demonstrated the best results for case-level validity, outperforming all other vendors as well as the internal benchmark. Accuracy improved from cycle I to cycle II for all three vendor algorithms, although the increase was greatest for vendor 1. Improvement in accuracy was lowest for vendor 3, with an increase of only one percentage point.

DISCUSSION

Although emerging AI tools carry the potential to automate or facilitate almost every aspect of a modern pharmaceutical PV department, including case processing, signal detection, risk
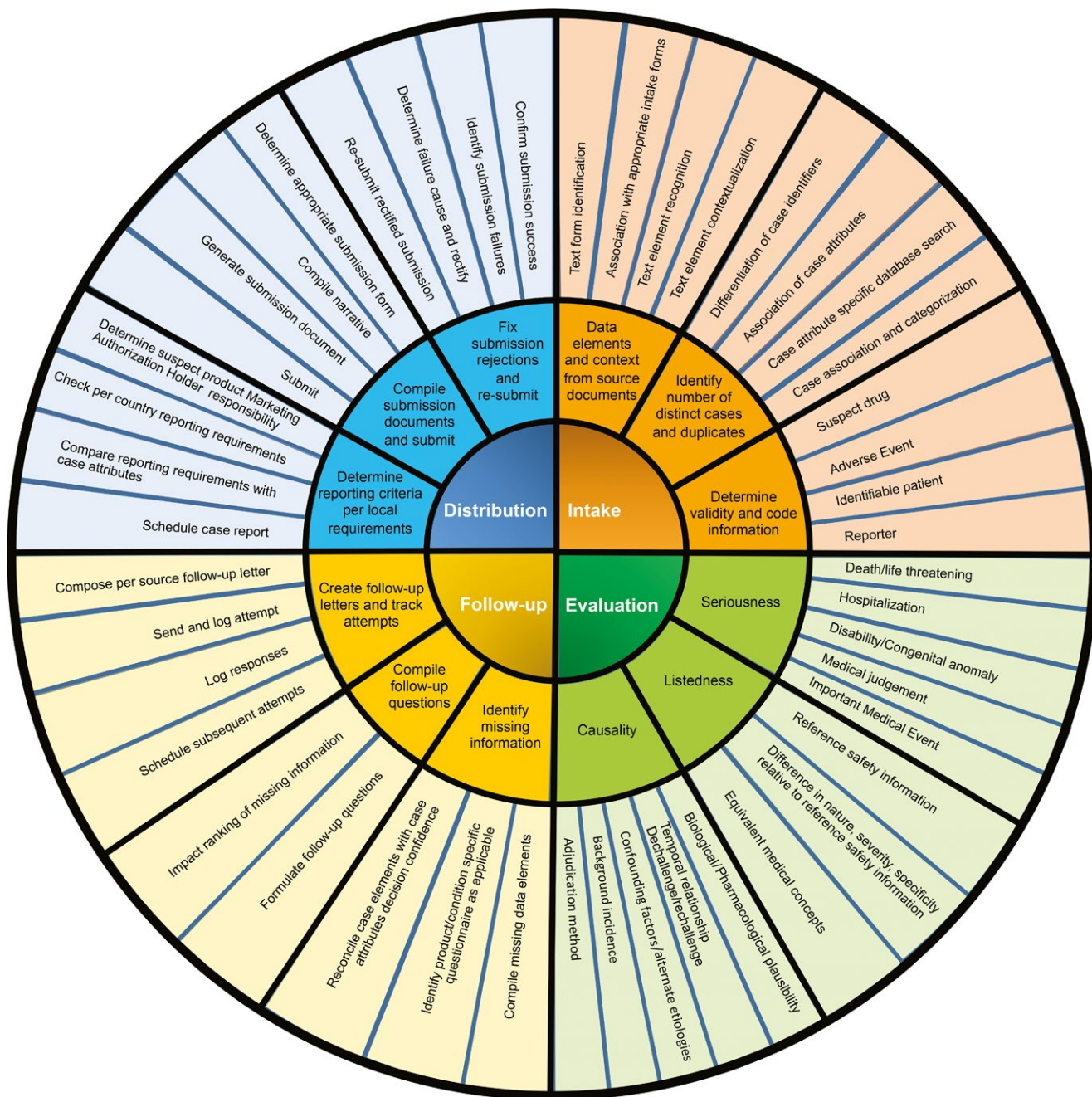
Figure 1 Case processing deliverables.

tracking, and risk contextualization, this pilot study focused on the case processing component, which currently represents the largest economic impact for a PV budget. The results from the pilot demonstrated that it is feasible to apply AI to automate safety case processing. The machine-learning algorithms used were able to successfully train solely on the basis of AE database content (i.e., no source document annotations), and the multiple combined accuracy measures allowed adjudication of the different vendor algorithms.

Within two training cycles, two of the vendors achieved overall F1 scores and case-level accuracy in excess of the Pfizer internal

benchmark. The F1 scores of 0.72–0.74 for these vendors, along with the observed case-level accuracy allowing processing of ≈33.3% of the cases to at least 80% completion, demonstrate that determining case validity during case intake can be performed using machine learning. They also confirm that viable levels of precision and accuracy in machine learning can be achieved using extracted case content from source documents, as captured in the safety database as a surrogate for annotation. A recent study by Comfort *et al.*[19] demonstrated the benefits of a rule-based approach enhanced through machine learning to identify valid AE cases from a social digital media data sources in a miniscule

Figure 2 Summary of F1 scores for nine entity types. Overall composite scores were 0.72, 0.52, 0.74, and 0.69 for vendor 1, vendor 2, vendor 3, and the Pfizer Artificial Intelligence Center of Excellence, respectively. AE, adverse event; DOB, date of birth.

Cycle		≥100%	≥80%	≥70%	≥60%	≥50%	≥40%	≥30%	≥20%	≥10%	≥0%
Vendor 1	I	287	963	651	845	1,325	336	367	185	113	30
	II	524	1,231	415	846	1,140	306	281	145	67	147
Vendor 2	I	70	690	300	776	1,202	595	601	386	319	163
	II	55	634	263	741	1,235	598	638	447	323	168
Vendor 3	I	214	1,084	487	979	1,083	306	232	203	222	292
	II	319	1,312	414	937	965	267	235	160	227	266
Pfizer AI CoE	I	108	768	334	1,195	1,385	371	362	337	173	69
	II	157	909	353	1,127	1,345	329	469	214	129	70

Figure 3 Heat map for case-level accuracy. AI CoE, Artificial Intelligence Center of Excellence.

fraction of the time it would have taken human experts to perform this task and with high sensitivity and specificity. The task of identifying AE cases from large data sources is somewhat similar to the scope of this study. However, a key difference is the method used for machine learning. Comfort *et al.* used annotated data to train the algorithm, this pilot tested the concept of using source documents and the extracted elements from the source documents, as reflected in the safety database, instead of revisiting the source

Table 1 Case-level validity for test cycle I and cycle II

Variable	Correct prediction (%)			Incorrect prediction (%)			
	Valid	Invalid	Total	Valid	Invalid	Total	
Pfizer AI CoE							No prediction (%)
Baseline	68	11	79	12	8	20	<1
Vendor 1							
Test cycle I	68	5	73	18	8	26	<1
Test cycle II	66	15	81	9	9	18	2
Vendor 2							
Test cycle I	52	18	70	5	24	29	<1
Test cycle II	53	20	73	3	19	22	5
Vendor 3							
Test cycle I	44	18	62	5	33	38	0
Test cycle II	45	18	63	5	31	36	0

AI CoE, Artificial Intelligence Center of Excellence.

documentation and annotating them in a labor-intensive manual process. Although this approach may potentially require a larger amount of training data to compensate for the lack of direct situational in-text word association, this type of training data usually is available in abundance.

Only the accurate extraction of information on AEs, suspect drug, patient information, and reporter information contained in source documents allows identification of valid cases, as required by regulatory authorities. The algorithms used by vendor 1 were able to correctly predict the highest percentage of cases as either valid or invalid, with 81% correct predictions after two training cycles, outperforming vendors 2 and 3 as well as the Pfizer internal benchmark of 79%.

Using the combination of multiple measures of precision, recall, and accuracy, it was possible to clearly differentiate vendor 1 from the other vendor proposals. Caution is advised when basing such a critical selection decision on the F1 score alone, which does not sufficiently reflect all important parameters to be considered in the evaluation of machine-learning algorithms.

There are several limitations to this study to be considered. This pilot was not conducted with the intention of identifying cases for regulatory submission. Rather, it was designed to test the concept of viability of automation of safety case processing using machine learning, an AI tool, and to help select the best performing among the three vendors. In this pilot, a combined scoring on the basis of F1, accuracy, and case validity assessment was used to judge overall performance of each of the four systems. However, for a production system, to be used in a regulatory environment, additional criteria and specifically high sensitivity thresholds would have to be used to ensure no valid AE information is missed. The rate of false negatives will need to be kept to a minimum.

Because this pilot also tested the viability of using machine-learning algorithms trained with source documents and database content rather than annotated source documents, manually reviewed cases are taken as the gold standard. The validity of using this gold standard is dependent on the quality of the manual case processing, which was performed with generally accepted

benchmark processes, including peer review and quality control. Furthermore, the availability of a sufficiently large volume of training data is critically important for machine-learning algorithms and will play an important role in achieving a higher degree of automation within the context of the current manual process of managing safety case processing.

Optical character recognition technologies play an important role in the development of intelligent automation solutions, as in this pilot program, in which optical character recognition software technologies were critical in making source document images machine readable. Because this technology is constantly evolving, the choice in optical character recognition software has a significant effect on consistency and accuracy of the AI interpretation.

Finally, there is a limitation to understanding the differences observed in the performance of the three vendors, because the pilot merely evaluated the composite performance of the vendors' systems, whereas the investigators remained agnostic to the algorithms used in their respective proprietary systems. Accordingly, no algorithm-specific data were collected. Consequently, this pilot does not allow performance comparisons of the underlying algorithmic components.
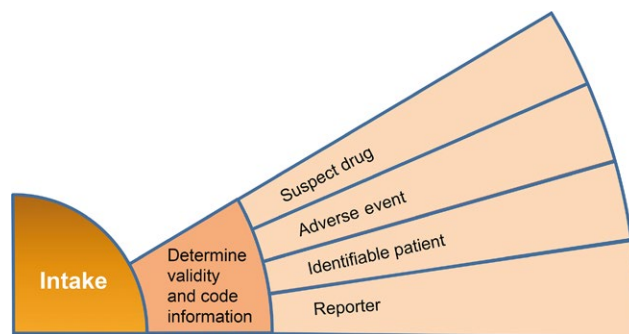

Figure 4 Process element selected for proof of concept.

In conclusion, the pilot was successful in confirming the feasibility of using AI-based tools to support PV operations and in demonstrating the viability of an efficient training method that does not require time-consuming and costly annotations. Finally, the evaluation and scoring method used in the pilot was able to differentiate vendor capabilities and identify vendor 1 as the best candidate to move into the discovery phase.

METHODS

As shown in **Figure 4**, a single process element within case intake (i.e., determination of case validity and code information) was selected for this test, and an internal standard was established for this process element by the internal Pfizer AI Center of Excellence, serving as a comparison benchmark. Vendors were requested to develop their algorithms to optimize performance for the selected test element.

Pilot design

The pilot was designed to simultaneously test proposed solutions of three commercial vendors for the ability to extract case critical information from source documents to identify valid AE cases after training the machine-learning algorithms with source documents and database content rather than annotated source documents. Validity was established by the presence of four elements (i.e., an AE (suspected adverse drug reaction), putative causal drug, patient, and reporter), which had to be extracted and specifically coded into the respective fields. The pilot protocol consisted of two cycles, depicted in **Figure 5**. During cycle I, an initial set of "training" data, consisting of ≈50,000 case source documents and associated safety database records, was fed into the respective test algorithms during a baseline machine-learning phase, followed by a novel set of "test" data consisting only of source documents from 5,000 case source documents to be evaluated by the algorithm. During cycle II, the original plus an additional set of training data of 50,000 case source documents were fed into the test algorithms during a second machine-learning phase, followed by the same set of test data used in cycle I. This design allowed testing of the viability of using the selected machine-learning algorithms and using noncontextual source document extractions in the form of database content, as well as comparison of test scores between cycle I and cycle II to assess the incremental machine learning. Accordingly, the training paradigm chosen for the pilot required the machine to learn on the basis of the information contained in case source documents and corresponding case content that had previously been entered into the safety database (i.e., the "answer keys"). This training method was selected because it does not require additional, time-consuming annotations to be made in support of machine learning. This paradigm represents a higher bar than would be imposed by rule-based training, but it carries an important benefit of increased efficiency.

Machine-learning algorithms and techniques used for establishing the internal benchmark

To support the pilot design, case documentation first had to be converted from pdf file format to machine-readable text documents. Optical character recognition was used to digitize pdf file content using open source techniques.[20]

Next, several different machine-learning algorithms were used to extract data from the digitized documentation:

- Table pattern recognition was used to predict if a specific table cell contained a certain type of information of interest (e.g., patient name or case narrative). This was accomplished by extracting various types of features, such as the location of the cell and the contents of the cell, and then feeding these features into a conditional random field model to predict the label of the current cell.[21]
- Sentence classification was used to predict if a sentence within a case narrative was related to AEs. The machine was subject to supervised learning, by which case narrative sections were extracted from AE reporting forms (i.e., the answer key) and were split into sentences and evaluated (e.g., words appearing around the sentence were identified).[22–24]
- Named entity recognition was used to predict AEs at a token level. A conditional random field (sequence labeling) model was used to detect adverse drug reactions from case narratives.[18,25]
- Rule-based pattern matching used various predefined rules to extract information of interest, including patient name (initials), sex, age, and date of birth.
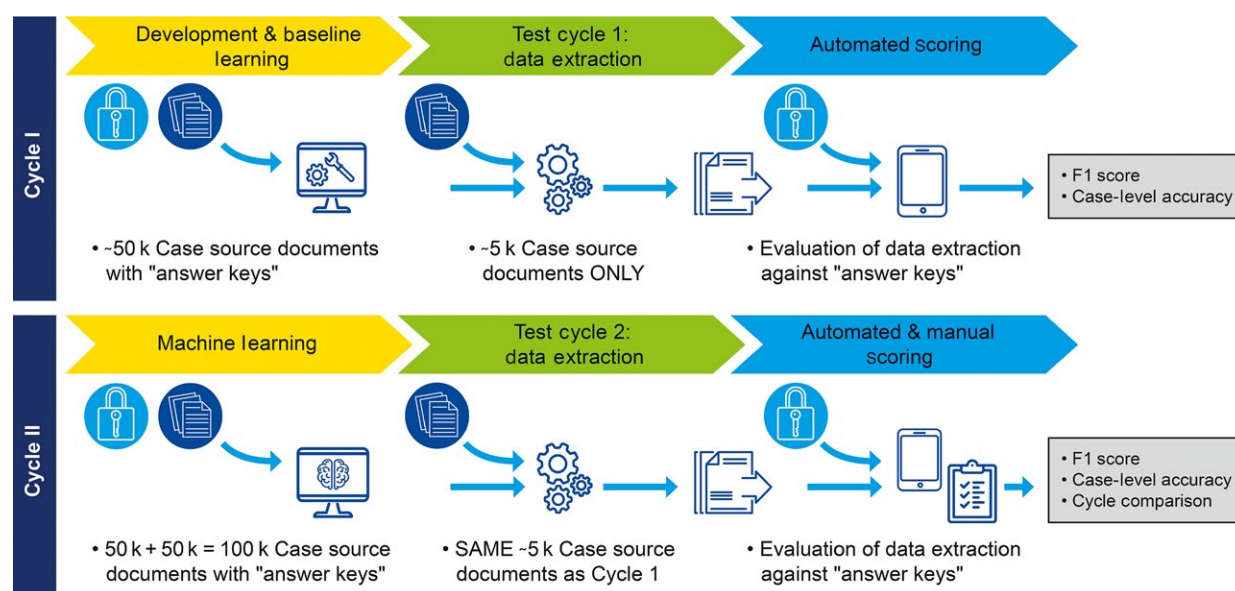

Figure 5 Pilot design.

Scoring and evaluation of results

Overall accuracy of information extraction. Accuracy was evaluated on the basis of precision and recall and the computed F1 score.[26] Precision is the ratio of correctly predicted positive values/the total predicted positive values, and it highlights the correct positive predictions of all the positive predictions. High precision indicates a low false-positive rate. Recall is the ratio of correctly predicted positive values/the actual positive values, and it highlights the sensitivity of the algorithm (i.e., of all the actual positives, how many were identified). F1 score is the harmonic mean of precision and recall, taking into account both false positives and false negatives, and was the primary data element evaluation measure used in the pilot.

$$\text{Precision} = \frac{\#\ \text{Correct Entities Found}}{\#\ \text{Total Entities Found}} \rightarrow \text{If a prediction is made, how likely is it to be correct?}$$

$$\text{Recall} = \frac{\#\ \text{Correct Entities Found}}{\#\ \text{Total Entities}} \rightarrow \text{If there is an entity, how likely is it captured?}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \rightarrow \text{Standard measure for balancing precision + recall.}$$

Precision, recall, and F1 score were calculated for each of nine entity types:

1. AE verbatim text: the verbatim sentence(s), from the original document, describing the reported event(s)
2. Suspect drug
3. Concomitant drug
4. Patient's age
5. Patient's sex
6. Patient's date of birth
7. Patient's initials
8. Reporter type
9. Reporter occupation

For each of these nine categories, true positives, false positives, and false negatives were calculated, and from these metrics, the precision, recall, and F1 score were computed. The F1 scores for each of the nine data elements were averaged into a composite score for all the data elements combined.

Case-level accuracy. Case-level accuracy was determined by the degree of case completeness, defined as the correct predictions (i.e., matches) among the nine entity types. For example, a case with nine of nine matches was considered 100% complete, a case with eight of nine matches was ≥89% complete, and a case with seven of nine matches was ≥78% complete.

Case validity. Manual case processing follows a rule-based approach to identify a regulatory valid case adhering to a standard method that includes peer review and quality control,[27] whereas the automated case processing in this pilot used a binary classification algorithm in which there are two possible predicted classes, "valid" (submitted documentation contains a valid case) and "invalid" (submitted documentation does not have a valid case). Confusion matrices were used to describe the performance of this classification model (or "classifier") on a set of test data for which the true values are known. The results are displayed in **Table 1**.

CONFLICT OF INTEREST
The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS
K.N. and J.S. wrote the article; K.K., J.S., B.S., and C.L. designed the research; K.K., J.S., B.S., and C.L. performed the research; K.K. analyzed the data.

1. Navitas Life Sciences PVNET Benchmark Survey 2016.
2. Yang, C.C., Yang, H. & Jiang, L. Postmarketing drug safety surveillance using publicly available health consumer contributed content in social media. *ACM Trans. Manage. Inf. Syst.* **5**, 2–21 (2014).
3. Carreiro, A.V., Amaral, P.M.T., Pinto, S., Tomas, P., de Carvalho, M. & Madeira, S.C. Prognostic models based on patient snapshots and time windows: predicting disease progression to assisted ventilation in amyotrophic lateral sclerosis. *J. Biomed. Inform.* **58**, 133–144 (2015).
4. Pivovarov, R., Perotte, A.J., Brave, E., Angiolillo, J., Wiggins, C.H. & Elhadad, N. Learning probabilistic phenotypes from heterogeneous EHR data. *J. Biomed. Inform.* **58**, 156–165 (2015).
5. Huang, Z., Dong, W. & Duan, H. A probabilistic topic model for clinical risk stratification from electronic health records. *J. Biomed. Inform.* **58**, 28–36 (2015).
6. Zhang, Y., Padma, R. & Patel, N. Paving the COWpath: learning and visualizing clinical pathways from electronic health records. *J. Biomed. Inform.* **58**, 186–197 (2015).
7. Segura-Bedmar, I. & Martinez, P. Pharmacovigilance through the development of text mining and natural language processing techniques. *J. Biomed. Inform.* **58**, 288–291 (2015).
8. Botsis, T., Nguyen, M.D., Woo, E.J., Markatou, M. & Ball, R. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. *J. Am. Med. Inform. Assoc.* **18**, 631–638 (2011).
9. Alvaro, N., Conway, M., Doan, S., Lofi, C., Overington, J. & Collier, N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *J. Biomed. Inform.* **58**, 280–287 (2015).
10. Ben Abacha, A. & Zweigenbaum, P. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. 12th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2011, Tokyo, Japan, 2011. <https://rd.springer.com/book/10.1007%2F978-3-642-19400-9>.
11. Fong, A., Hettinger, A.Z. & Ratwani, R.M. Exploring methods for identifying related patient safety events using structured and unstructured data. *J. Biomed. Inform.* **58**, 89–95 (2015).
12. Ben Abacha, A., Mahbub Chowdhury, F., Karanasiou, A., Mrabet, Y., Lavelli, A. & Zweigenbaum, P. Text mining for pharmacovigilance: using machine learning for drug name recognition and drug-drug interaction extraction and classification. *J. Biomed. Inform.* **58**, 122–132 (2015).
13. Kim, S., Liu, H., Yeganova, L. & Wilbur, J.W. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. *J. Biomed. Inform.* **58**, 23–30 (2015).
14. Lopez Pineda, A., Ye, Y., Visweswaran, S., Cooper, G.F. & Wagner, M.M. Comparison of machine learning classifiers for influenza

detection from emergency department free-text reports. *J. Biomed. Inform.* **58**, 60–69 (2015).

15. Yang, M., Kiang, M. & Shang, W. Filtering big data from social media: building an early warning system for adverse drug reactions. *J. Biomed. Inform.* **54**, 230–240 (2015).

16. Liu, X. & Chen, H. A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports. *J. Biomed. Inform.* **58**, 268–279 (2015).

17. Weaver Fisher, A. & McKenney, J.L. The development of the ERMA Banking System: lessons from history. *IEEE Ann. Hist. Comput.* **15**, 1 (1993).

18. Aggour, K.S., Bonissone, P.P., Cheetham, W.E. & Messmer, R.P. Automating the underwriting of insurance applications. *AI Magazine* **27**, 3 (2006).

19. Comfort, S. *et al.* Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media. *Drug Saf.* **41**, 579–590 (2018).

20. Python Package, pdfminer. <https://github.com/euske/pdfminer/>. Accessed 7 March 2017.

21. Sutton, C. & McCallum, A. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning* **4**, 267–373 (2011).

22. Gurulingappa, H., Fluck, J., Hofmann-Apitius, M. & Toldo, L. Identification of adverse drug event assertive sentences in medical case reports. Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery 2-11 Workshop on Knowledge Discovery in Health Care and Medicine, Athens, Greece, 2011.

23. Kim, S., Martinez, D., Cavedon, L. & Yencken, L. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*. **12** (suppl 2), S5 (2012).

24. Boser, B.E., Guyon, I.M. & Vapnik, V.N. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory; Pittsburgh, PA, 144–152 (1992).

25. Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R. & Gonzalez, G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J. Am. Med. Inform. Assoc.* **22**, 671–681 (2015).

26. Freitag, D. Machine learning for information extraction in informal domains. *Mach. Learn.* **39**, 169–202 (2000).

27. European Medicines Agency. EMA/873138/2011 rev 2-guideline on good pharmacovigilance practices (GVP). Module VI–collection, management and submission of reports of suspected adverse reactions to medicinal products. <http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2017/08/WC500232767.pdf> (July 2017).