Drug Safety (2022) 45:429–438
https://doi.org/10.1007/s40264-022-01157-4

CURRENT OPINION

"Artificial Intelligence" for Pharmacovigilance: Ready for Prime Time?

Robert Ball[1] · Gerald Dal Pan[1]

Accepted: 10 February 2022

Abstract

There is great interest in the application of 'artificial intelligence' (AI) to pharmacovigilance (PV). Although US FDA is broadly exploring the use of AI for PV, we focus on the application of AI to the processing and evaluation of Individual Case Safety Reports (ICSRs) submitted to the FDA Adverse Event Reporting System (FAERS). We describe a general framework for considering the readiness of AI for PV, followed by some examples of the application of AI to ICSR processing and evaluation in industry and FDA. We conclude that AI can usefully be applied to some aspects of ICSR processing and evaluation, but the performance of current AI algorithms requires a 'human-in-the-loop' to ensure good quality. We identify outstanding scientific and policy issues to be addressed before the full potential of AI can be exploited for ICSR processing and evaluation, including approaches to quality assurance of 'human-in-the-loop' AI systems, large-scale, publicly available training datasets, a well-defined and computable 'cognitive framework', a formal sociotechnical framework for applying AI to PV, and development of best practices for applying AI to PV. Practical experience with stepwise implementation of AI for ICSR processing and evaluation will likely provide important lessons that will inform the necessary policy and regulatory framework to facilitate widespread adoption and provide a foundation for further development of AI approaches to other aspects of PV.

Key Points

Application of "artificial intelligence" (AI) to pharmacovigilance (PV) might fruitfully begin with the processing and evaluation of Individual Case Safety Reports (ICSRs) as the number of ICSRs that are processed, submitted, and assessed for safety signals continues to grow and ICSRs will likely remain an important part of PV for the foreseeable future.

The performance of current AI algorithms applied to processing and evaluation of ICSRs, while generally not sufficient for complete automation, can likely be applied to improve efficiency, value, and consistency if integrated into a system with a "human-in-the-loop" for careful quality control.

✉ Robert Ball
Robert.Ball@fda.hhs.gov

[1] US Food and Drug Administration, Center for Drug Evaluation and Research, Office of Surveillance and Epidemiology, Silver Spring, MD, USA

1 Introduction—The Need for AI in Pharmacovigilance

There is much excitement about the application of 'artificial intelligence' (AI) approaches to drug[1] development and lifecycle drug management, including pharmacovigilance (PV) [1]. The US FDA defines PV as "all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events" [2]. FDA's definition of PV is broad and includes the use of a wide range of scientific inquiry, such as Individual Case Safety Reports (ICSRs), pharmacoepidemiologic studies, registries, clinical pharmacology studies, and other approaches. Although FDA is exploring the use of AI in many of these areas [1, 3–7], research in these areas is not yet mature enough to consider widespread implementation from a regulatory perspective. We focus here on the application of AI to the processing of data from multiple sources to identify adverse events (AEs) meeting regulatory reporting requirements, the preparation of these AEs as ICSRs, and their further reporting and evaluation. We take this focus because of the following.

[1] All references to 'drugs' include both human drugs and biological drug products.

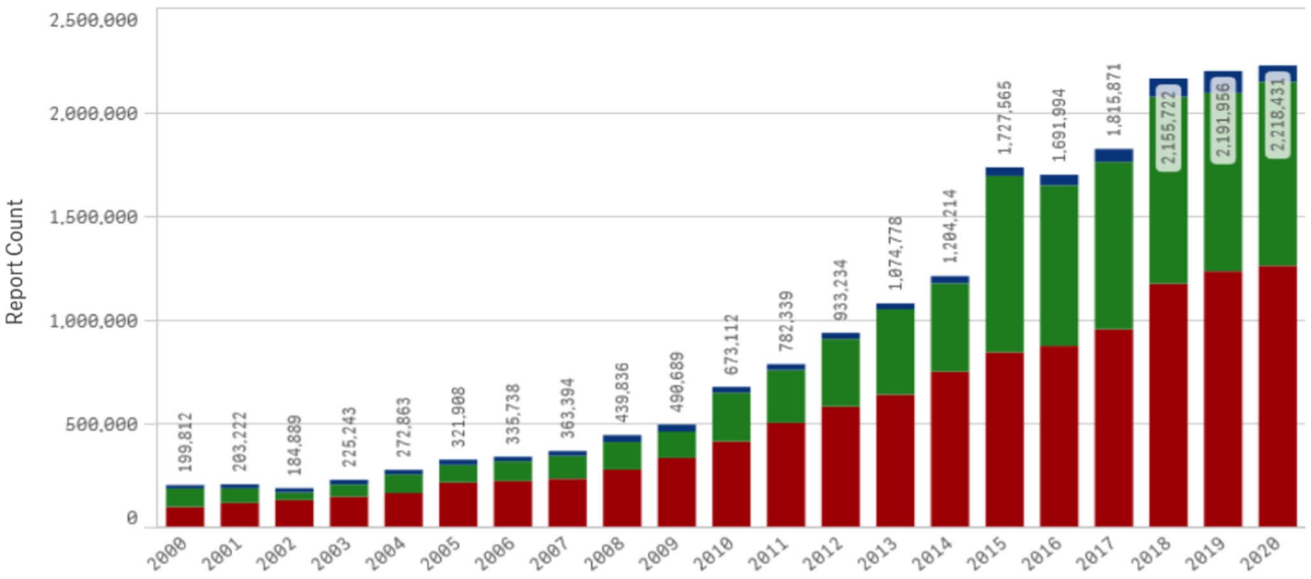


Fig. 1 Individual case safety reports received by the US FDA adverse event reporting system (FAERS) have increased dramatically in the past two decades

1. New safety issues arise frequently after a drug is approved [8, 9] and ICSRs have a long, proven track record of identifying safety issues and remain the source of important new safety information [10].
2. There are an increasing number and variety of data sources that need to be evaluated for safety information that result in a growing volume of ICSRs that are processed, submitted, and assessed for safety signals by industry and regulators (see Fig. 1), leading to increased costs and workloads for a limited supply of human safety experts. This general trend has been accentuated by increased reporting for products used for prophylaxis and treatment of coronavirus disease 2019 (COVID-19).
3. Submission of ICSRs is required by regulators globally and harmonization of approaches improves efficiencies and promotes standardization.
4. Despite the growing interest in identifying and assessing safety signals based on analyses of population-based data sources [11–14], a full assessment of how these approaches will best fit in PV remains to be completed, therefore ICSRs will likely continue to play an important role as an early warning system of drug safety signals, especially for rare events, and will remain a substantive component of the PV enterprise for the foreseeable future.
5. While modifications to existing approaches to reporting ICSRs have been proposed [15], it is not likely that such changes alone will be sufficient to address the increased number of data sources to be evaluated and ICSRs to be submitted.

AI potentially plays an important role in improving the efficiency and scientific value of ICSRs. In this paper we review the landscape of approaches inside and outside of FDA that are being taken to address this issue.

While FDA has not adopted a formal definition of AI for PV, the FDA document on “Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan” [16] notes “Artificial intelligence has been broadly defined as the science and engineering of making intelligent machines, especially intelligent computer programs” [17]. While the Action Plan focuses on the application of AI to medical devices, the scientific framework it articulates may also be usefully applied to AI for PV. Many technologies have been placed under the ‘AI’ umbrella; for PV, machine learning (ML) and natural language processing (NLP) are two of the most common being applied. ML is defined as a “… technique that can be used to design and train software algorithms to learn from and act on data ...” [16] and NLP is defined as “the application of computational techniques to the analysis and synthesis of natural language and speech” [18]. We first describe the ICSR-related processes and workflows where different AI approaches might be most fruitfully applied. We then describe a framework for considering the readiness of AI for ICSR processing and evaluation, followed by some examples of the application of AI to ICSR processing and evaluation in industry and FDA, with a comparison to the readiness framework, and identify outstanding scientific and policy issues to be addressed before the full potential of AI can be exploited for ICSR processing and evaluation and PV more generally.

2 Pharmacovigilance Processes and Individual Case Safety Reports

An ICSR contains information on the patient, the AE, the suspect medical products, the reporter, and, for ICSRs submitted by industry, information on the company that holds the application or license for the drug [19]. According to FDA regulations, ICSRs might need to be generated "from any source, foreign or domestic, including information derived from commercial marketing experience, postmarketing clinical investigations, postmarketing epidemiological/surveillance studies, reports in the scientific literature, and unpublished scientific papers" [20]. Case processing and evaluation begin once the company has made a determination that they must make an assessment of reportability of case information from any source. We outline case processing and evaluation phases, with two principal divisions between the work conducted by the regulated industry and that conducted by FDA.

2.1 Case Processing

Case processing has been described as having four activities, including intake, evaluation, follow-up, and distribution, with many subprocesses for each activity [21]. Intake of cases potentially requiring submission to FDA includes identification of the four elements ("an identifiable reporter, an identifiable patient, an adverse reaction, and a suspect product") that, when present, indicate that an ICSR must be prepared and submitted [22]. While these four elements are the minimum elements of an ICSR, an ICSR must also include all relevant information when such information is available [23]. Additional steps involve determination of important regulatory categories such as seriousness of the AE, whether the AE is already in the FDA prescribing information for the product (expectedness), and, for certain ICSRs (AEs from a study), likelihood of a causal association. These determinations depend on information in the report, the product label, and the source of information [23]. Report follow-up to obtain missing information is also conducted and the report is transmitted to regulators. Currently, the principal means of standardization for transmitting ICSRs from industry to regulatory agencies is specified in the International Council for Harmonisation (ICH) E2B guideline [24]. Importantly, this standardization encompasses many data elements that are placed in structured fields, as well as an unstructured narrative description of the case that often contains valuable information not codified in the structured data.

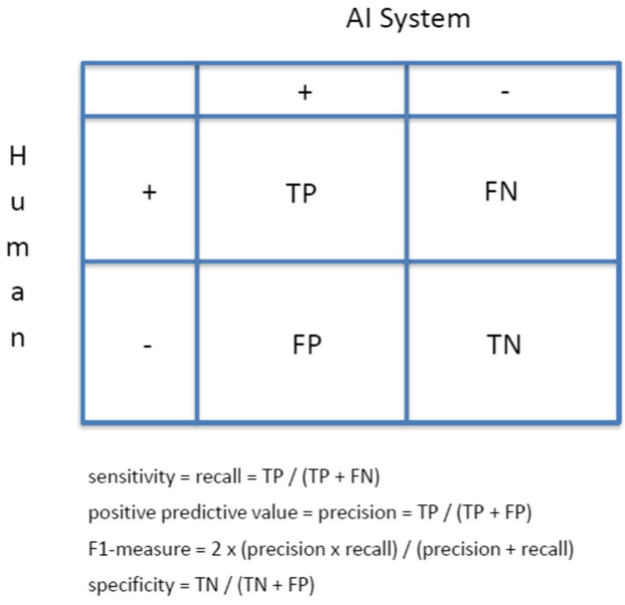
2.2 Case Causality Assessment

Case causality assessment—the determination of whether the drug is likely to have caused the reported AE—takes place at both the industry and FDA. Assessment of ICSRs for causality still relies primarily on expert judgment and global introspection [25, 26]. Although FDA requires companies to have "written procedures for the surveillance, receipt, evaluation, and reporting of postmarketing adverse drug experiences" [23] and has defined best practices [27] and workflows [28] for its own work, the ICSR case assessment workflow is not fully standardized to a level required for computation [29]. More importantly, any effort to standardize the workflow for purposes of computation must acknowledge the need for expert judgment and flexibility. This requirement means that understanding both the individual tasks that are performed and how they are then assembled into a cognitive framework for assessment to support human efforts, in multiple and difficult-to-describe scenarios, are necessary for AI approaches to be applied [30].

3 Framework for Considering the Readiness of Artificial Intelligence (AI) for Pharmacovigilance

Several factors must be considered when deciding whether an AI algorithm might be ready for implementation. Algorithm performance (e.g., validity, generalizability, absence of bias, and robustness in real-world settings with changing inputs) is arguably the essential first step, but documentation, transparency, explainability (i.e., the reasons for an algorithm's prediction), quality control with real-world data collection and monitoring, and algorithm change control are all needed. AI best practices around data management, feature extraction, training, interpretability, evaluation, and documentation are still in development and harmonization of the numerous efforts around best practices, including through consensus standards efforts, leveraging already existing workstreams, and involvement of other communities focused on AI/ML, will be needed [16]. There is still a need to standardize terminologies used in AI frameworks, with similar concepts being represented using different words depending on the context. AI for PV will have to be aligned with these emerging best practices for the field to reach a state of maturity [31].

While it is beyond the scope of this paper for a complete discussion of all of these issues, a discussion of some of the core aspects of algorithm performance will be helpful as a first step in assessing the readiness of current AI algorithms for ICSR processing and evaluation. The key factors in algorithm performance are the metrics chosen



AI System

Human	+	-
+	TP	FN
-	FP	TN

sensitivity = recall = TP / (TP + FN)

positive predictive value = precision = TP / (TP + FP)

F1-measure = 2 x (precision x recall) / (precision + recall)

specificity = TN / (TN + FP)

Fig. 2 Standard metrics of AI algorithm performance. *AI* artificial intelligence, *TP* true positive, *FP* false positive, *FN* false negative, *TN* true negative

to measure that performance and the implications of the values of those metrics for implementation. Some standard metrics of AI algorithm performance are shown in Fig. 2. Recall (sensitivity), precision (positive predictive value [PPV]), and F1 score are commonly used metrics. The F1 score is a summary measure of recall and precision and we will use it in this paper as a means for illustration and comparison, but it is not necessarily the metric of choice for all purposes. For example, recall (sensitivity) might be a very important metric to use in the context of identifying AEs that meet the criteria for reporting to a regulator, as we discuss later.

As a practical matter, the lower bound of an AI system's F1 score is not hard and fast and will depend on the ability of the AI system to add sufficient value to be worth implementing in the context in which it is being operationalized. On the other hand, full automation, or use of the AI system's output without human review, would require an F1 score approaching 1.0. How close to an F1 score of 1.0 system performance needs to be depends on the risks associated with erroneous classification. For example, if misclassification by the algorithm was to lead to missing an important safety signal, a near perfect F1 score would likely be required. Some general criteria for qualitatively assessing whether an AI system is performing at least as well as human experts and might be a candidate for full automation include whether human experts see no obvious patterns in an analysis of any erroneous classification, and, in human review of algorithm outputs, whether any perceived errors are not obvious misclassifications and are similar to the differences of opinion that might arise among human experts. The exact manner by which a determination of readiness for full automation can be achieved remains an open question, but human expert quality assessment will likely be required for the foreseeable future.

We shift now to a discussion of a few published examples from the literature to illustrate how the above framework might be applied.

4 Examples from Industry

A major area of interest of the pharmaceutical industry is in case processing [32, 33]. Current areas of AI activities in assessment, proof of concept, or development for production stages include digital media screening; extracting and classifying data from source documents; checking for duplicate reports; case validation (e.g. minimum reporting requirements); triage and initial assessment (e.g. seriousness, expectedness); data entry (e.g. structured fields accurately populated with available information); medical assessment, including causality; narrative writing; and coding AE concepts into standardized terminology [32].

Published examples of applications of AI by the pharmaceutical industry include ICSR processing [21], determination of seriousness [34], and causality assessment [35]. In the case processing example, F1 scores ranged widely from 0.36 for identifying the 'AE verbatim' (defined as the "verbatim sentence(s), from the original document, describing the reported event(s)") as part of the process for determining whether an AE is present in the original document, to a high of 0.91 for 'reporter occupation' across multiple algorithms and tests [21]. The seriousness classifier achieved F1 scores for categorization ranging from 0.76 to 0.79 [34]. Comparison of MONARCSi and Roche safety professionals' assessments of causality had an F1 score of 0.71 [35]. It is important to note that FDA has not endorsed any specific approach to applying AI to case processing and evaluation or any quantitative metric of algorithm performance; these examples are provided because they illustrate the approaches being taken and the performance of the algorithms.

5 Application of the Readiness Framework to Case Processing

As an example of how the general framework for considering the readiness of AI for ICSR processing and evaluation might be applied, we focus on the potential automation of the identification of cases required to be submitted to FDA. At the highest level, scientifically rigorous procedures must be in place to ensure that processes for AE identification

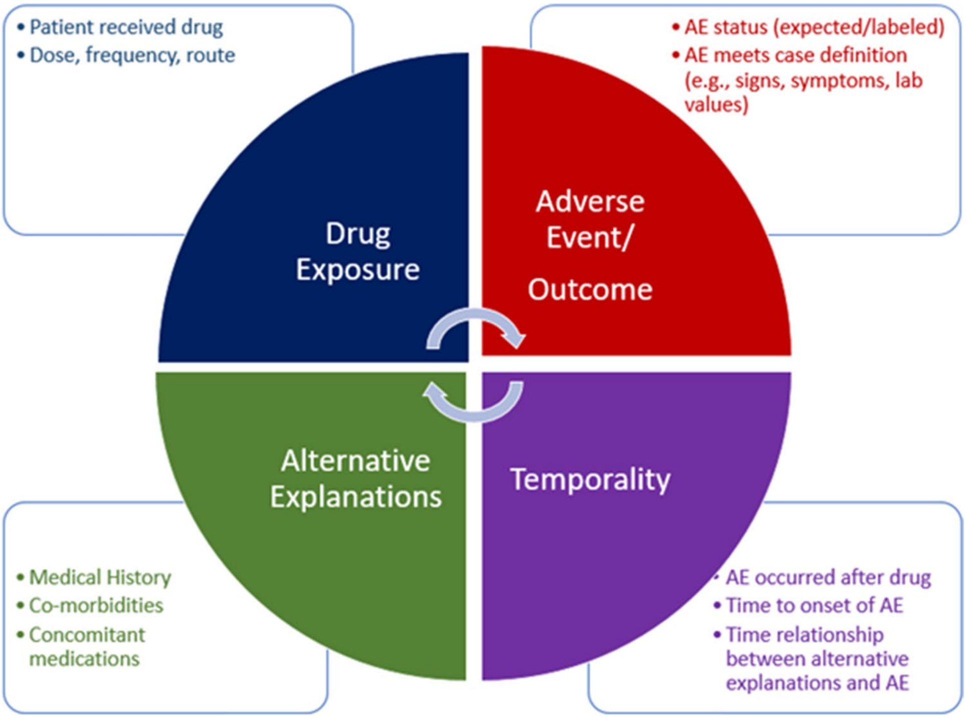


Fig. 3 Elements of the cognitive framework for ICSR causality assessment. *AE* adverse events, *ICSR* Individual Case Safety Reports

have both high sensitivity and high specificity. Without high sensitivity, AEs would not be identified and thus potentially important safety signals would be missed. Without high specificity, ICSRs would be generated for events that are not AEs, but are identified as such by the AI system, a situation that would potentially result in submitting more reports than necessary and creating noise that would make safety signal detection more difficult.

As mentioned earlier, the published performance of different AI algorithms for key aspects of the case identification process does not achieve the threshold likely needed for full automation (F1 scores approaching 1.0). The performance is sufficiently good that it might justify implementation to improve the efficiency of case processing in certain situations. The decision to implement a less than perfect AI algorithm should be made by the organization that manages the process according to its own analysis, but an overarching consideration is that important quality checks will be needed to ensure the performance of the combined human–AI system is at least as good as the human-only system it is replacing. This approach to using AI to support, rather than supplant, human experts is sometimes referred to as augmented intelligence [30] and includes a 'human-in-the-loop'. In the example of AI for automation of the identification of cases required to be submitted to FDA discussed earlier, human review of an AI algorithm's output would be needed to ensure no true AEs are missed, and no 'non-AEs' are submitted. In addition, with ML approaches, it is anticipated that the algorithm will be periodically retrained on new data or new algorithms will be developed. Each time an algorithm is retrained, a formal validation of system performance will be required. The performance of any given system should be evaluated in the 'real world' within the workflow where it is to be employed.

6 FDA's Experience

FDA has its own interests in applying AI to its PV processes to improve the efficiency and scientific value of its analyses of ICSRs. In addition to the nearly two million reports to the FAERS that FDA receives each year from industry, FDA processes into FAERS several hundred thousand reports that the public submits directly to FDA. Thus, FDA shares some of the challenges faced by the industry for case processing.

FDA further believes that experts' time is best spent on complex tasks that have public health impact, rather than on extracting and organizing the information from ICSRs that is needed to make an assessment, especially important clinical information that is contained mostly in unstructured narrative text. Given the number of ICSRs that FDA receives each year, FDA's research and development activities have focused on applying AI to address this challenge as part of the causality assessment of ICSRs. Figure 3 highlights some of the key elements that are included in an ICSR causality assessment. The elements included in the assessment are superficially straightforward and include identification of the key features involved in the assessment, namely the drug, AE outcome, their temporal relationship, and alternative explanations for the AE besides a causal relationship

Table 1 Key FDA efforts applying AI to PV from 2011 to the present

Developed NLP for clinical feature extraction and ML classification for a specific case definition (e.g., anaphylaxis) [36–40]
Applied NLP with statistical clustering algorithm/network analysis to identify reports of similar medical condition [41–45, 49]
Used NLP to extract temporal information [46]
Extraction of demographic information and clinical concepts [36, 47, 48]
Applied NLP and ML to summarize key features of ICSRs [50–52]
Use of ML to predict which ICSRs are most useful for causality assessment [29, 54, 55]
Extract and code AEs from drug product label/package insert [56–58]
Developed deduplication algorithm for ICSRs [53]
NLP extraction and visualization of clinical data (e.g., temporality) to support cases series analyses for causality assessment [59]
ML algorithm to identify unassessable cases (e.g., ICSRs containing insufficient information to support causality assessment) [29]

AEs adverse events, *AI* artificial intelligence, *ICSRs* individual case safety reports, *ML* machine learning, *NLP* natural language processing, *PV* pharmacovigilance

with the drug being assessed. While there are challenges in automating the extraction of the details about these features from ICSR narratives (e.g., the signs and symptoms needed to apply a case definition), the larger issue is that the cognitive processes for feature integration are complex, primarily conducted through global introspection, iterative in nature (as reflected by the circular arrows at the center of the figure), and not defined in sufficient detail to make computable. Table 1 categorizes the efforts FDA has taken over the past decade to apply AI to this complex process [29, 36–61]. As can be seen in the description of the efforts, most have involved automating the extraction of the key features from ICSR narratives using NLP, with a few attempting to develop predictive ML algorithms that attempt to automate the cognitive processes for feature integration. While these efforts have resulted in successful development of discrete algorithms, algorithm performance does not yet achieve levels required for full automation.

For example, a common step in evaluating a series of ICSRs to determine whether they support a causal relationship between a drug and AE is the development of a 'case definition' describing the clinical features that are consistent with a particular AE. This case definition is then compared with clinical information in the ICSRs. The first application of AI to PV at FDA involved using NLP and ML to classify AEs identified in ICSRs as possible anaphylaxis after H1N1 influenza vaccination [36]. The best performing anaphylaxis classification algorithm had an F1 score of 0.758 compared with human experts [36].

In a second example, the best performing algorithms for the identification of assessable reports (i.e., those containing enough information to make an informed causality assessment) achieved F1 scores above 0.80 [29]. This was accomplished by training the algorithm on reports classified as either 'assessable' or 'unassessable' (i.e., non-informative for causality assessment). Algorithms attempting to address other aspects of causality assessment did not perform as well, suggesting that identification of low-value reports might be a first step in applying AI to causality assessment of ICSRs.

Integrating these imperfect algorithms into the existing workflow and into information technology (IT) systems is an ongoing challenge. In the production setting, extraction of key features (e.g., age) from the ICSR narrative has been implemented; integration into traditional workflows and IT systems of more complex algorithms, such as identifying and removing duplicate ICSRs based on both structured fields and narrative text, is underway. Development of a general platform that breaks down the case evaluation process into computable steps and would allow for insertion of improved algorithms for a given task (e.g., automating the application of a case definition) is an active area of research [59], along with application of AI-based language models to ICSR narratives to improve extraction of key features and their relationships [60, 61].

7 Approaches to Quality Assurance of "Human-in-the-Loop" AI Systems

If an AI algorithm does not achieve performance levels required for full automation, the key challenge of including a 'human-in-the-loop' is to ensure quality without reducing the efficiency gained from the AI algorithm. Stated otherwise, the human expert should not do the work that the machine can do well and efficiently, and the machine should not do (poorly) the work that the human expert can do well. General considerations for the characteristics of quality assurance that might be applied to a human-in-the-loop approach to an imperfect AI system include (1) a risk-based approach in which effort is proportional to the implications of misclassification on the overall evaluation goals; (2) incorporation of the reliability of the AI algorithm's performance through carefully applied principles of algorithm development or formal confidence metrics; and (3) selection of quality assurance techniques such as sampling, simultaneous independent algorithm application, and incorporating the AI algorithm in a general evaluation process that includes other means of quality assurance.

To illustrate the challenge and areas where further research might be helpful to turn these general considerations into concrete approaches, consider the example of an AI algorithm that predicts whether a report has valuable information needed for making a causality assessment. Efficiency can be gained if such an algorithm's prediction and an appropriate threshold would identify many potentially low-value reports that would not need human expert review. One approach to ensure quality (i.e., in this example, correct classification of high-value reports) would be to set the threshold so there are no high-value reports falsely classified as being low-value (the algorithm would have perfect PPV for identifying low-value reports). Typically, there is a trade-off between PPV and sensitivity, therefore having a perfect PPV would likely lead to a lower sensitivity for identifying low-value reports. This would result in some low-value reports being incorrectly classified as high-value reports and fewer low-value reports excluded from human review, thus undermining the efficiency gains (i.e., in this example, sparing the human expert from reviewing low-value reports) from the AI algorithm. On the other hand, if the threshold was adjusted so the algorithm had a lower PPV and likely a higher sensitivity, some high-value reports would be incorrectly classified as being low-value, therefore the efficiency of the entire process would be improved (i.e., the human expert has fewer low-value reports to review) but at the cost of lower quality because high-value reports would be missed unless additional quality assurance procedures were in place.

In this scenario, quality might be assured by human expert assessment of a random sample of the excluded reports. To maintain efficiency, the size of the sample could not be large, therefore such a random sampling process would likely not find all high-value reports misclassified as low-value. Thoughtful design of the sampling process (e.g., oversampling reports with algorithm scores close to the threshold, reports with algorithm scores for which the algorithm's predictions are known to be less reliable using a confidence metric, or with drugs or AEs of particular concern or relative rareness) might be considered. Simultaneous use of an independent rule-based algorithm specifically designed to identify important reports (e.g., reports of anaphylaxis, drug-induced liver injury, Stevens–Johnson syndrome) might provide additional assurance. Embedding the specific human-AI system in a more general evaluation process that uses other techniques to ensure the overall goals of the process are not compromised might also be an option. For example, applying such an algorithm only to evaluations with large numbers of reports of a drug–AE combination being evaluated would reduce the chance that a small number of misclassified high-value reports would change the overall conclusions of the case series evaluation. Additional research is needed to determine which of these, or other techniques, might best address the challenge of a human-in-the-loop approach.

8 Challenges

With the exception of the US Vaccine Adverse Event Reporting System (VAERS), large-scale, publicly available datasets of ICSRs with complete information, including narrative descriptions of AEs, are not available because of the need to protect personal health information. Only small ICSR datasets annotated by human experts for the purposes of causality assessment are available due to the expense of annotating and anonymizing the data. Developing a mechanism for sharing datasets with narrative text and appropriate annotations would accelerate progress in applying supervised ML to ICSR processing and evaluation, as well as facilitate harmonization and building trust among stakeholders.

Human expert processes for causality assessment of ICSRs use information that is both internal and external to the report. The development of a well-defined 'cognitive framework' that can be made computable and fit into existing workflows will be needed to further the application of AI to ICSR processing and evaluation. Direct engagement with human PV experts to describe in detail how they do their work and the development of transparent and explainable ML algorithms that identify the key features and their interrelationships in achieving certain goals could converge on a detailed description of the PV cognitive framework that has long eluded the field. Currently, statistical disproportionality analyses [62] and case-series evaluations are largely separate activities. The development of a computable cognitive framework might identify ways in which traditional statistical methods can be integrated with NLP and ML algorithms to more rigorously identify unusual patterns [41, 44] in case series.

Successful implementation of information technology systems requires an understanding of the complex interrelationships among hardware, software, information content, and the human–computer interface [63]. The implementation of systems purporting to introduce AI into the workplaces of a highly regulated industry brings additional dimensions to an already difficult challenge. One approach to addressing this challenge, which has been applied to health information technology for health care delivery, is the development of a formal sociotechnical framework that integrates technology and evaluation with people, workflow, communication, organizational policies, and external rules and regulations [63]. Applying such a framework to fully understand all the steps needed to integrate AI into existing PV processes across the PV enterprise, from patients and providers to pharmaceutical companies to regulators and back to providers and patients, would be a useful next step in creating a roadmap for implementation.

A related challenge is that PV professionals have traditionally been recruited primarily from clinical disciplines, with limited training in quantitative and computational approaches to data analysis. Both in industry and regulatory agencies, the education of PV staff who are not specialists in AI, and targeted recruitment of AI specialists to support AI application for PV, will be critical components of bringing about successful implementation of AI systems for PV.

9 Summary

Some aspects of ICSR case processing and assessment have been shown to be amenable to NLP and ML to augment human expertise. Implementation of some approaches is underway and has been described in the published literature. The likelihood that AI systems will reach a level of performance likely necessary for full automation (with an F1 score approaching 1.0) in the near term is low. Including a 'human-in-the-loop' will likely not only be necessary but also desirable for the foreseeable future. Experience with automation in aviation [64] suggests that thinking of automation as supporting rather than supplanting human expertise provides many benefits. Such benefits include better acceptance, reduced risks of errors, improved understanding of the process human experts actually use, and improved human expert performance [64]. Fully articulating a sociotechnical framework for AI in PV would likely further elucidate similarities and differences between PV and other fields, such as aviation, that have successfully introduced AI and aid in identification of additional measures that might be taken in implementing AI for PV. ICSR evaluation remains an art as much as a science. A potential advantage of efforts at automation is that existing inconsistencies in assessment processes will be revealed, leading to general improvements in decision making.

Key policy and regulatory approaches await more scientific study and development of best practices in AI generally, and for its application to ICSR processing and evaluation. Practical experience with stepwise implementation will likely provide important lessons learned that will inform the necessary policy and regulatory framework that will facilitate widespread adoption of AI for ICSR processing and evaluation in the future. This experience will provide a valuable foundation for further development of AI approaches to other aspects of PV.

Acknowledgements The authors wish to acknowledge Suranjan De and Oanh Dang for assistance with the figures and table.

Declarations

The views expressed are those of the authors and not necessarily those of the US FDA.

Funding This work was conducted by employees of the US FDA as part of their official duties.

Conflict of interest Robert Ball is an author on US Patent 9,075,796, 'Text mining for large medical text datasets and corresponding medical text classification using informative feature selection'. At present, this patent is not licensed and does not generate royalties. Gerald Dal Pan has no conflicts of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material Not applicable.

Code availability Not applicable.

Author contributions RB conceived the manuscript structure, wrote the first draft, and made subsequent revisions. GDP contributed to the manuscript structure and revised multiple versions. Both authors read and approved the final version.

References

1. Liu Q, Zhu H, Liu C, et al. Application of machine learning in drug development and regulation: current status and future potential. Clin Pharmacol Ther. 2020;107:726–9.
2. US FDA. Guidance for industry—good pharmacovigilance practices and pharmacoepidemiologic assessment. 2005. https://www.fda.gov/media/71546/download. Accessed 30 Nov 2021.
3. Platt R, Brown JS, Robb M, McClellan M, Ball R, Nguyen M, et al. The FDA sentinel initiative—an evolving national resource. N Engl J Med. 2018;379:2091–3.
4. Ball R, Toh S, Nolan J, Haynes K, Forshee R, Botsis T. Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System. Pharmacoepidemiol Drug Saf. 2018;27:1077–84.
5. Brown JS, Maro JC, Nguyen MD, Ball R. Using and improving distributed data networks to generate actionable evidence: The case of real-world outcomes in the Food and Drug Administration's Sentinel System. J Am Med Inform Assoc. 2020;27:793–7.
6. Gibson TB, Nguyen M, Burrell T, Yoon F, Wong J, Dharmarajan S, et al. Electronic phenotyping of health outcomes of interest

using a linked claims-electronic health record database: findings from a machine learning pilot project. J Am Med Inform Assoc. 2021;28:1507–17.

7. Ball R. Artificial Intelligence in the FDA's Sentinel System. In: Anklamm E, Bahlb MI, Ball R, et al. Emerging technologies and their application to regulatory science. Experimental biology and medicine. 2021;246:15–18.
8. Pinnow E, Amr S, Bentzen SM, et al. Postmarket safety outcomes for new molecular entity (NME) drugs approved by the Food and Drug Administration between 2002 and 2014. Clin Pharmacol Ther. 2018;104:390–400.
9. Bulatao I, Pinnow E, Day B, Cherkaoui S, Kalaria M, Brajovic S, et al. Postmarketing safety-related regulatory actions for new therapeutic biologics approved in the United States 2002–2014: similarities and differences with new molecular entities. Clin Pharmacol Ther. 2020;108:1243–53.
10. Lester J, Neyarapally GA, Lipowski E, Graham CF, Hall M, Dal Pan G. Evaluation of FDA safety-related drug label changes in 2010. Pharmacoepidemiol Drug Saf. 2013;22:302–5.
11. Kulldorff M, Dashevsky I, Avery TR, Chan KA, Davis RL, Graham D, et al. Drug safety data mining with a tree-based scan statistic. Pharmacoepidemiol Drug Saf. 2013;22:517–23.
12. Bate A, Hornbuckle K, Juhaeri J, Motsko SP, Reynolds RF. Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance. Ther Adv Drug Saf. 2019;10:1–9.
13. Arnaud M, Bégaud B, Thurin N, Moore N, Pariente A, Salvo F. Methods for safety signal detection in healthcare databases: a literature review. Expert Opin Drug Saf. 2017;16:721–32.
14. Harpaz R, DuMouchel W, Schuemie M, et al. Toward multimodal signal detection of adverse drug reactions. J Biomed Informat. 2017;76:41–9.
15. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, E2D(R1), Final Concept Paper for E2D(R1). 2021. https://database.ich.org/sites/default/files/E2D-R1_ConceptPaper_Final_2020_0115.pdf. Accessed 30 Nov 2021.
16. US FDA. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. 2021. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Accessed 30 Nov 2021.
17. McCarthy J. What Is Artificial Intelligence? Stanford, CA: Stanford University. 2007. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf. Accessed 30 Nov 2021.
18. "Definition of natural language processing". Oxford University Press. Lexico.com. 2021. https://www.lexico.com/definition/natural_language_processing. Accessed 30 Nov 2021.
19. Electronic Code of Federal Regulations. Title 21: Food and Drugs Part 314.80 (f). 2021. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.80. Accessed 30 Nov 2021.
20. Electronic Code of Federal Regulations. Title 21: Food and Drugs Part 314.80 (b). 2021. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.80. Accessed 30 Nov 2021.
21. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharmacol Ther. 2019;105:954–61.
22. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, E2D(R1). 2021. https://database.ich.org/sites/default/files/E2D_Guideline.pdf. Accessed 30 Nov 2021.
23. Electronic Code of Federal Regulations. Title 21: Food and Drugs Part 314.80. 2021. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.80. Accessed 30 Nov 2021.
24. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, E2B(R3) clinical safety data management: data elements for transmission of individual case safety reports. 2021. https://ich.org/page/e2br3-individual-case-safety-report-icsr-specification-and-related-files. Accessed 30 Nov 2021.
25. Taofikat B, Agbabiaka JS, Edzard E. Methods for causality assessment of adverse drug reactions: a systematic review. Drug Saf. 2008;31:21–37.
26. Edwards R. Causality assessment in pharmacovigilance: still a challenge. Drug Saf. 2017;40:365–72.
27. US FDA. Best practices in drug and biological product postmarket safety surveillance for FDA Staff, MAPP 4121.3. 2021. https://www.fda.gov/media/130216/download. Accessed 30 Nov 2021.
28. US FDA. Collaborative identification, evaluation, and resolution of a newly identified safety signal (NISS). 2021. https://www.fda.gov/media/137475/download. Accessed 30 Nov 2021.
29. Kreimeyer K, Dang O, Spiker J, Muñoz M, Rosner G, Ball R, et al. Feature engineering and machine learning for causality assessment in pharmacovigilance: lessons learned from application to the FDA adverse event reporting system (FAERS). Comput Biol Med. 2021;135: 104517. https://doi.org/10.1016/j.compbiomed.2021.104517.
30. Zheng N, Liu Z, Ren P, Ma Y, Chen S, Yu S, et al. Hybrid-augmented intelligence: collaboration and cognition. Front Inform Technol Electron Eng. 2017;18:153–79.
31. Huysentruyt K, Kjoersvik O, Dobracki P, Savage E, Mishalov E, Cherry M, et al. Validating intelligent automation systems in pharmacovigilance: insights from good manufacturing practices. Drug Saf. 2021;44:261–72.
32. Ghosh R, Kempf D, Pufko A, Barrios Martinez LF, Davis CM, et al. Automation opportunities in pharmacovigilance: an industry survey. Pharmaceut Med. 2020;34:7–18.
33. Lewis DJ, McCallum JF. Utilizing advanced technologies to augment pharmacovigilance systems: challenges and opportunities. Ther Inn Reg Sci. 2020;54:888–99.
34. Routray R, Tetarenko N, Abu-Assal C, et al. Application of augmented intelligence for pharmacovigilance case seriousness determination. Drug Saf. 2020;43:57–66.
35. Comfort S, Dorrell D, Meireis S, Fine J. MOdified NARanjo causality scale for ICSRs (MONARCSi): a decision support tool for safety scientists. Drug Saf. 2018;41:1073–85.
36. Botsis T, Nguyen MD, Ball R, et al. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. J Am Med Inform Assoc. 2011;18:631–8.
37. Botsis T, Nguyen MD, Woo EJ, Buttolph T, Winiecki S, Ball R. Vaccine Adverse Event Text Mining (VaeTM) system for extracting features from vaccine safety reports. J Am Med Inform Assoc. 2012;19:1011–8.
38. Botsis T, Ball R. Automating case definitions using literature-based reasoning. Appl Clin Inform. 2013;4:515–27.
39. Botsis T, Woo EJ, Ball R. The contribution of the vaccine adverse event text mining system to the classification of possible Guillain-Barre syndrome reports. Appl Clin Inform. 2013;4:88–99.
40. Botsis T, Woo EJ, Ball R. Application of information retrieval approaches to case classification in the vaccine adverse event reporting system. Drug Saf. 2013;36:573–82.
41. Ball R, Botsis T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? Clin Pharmacol Ther. 2011;90:271–8.

42. Botsis T, Ball R. Network analysis of possible anaphylaxis cases reported to the US Vaccine Adverse Event Reporting System after H1N1 influenza vaccine. Stud Health Technol Inform. 2011;169:564–8.
43. Botsis T, Scott J, Goud R, Toman P, Sutherland A, Ball R. Novel algorithms for improved pattern recognition using the US FDA adverse event network analyzer. Stud Health Technol Inform. 2014;205:1178–82.
44. Markatou M, Ball R. A pattern discovery framework for adverse event evaluation and inference in spontaneous reporting systems. Stat Anal Dat Min. 2014;7:352–67.
45. Botsis T, Scott J, Woo EJ, Ball R. Identifying similar cases in document networks using cross-reference structures. IEEE J Biomed Health Inform. 2015;19:1906–17.
46. Wang W, Kreimeyer K, Woo EJ, Ball R, Foster M, Pandey A, et al. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. J Biomed Inform. 2016;62:78–89.
47. Wunnava S, Qin X, Rundensteiner EA, et al. Towards transforming FDA adverse event narratives into actionable structured data for improved pharmacovigilance. Proc Symp Appl Comput (SAC). 2017;2017:777–82.
48. Wunnava S, Qin X, Kakar T, Kong X, Rundensteiner EA, Sahoo SK, et al. One Size does not fit all: an ensemble approach towards information extraction from adverse drug event narratives. In: Proceedings of the 11th international joint conference on biomedical engineering systems and technologies—volume 5: HEALTHINF; 2018. pp. 176–188.
49. Botsis T, Foster M, Arya N, Kreimeyer K, Pandey A, Arya D. Application of natural language processing and network analysis techniques to post-market reports for the evaluation of dose-related anti-thymocyte globulin safety patterns. App Clin Inform. 2017;8:396–411.
50. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform. 2017;73:14–29.
51. Baer B, Nguyen M, Woo EJ, Winiecki S, Scott J, Martin D, et al. Can natural language processing improve the efficiency of vaccine adverse event report review? Meth Inform Med. 2016;55:144–50.
52. Botsis T, Jankosky C, Arya D, et al. Decision support environment for medical product safety surveillance. J Biomed Inform. 2016;64:354–62.
53. Kreimeyer K, Menschik D, Winiecki S, et al. Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems. Drug Saf. 2017;40:571–82.
54. Han L, Ball R, Pamer CA, et al. Development of an automated assessment tool for MedWatch reports in the FDA adverse event reporting system. J Am Med Inform Assoc. 2017;24:913–20.
55. Muñoz MA, Dal Pan GJ, Wei YJ, et al. Towards automating adverse event review: a prediction model for case report utility. Drug Saf. 2020;43:329–38.
56. Ly T, Pamer C, Dang O, Brajovic S, et al. Evaluation of natural language processing (NLP) systems to annotate drug product labeling with MedDRA terminology. J Biomed Inform. 2018;83:73–86.
57. Pandey A, Kreimeyer K, Foster M, Botsis T, Dang O, Ly T, et al. Adverse event extraction from structured product labels using the event-based textmining of health electronic records (ETHER) system. Health Inform J. 2019;25:1232–43.
58. Bayer S, Clark C, Dang O, et al. ADE Eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. Drug Saf. 2021;44:83–94.
59. Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, et al. Information visualization platform for post-market surveillance decision support. Drug Saf. 2020;43:905–15.
60. Wang X, Xu X, Tong W, Roberts R, Liu Z. InferBERT: a transformer-based causal inference framework for enhancing pharmacovigilance. Front Artif Intell. 2021;4: 659622. https://doi.org/10.3389/frai.2021.659622.
61. Liu Z, Roberts RA, Lal-Nag M, et al. AI-based language models powering drug discovery and development. Drug Discovery Today. 2021;26(11):2593–607.
62. Ding Y, Markatou M, Ball R. An evaluation of statistical approaches to post marketing surveillance. Stat Med. 2020;39:845–74.
63. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. 2010;19(Suppl 3):i68–74.
64. Mindell DA. Our robots, ourselves: robotics and the myths of autonomy. New York NY: Viking; 2015. ISBN: 978-0-525-42697-4.