Artificial intelligence in pharmacovigilance

CIOMS Working Group report
Draft, 1 May 2025

This report was posted for comment on 1 May 2025 at: https://cioms.ch/working_groups/working-group-xiv-artificial-intelligence-in-pharmacovigilance/.

Please note that the layout will be improved in the final version, and best efforts will be made to correct remaining typographical and/or grammatical errors, as well those pertaining to references.

Permissions are being sought to reproduce some of the illustrative materials included in this report. We welcome responses from organisations that own any of these materials and have not yet been contacted in this regard.

Please submit your comments using the form posted on the CIOMS website at https://cioms.ch/working_groups/working-group-xiv-artificial-intelligence-in-pharmacovigilance.

The timeline for submission of comments is 6 June 2025.

Thank you.

Suggested citation: Artificial intelligence in pharmacovigilance. CIOMS Working Group report. Geneva, Switzerland: Council for International Organizations of Medical Sciences (CIOMS), 2026.

DOI *number tbc*

This publication is freely available on the CIOMS website at:
https://cioms.ch/publications/product-category/recently-published/

Acknowledgements

(to follow)

1

Table of contents

List of figures and tables

List of figures

List of tables

1

Abbreviations

ADR	Adverse Drug Reaction
AE	Adverse Event
AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
AIDA	Artificial Intelligence and Data Act (Canada)
ALTAI	Assessment List for Trustworthy Artificial Intelligence
ATC	Anatomical Therapeutic Chemical
BLEU	Bilingual Evaluation Understudy
CDC	Centers for Disease Control and Prevention
CIOMS	Council for International Organizations of Medical Sciences
CNN	Convolutional Neural Networks
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSV	Computerized System Validation
EEA	European Economic Area
EHR	Electronic Health Records
EMA	European Medicines Agency
ETHER	Event-based Text-mining of Health Electronic Records
EU AI Act	European Artificial Intelligence Act
EU	European Union
FAERS	Food and Drug Administration Adverse Event Reporting System (USA)
GAMP 5	Good Automated Manufacturing Practice 5
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
GSD	Global Safety Database
GVP	Good Pharmacovigilance Practices
GxP	Good [x] Practices
HIC	Human-in-Command
HIPAA	Health Insurance Portability and Accountability Act
HITL	Human-in-the-Loop
HOTL	Human-on-the-Loop
ICH	The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

ICSR	Individual Case Safety Report
InfoViP	Information Visualization Platform
IRB	Institutional Review Board
ISIC	International Skin Imaging Collaboration
ISPE	International Society for Pharmaceutical Engineering
IT	Information Technology
KPI	Key Performance Indicator
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Models
MAH	Marketing Authorisation Holder
MedDRA	Medical Dictionary for Regulatory Activities
MHRA	Medicines and Healthcare products Regulatory Agency (UK)
ML	Machine Learning
ML-DSF	Machine Learning-Enabled Device Software Functions
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
OTC	Over the Counter
PASS	Post-Approval Safety Studies
PD	Pharmacodynamic
PDMP	Prescription Drug Monitoring Program
PHI	Protected Health Information
PK	Pharmacokinetic
PPV	Positive Predictive Value
PSMF	Pharmacovigilance System Master File
PSUR	Periodic Safety Update Reports
PV	Pharmacovigilance
QA	Quality Assurance
RCT	Randomized controlled trial
RMF	Risk Management Framework
RPA	Robotics Process Automation
RWD	Real-World Data
RWE	Real-World Evidence
SARAH	Smart Artificial Intelligence Resource Assistant for Health
SHAP	Shapley Additive exPlanations
SME	Subject Matter Expert

SmPC	Summary of product characteristics
SOP	Standard Operating Procedures
US FDA	U.S. Food and Drug Administration
UTC	United Therapeutics Corporation
WHO	World Health Organization
xAI	Explainable Artificial Intelligence

Preface

The Council for International Organizations of Medical Sciences (CIOMS) has played a pivotal role in the advancement of modern pharmacovigilance (PV) by developing guidelines that address ethical and scientific aspects of drug development and safety. Notably, CIOMS has published guidance documents that have supported a structured approach for the collection and reporting of adverse drug reactions (ADRs) in addition to guidance on practical aspects of signal detection in PV, fostering international collaboration and standardization in drug safety monitoring.

The thalidomide tragedy of the early 1960s exposed severe deficiencies in global drug safety practices, highlighting the need for comprehensive data collection and international harmonization. In response, the World Health Organization (WHO) established the Program for International Drug Monitoring in 1968, initiating efforts to share individual case reports between countries and harmonize data practices. Building on these foundational efforts, the late 1980s and the 1990s saw pivotal CIOMS reports like the *Monitoring and Assessment of Adverse Drug Effects* (1985) and the *International Reporting of Adverse Drug Reactions* (1987), both by the CIOMS Working Group I, and the *Current Challenges in Pharmacovigilance: Pragmatic Approaches* (1999) by the CIOMS Working Group V. Subsequent CIOMS Working Group reports and the establishment of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) aimed to address the fragmented approaches to drug safety identified decades earlier, providing a framework for standardized adverse event collection and reporting in addition to signal detection processes.

Advancements in technology have transformed PV, with artificial intelligence (AI) playing an increasing role in healthcare. At this juncture, regulatory harmonization and transparency remain critical to leveraging AI effectively, ensuring that data is reliable and that AI systems deliver accurate and trustworthy results, while ensuring safe and responsible AI use.

As AI continues to evolve and impact biomedical research, its increased integration in and impact on PV practice is inevitable. AI's potential for transformative disruption, compels us to engage in critical discourse: how do we wish to see AI developed, validated, and deployed within this domain?

Since the CIOMS expert Working Group XIV on AI in PV was established in early 2022, there has been significant progress in the field, marked by the rapid development and widespread availability of generative AI (GenAI). While there is growing interest in exploring GenAI for PV applications, we recognize the need to focus on its appropriate use, which brings specific challenges in highly regulated domains such as PV, and we look to distinguish where possible and beneficial from general issues of AI use. In light of this swiftly evolving context, this report aims to offer a general framework of principles and good practices for developing and using AI in PV. Rather than offering technical guidance, the aim is to ensure continued relevance as AI capabilities advance. The report focuses on applications that are specific to PV or issues of particular importance to PV rather than general use of AI, highlighting issues of high importance or priority to PV, even if they may not have widespread attention or relevance in other contexts.

This report aims to provide guidance to those working in PV, in addition to organisations and vendors developing AI solution for the PV domain.

Executive summary

The CIOMS report on artificial intelligence (AI) in pharmacovigilance (PV) tackles a rapidly emerging cross-disciplinary field that is at the intersection of PV, computer science, regulation, law, medicine, human rights, psychology and social science. Consequently, just as with medicinal products, it is important to establish the approved indications, posology, side effects, and warnings and precautions for use of AI in PV. This must be clearly defined and understood by many people from diverse backgrounds to propel research and practical implementation in an effective, safe and responsible manner. The diverse pool includes PV professionals, researchers, and decision makers working in PV in the industry, government, academia and software vendors, who are developing AI solutions in the PV space, including signal management and all aspects of Individual Case Safety Report (ICSR) processing. This report provides the requisite terminology and conceptual understanding to actively engage in this space, either by participating in the applied scientific research and public discourse, or by performing evaluations and making decisions at their respective organizations.

Perhaps more than other Council for International Organizations of Medical Sciences (CIOMS) report topics, the potential adverse effects of AI in PV and related points are major elements of our key results because rapidly evolving, advanced and often opaque technologies may generate a rush of excited promotion and initial over-estimation of utility, observed in so called technology “hype cycles”, that does not correspond with the practical realities. There is a corresponding safety net of core guiding principles for human protection elaborated by multiple organizations, through which AI PV must grow. This report provides a set of guiding principles and corresponding organizations that have elaborated each one. These principles form the bulk of the report: a risk-based approach, human oversight, validity and robustness, transparency, data privacy, and governance and accountability. Key points to consider for these guiding principles are elaborated throughout the report and summarized concisely below.

Similar to prior CIOMS reports, this one benefits from a consensus position from multiple stakeholders, including those based in regulatory agencies, academia, and industry. The Working Group (WG) recognized that the field of AI is progressing so rapidly, that a prescriptive document would likely be quickly outdated. Instead, the WG decided to focus upon a set of common principles that were expected to be useful for years to come for PV professionals. PV is but one of a myriad of AI applications that are now transforming many aspects of modern life. As such, this reports benefits as well from the increasing interest in AI by national governments, several of which have issued legislation and guidances not only on AI in drug development but also more broadly on the general use of AI.

Risk-based approach. Integrating AI into PV processes needs to take into account the potential inaccuracies and variability of AI algorithms, and corresponding impacts on the safety and well-being of individuals and society. The level of risk, and corresponding intensity of required oversight, will be a function of how high stakes the decision made by the AI is and whether the machine is intended to be used in an unchecked stand-alone mode or with human-machine interaction. A sound risk-based approach, in which the human oversight in the development and deployment of AI is commensurate with these risks, enables organisations to make the most of AI capabilities while ensuring that neither patient safety nor PV stakeholders are adversely affected. The risk-based approach applies to the human oversight modalities, the validity and robustness strategy, the level of transparency, and the efforts to uphold fairness and equity, and data privacy. The risk assessment should consider the AI system itself, the context of use, and the potential impact and likelihood of risks materialising. A risk-based approach should be reviewed at regular intervals and adapted if needed.

Human oversight. Human oversight supports performance optimization of AI in PV and increases trustworthiness and accountability. The extent and nature of human oversight for an AI solution

should be risk-based, incorporating quality assurance principles. The human oversight might be "human-in-the-loop" meaning that the decision is the end results of a human-machine interaction, while in "human-on-the-loop", the machine autonomously makes a decision or otherwise returns a result that is checked by a human. Human oversight is necessary to define fit-for-purpose levels of performance for the intended task (i.e. validity). It involves predefining acceptable performance levels, selecting appropriate data for model development in a realistic setting, an ongoing quality assessment process and retraining of the model as needed. Increased use of automation and AI in PV will transform traditional roles and competencies, requiring appropriate change management and training strategies.

Validity & Robustness. PV stakeholders must learn to critically appraise proposed AI solutions. Performance evaluation must demonstrate acceptable and robust results for intended use under realistic conditions. Such an evaluation should be both qualitative and quantitative, and a cross-disciplinary exercise and span a diverse range of relevant examples. Evaluations should use a sufficient representation of relevant data types to detect biases, promote adequate and generalizable performance across the intended deployment domain, assess usability, and identify circumstances associated with underperformance. Enrichment strategies to obtain representative test sets with high enough prevalence of the outcome may be required. Special care should be taken to ensure that performance evaluation results generalize to real-world settings.

Transparency. Declaring when and how AI solutions are used is critical for building trust among stakeholders. The nature of AI solutions deployed for core PV tasks should be communicated, including model architectures, expected inputs and outputs, and the nature of human-computer interaction. To fully characterize an AI solution's effectiveness and limitations, performance evaluation results should describe the scope and nature of the test set(s) used including reference standards and sampling strategies. Performance metrics should be relevant for the intended tasks, compared with relevant benchmarks, and complemented by qualitative review of representative examples of correct and incorrect output. Explainability is an important concept relevant to those models whose internal decision pathways are so intricate and non-linear that they remain inscrutable even to technically literate persons – so called black boxes of the first kind. Explainable AI are a set of techniques that "look under the hood" and return plausible hypothesis about these pathways – roughly how the black box arrived at its outputs. To be able to do this can be advantageous to model building/trouble shooting, building trust, establishing auditability and accountability, including providing a basis for a human to challenge an AI result that may be adversely impacting them, and regulatory compliance and scientific hypothesis generation. However, explainable AI methods have limitations, and they only provide plausible hypothesis, but are no guarantee that the AI in fact used the hypothesized decision pathways.

Data Privacy. The ethical framework to evaluate the use of AI in PV is embedded within the standard principles for research activities involving human subjects. A crucial principle for the use of AI in routine PV is the sanctity of data privacy. With the increasing power of both the hardware and software that power AI, there is a vast potential to build large, linked databases, and the potential inherent in LLMs for patient reidentification. These may pose an ongoing challenge to the traditional safeguards that protect data privacy. In this context, there are multiple opportunities to reveal highly sensitive personal and health information to a broad, cross-disciplinary range of stakeholders throughout the AI development and deployment workflow. Consequently, countries have been enacting legislation and guidances intended to protect these data. PV professionals should recognize that existing procedures used to assure regulatory compliance may need to be reevaluated due to the heightened risks of GenAI to compromise data privacy.

Fairness & Equity. Supporting fairness and equity, avoiding propagating or amplifying harmful explicit biases underserving certain subpopulations, discrimination and inaccurate results during model development and deployment are regulatory and ethical imperatives. Equity may be advanced by taking measures to assure that AI in PV returns outputs that are relevant to populations

anticipated to have exposure to the specific medicinal product being evaluated. Screening, identifying and excising explicit or potential bias when possible is key to mitigating risk, determining AI applicability and limitations, and defining acceptable performance. Scrutinize training and performance evaluation of reference data sets for adequate representation and evaluate performance in relevant subgroups when possible. Inadequate reference data is often the cause of inadequate fairness and equity.

Governance & Accountability. Robust governance and clear accountability are crucial for the success of AI initiatives. These principles help ensure that AI systems are used safely, responsibly and ethically, and in compliance of all applicable legal and regulatory mandates while fostering trust and transparency among stakeholders. Clearly defined roles and responsibilities are crucial to enable all stakeholders to understand their accountability and obligations in order to effectively oversee AI systems.

As AI technology evolves, governance and accountability frameworks will need to be adapted. New risks and challenges will emerge, requiring updated principles and practices. Continuous review and adaptation are essential for staying ahead of these changes. This includes the adaptation refinement of the proposed grid for practical use.

Future considerations for development and deployment of artificial intelligence in pharmacovigilance. Increasing deployment of AI in PV is expected to prioritize and accommodate rapid data collection, assessment and reporting for signal detection in real or quasi real time. This may also be accompanied by a relative shift from warm-start to cold-start prediction scenarios (i.e. post-approval to early-stage drug development). This could fundamentally change the way we work to take advantage of these technological advances, for example, streamlining processes and causing changes in the wider healthcare environment and beyond, including patient privacy. We also expect to see increasing deployment of AI in PV in the clinic, where it will support primary, secondary and tertiary prevention of adverse drug reactions. The extent to which humans remain in or on the loop will be determined by the nature of the task (e.g. routinized tasks versus those requiring expert clinical and scientific judgement), consistent with the elaborated risk-based approach, but it is possible that some AI-based expert systems will eventually develop refined medical and scientific judgement.

It is critical that the guiding principles outlined in this report remain as core considerations, but they will need to evolve and adapt with advancements and application of AI in PV and medicine in general. This is to ensure AI use in PV remains unbiased, transparent, and secure to prevent misuse or accidental harm. The appropriate human oversight, including regulatory and ethical safeguards, will be as crucial as the technological advancements being applied.

Chapter 1: Introduction

An artificial intelligence (AI) system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[1] Different AI systems vary in their levels of autonomy and adaptiveness after deployment. In the context of pharmacovigilance, the use of AI systems and activities is aimed at enhancing drug safety monitoring, patient safety and regulatory compliance. PV is practiced not only at pharmaceutical companies, health authorities, drug monitoring centres and academia, but also in the clinic, and AI is finding applications to PV in all these settings.[2]

An AI solution is designed to address specific objectives within PV. The overall AI solution could be developed with one or many AI systems. An AI system encompasses not only the model itself but also includes the components necessary for utilization including user interfaces and data processing pipelines. At the core of these systems are AI models. These models utilize parameters to learn relationships within data, enabling the systems to adapt and improve model performance over time.

Simpler AI systems, such as statistical methods for signal detection, have been widely utilized in PV for decades.[3] However, the past decade(s) have seen drastic improvements in AI capabilities, particularly in image analysis and natural language processing. These advancements have resulted in a significant increase in their use. In addition, tremendous and continual advances in computing power and model architectures have enabled the development and aggregation of large electronic databases with potential for linkage. These have enabled the field of AI to be applied to an increasing number of disciplines, including the life sciences.[4] Within the life sciences, AI is being applied to a growing number of areas, such as drug discovery and development, medical imaging and diagnostics, genomics, precision medicine, public health, and healthcare delivery.[5]

Partly due to advances in AI, the pharmaceutical field is poised for rapid transformation across clinical, regulatory and PV practices, aiming to streamline end-to-end processes to accelerate product development and market delivery. Similarly, there is a growing emphasis focusing on enhancing clinical and post marketing safety and risk management activities to enable proactive identification (or even prediction) of safety signals and benefit-risk evaluation. In the clinic, AI is being tested or deployed for early diagnosis (and thus secondary and tertiary prevention) of various adverse drug reactions. Examples include early detection of hydroxychloroquine retinopathy,[6] digoxin toxicity,[7] and drug-induced movement disorders in Parkinsons patients.[8]

These advancements leverage massive integrated datasets and inductive logic, enabling AI models to make informed decisions by utilizing accumulated data, rather than relying solely on explicit rules or human intervention. This approach facilitates the development of AI tools that provide new, improved, or complementary solutions. A critical enabler for AI success within PV will be the ability to link and analyze large volumes of heterogeneous data of varying quality from diverse data sources, such as electronic health records (EHRs), claims databases, registries, Internet of Things (IoTs), and connected devices. The ability to leverage health data can lead to potentially faster development of new treatments, improved patient outcomes, and reduced healthcare costs, including the potential for unlocking novel, useful, and actionable insights that might not have been identified otherwise. Hence, there is an acute need to effectively communicate the key importance of data access to support patient safety outcomes.

Incorporating AI into PV necessitates a thorough assessment of its potential benefits and risks, helping stakeholders understand its implications for existing practices. Given the rapid pace of

change, this document does not prescribe specific uses for AI in PV but rather establishes and promotes guiding principles for utilizing AI including ML.

The start of systematic safety surveillance predated the advent of the internet and widespread electronic reporting capabilities. As such, it was a largely manual process that relied upon computing for purposes such as summarizing data.

ICSRs are a key component of PV and remain a cornerstone of post-market safety surveillance as they provide crucial safety information for an approved pharmaceutical product, which is important to mitigate patient harm when assessed within a broader signal management system.

The processing of ICSRs involves several steps: collection, triage, data entry, quality review, medical assessment, a with further transmission to other safety databases (e.g. regulatory authorities). As the number of product approvals and the patient exposure grow, so do the number of reported adverse events. The increased volume of ICSRs, coupled with stringent safety regulations, create significant challenges in ICSR processing and compliance.

Once a signal is detected as a result of individual or aggregate analysis of adverse event reports, it needs to be systematically investigated through sequential steps, which include signal triage, validation, and, based on scientific assessment, formal evaluation using independent data sets, such as hypothesis-testing research studies.[9] Such investigation must be conducted in an integrated, holistic fashion with all available scientific evidence and logic, offering wider opportunities for use of AI for data insights (see Figure 1).

Traditional PV methods for analysis of adverse event reports include:[10,11]

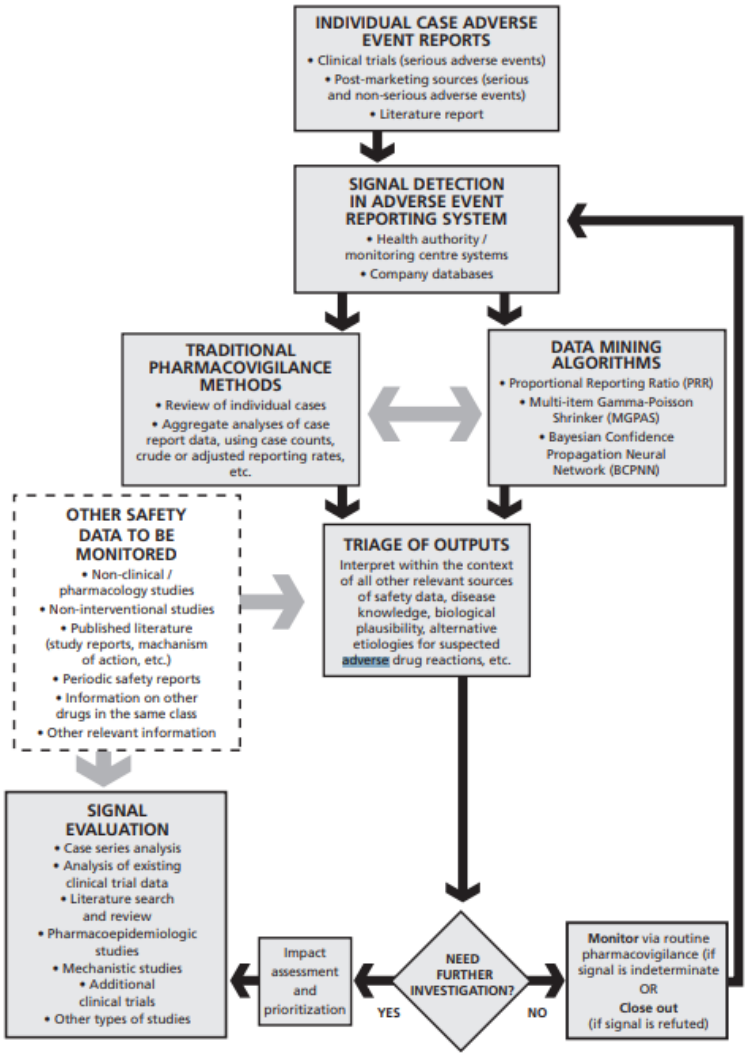
- Review of individual cases safety reports (ICSRs) or case series in a PV database or in published medical or scientific literature; and
- Aggregate analyses of case reports using absolute case counts, simple reporting rates, proportions or estimated exposure-adjusted reporting rates.

While ICSRs are fundamental to PV, other data streams are also considered throughout the PV lifecycle. These streams may be directly linked or conceptually related and include pharmacokinetic/ pharmacodynamic (PK-PD) data, other real-world data (RWD), literature, and information from clinical trials.

Once safety concerns (important identified risks or important potential risks) are identified, it is essential to communicate them appropriately to a wide range of stakeholders. This is achieved through documents such as aggregate reports, risk management plans, labelling information and Dear Healthcare Professional communications (DHCPs).

[Figure 1]: Traditional signal management process

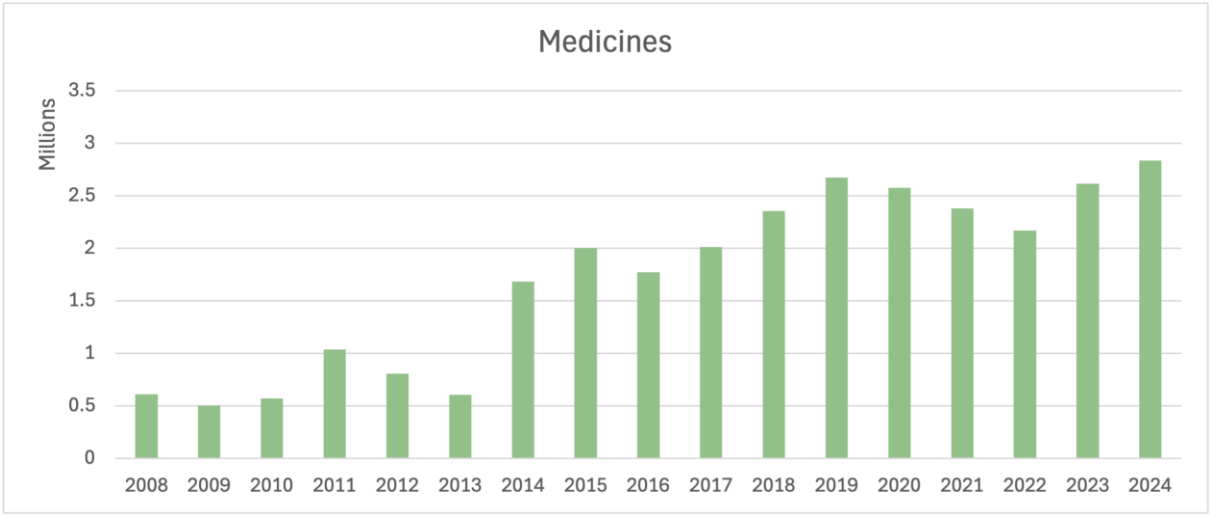
Source[12]



The COVID-19 pandemic has further emphasized the need for advanced methods in PV, as it has led to a significant rise in safety reports (see Figures 2 and 3).[13,14] As public awareness and expectations regarding drug safety continue to rise, there is a greater demand for robust PV systems that can effectively identify and mitigate potential risks associated with medications.

Figure 2: Growth over time of VigiBase, the World Health Organization global database of adverse event reports for medicines and vaccines

Source: VigiBase accessed April 2025.



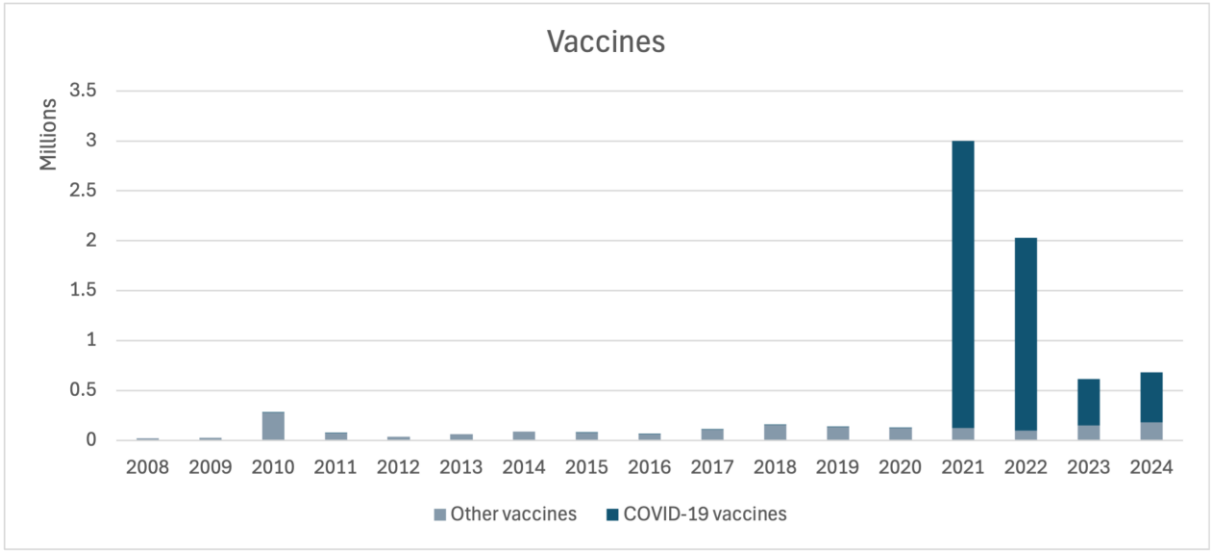
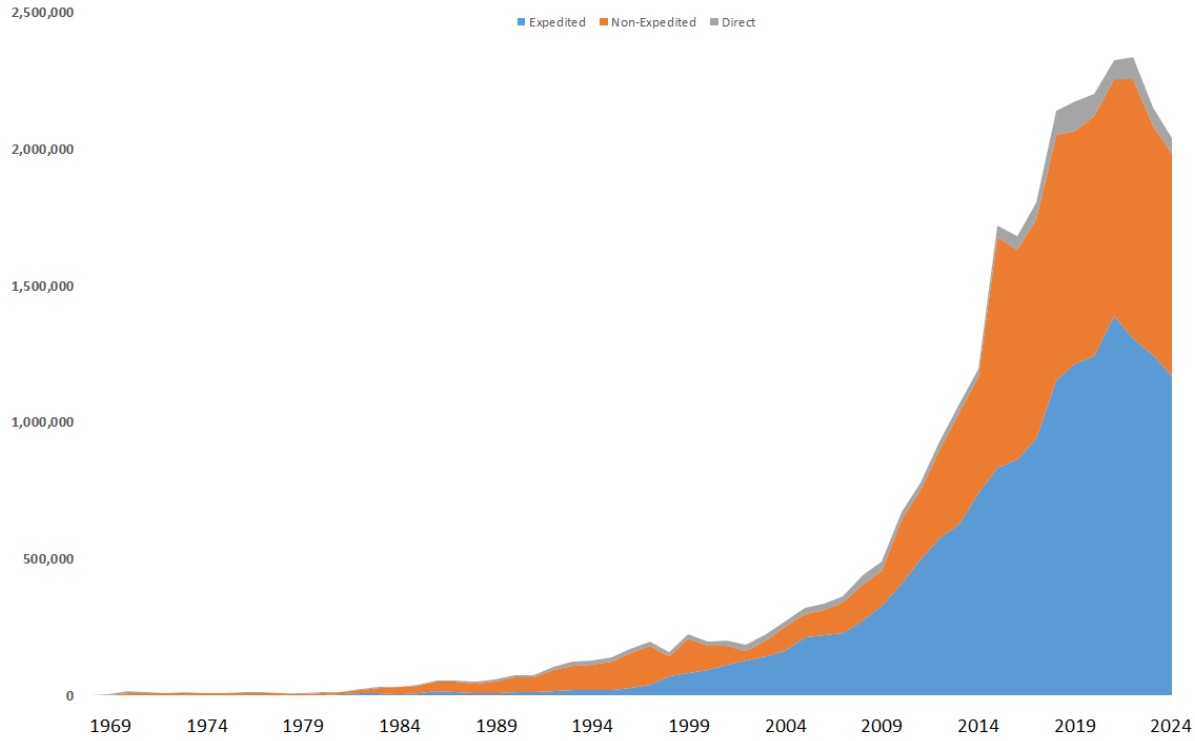


Figure 3: Growth over time of the FDA Adverse Event Reporting System (FAERS) database

Source: Constructed using FDA FAERS database.[15]



Nevertheless, the challenges of establishing and maintaining progressively more complex PV systems in a globally diverse and evolving regulatory environment are increasing. There is a need to rethink traditional PV strategies based on existing pressures on the one hand (e.g. managing increasing volumes and increasing regulatory complexity), and increasing and data sources on the other.

Technology solutions are already vital for the evolution of PV. While this notion of technology as a transformative enabler spans across all areas of product development, it is evident that applying innovative automation tools and processes to PV is no longer an option but an essential capability.

Rapid evolution of artificial intelligence

Traditional AI methods (e.g. K-means clustering, decision trees, support vector machines etc.) have traditionally been tailored for specific tasks, primarily utilizing supervised learning techniques. In contrast, deep neural networks such as BERT[1] have played a significant role in natural language processing, where they are pre-trained on large datasets and subsequently fine-tuned for specific applications delivering predictable outputs.

However, the landscape is evolving beyond this framework thanks to emerging technologies like Generative AI (GenAI). GenAI models are trained on expansive and varied text corpuses, often incorporating phases of human reinforcement learning. These models can perform specific tasks using sophisticated prompts, adopting zero-shot or few-shot learning techniques.

[1] BERT: Bidirectional Encoder Representations from Transformers

Scope

This document aims to guide those working in PV in addition to organisations developing AI solutions for the PV domain, such as medicinal product regulators, medicinal products industry professionals, software vendors, international and national PV organizations, researchers, and health care professionals.

This report proposes a broad framework of principles and best practices for integrating and implementing AI within PV, not technical guidance. Recognizing the rapid evolution and application of AI technology, the CIOMS Working Group XIV developed this document to guide the development and integration of AI tools into PV activities.

Our scope focuses on all aspects, direct and indirect, of the optimal collection, organization, analysis, and communication of ICSRs from any source, including RWD, medical literature, randomized controlled trials, and social media). Additionally, it includes productivity enhancers closely linked to PV, such as tools that improve querying of safety databases[16] or capabilities that enable faster, more effective, or consistent data entry into a safety database which indirectly contributes to better safety surveillance.[17]

The scope deliberately excludes broader healthcare data applications outside the direct purview of safety, such as pharmacoepidemiology and other real world evidence study designs and conduct that fall outside the realm of ICSRs. Similarly, the general use of AI as a productivity enhancer, if not directly connected to PV activities (e.g. for email support), is excluded, as considerations may differ.

The scope has been intentionally limited to provide a practical guidance organised as principles and their applications of AI in PV, rather than detailed guidance to ensure longevity. As AI is progressing extremely rapidly, future opportunities and considerations are described in a later chapter.

References

[1] *OECD (2024), "Explanatory memorandum on the updated OECD definition of an AI system", OECD Artificial Intelligence Papers, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/623da898-en.*

[2] *Hauben M. Artificial Intelligence in pharmacovigilance: Do we need explainability? Pharmacopidemiol Drug Saf. 2022;Dec;31(12):1311-1316. (Journal full text)*

https://doi.org/10.1002/pds.5501

[3] *Kompa B, Hakim JB, Palepu A, Kompa KG, Smith M, Bain PA, Woloszynek S, Painter JL, Bate A, Beam AL. Artificial intelligence based on machine learning in pharmacovigilance: A scoping review. Drug Safety. 2023;Apr46(4):433. (Journal full text) https://doi.org/10.1007/s40264-023-01273-9*

[4] *Webb S. Deep learning for biology. Nature. 2018;Feb22;554:555-557 (Journal abstract) https://doi.org/10.1038/d41586-018-02174-z*

[5] *Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American Journal of Medicine. 2019;Jul1;132(7):795-801. (Journal full text) https://doi.org/10.1016/j.amjmed.2019.01.017*

[6] *Wright T, Yan P, Easterbrook M. Machine learning to identify multifocal ERG deficits in patients taking hydroxychloroquine. Investigative Ophthalmology & Visual Science. 2019;Jul22;60(9):5959. (Journal full text)*

[7] *Chang DW, Lin CS, Tsao TP, et al. Detecting digoxin toxicity by artificial intelligence-assisted electrocardiography. Int J Environ Res Public Health. 2021;Apr6;18(7). (Journal full text) https://doi.org/10.3390/ijerph18073839*

[8] *Li MH, Mestre TA, Fox SH, Taati B. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. J Neuro Engineering Rehabil. 2018;Dec15:1-3. (Journal full text) HYPERLINK "https://doi.org/10.1186/s12984-018-0446-z"https://doi.org/10.1186/s12984-018-0446-z*

[9] *Practical Aspects of Signal Detection in Pharmacovigilance. CIOMS Working Group VIII report. Council for International Organizations of Medical Sciences (CIOMS), 2010.*

[10] *U.S. Food and Drug Administration. Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment. Guidance for Industry. March 2005. (Full text accessed 21 March 2025).*

[11] *Practical Aspects of Signal Detection in Pharmacovigilance: Report of CIOMS Working Group VIII 2010*

[12] *Practical Aspects of Signal Detection in Pharmacovigilance. CIOMS Working Group VIII report. Council for International Organizations of Medical Sciences (CIOMS), 2010.*

[13] *V. Këpuska and G. Bohouta, "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2018, pp. 99-103, doi: 10.1109/CCWC.2018.8301638.*

[14] *Alvager T, Smith TJ, Vijai F. Neural-network applications for analysis of adverse drug reactions. Biomed Instrum Technol. 1993 Sep-Oct;27(5):408-11. PMID: 8220635.*

[15] *U.S. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS) Public Dashboard. (Webpage accessed 21 March 2025)*

[16] *Painter JL, Chalamalasetti VR, Kassekert R, Bate A. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. JAMIA open. 2025;Feb;8(1):ooaf003. (Journal full text) https://doi.org/10.1093/jamiaopen/ooaf003*

[17] *Painter JL, Mahaux O, Vanini M, Kara V, Roshan C, Karwowski M, Chalamalasetti VR, Bate A. Enhancing drug safety documentation search capabilities with large language models: a user-centric approach. In 2023 International Conference on Computational Science and Computational Intelligence (CSCI). 2023;Dec13:49-56. (Journal abstract) https://doi:10.1109/CSCI62032.2023.00015*

401

402

Chapter 2: Landscape analysis

Use of artificial intelligence in pharmacovigilance to date

AI may directly or indirectly impact all aspects of PV (see Figure 1: Traditional signal management process). In this chapter, we discuss solutions that incorporate elements of AI and have been developed or deployed for a variety of tasks across PV, focusing on those that have been implemented specifically for PV or have accounted for attributes or features especially prominent in PV applications. For example, AI solutions for general translation tasks are out of scope, but PV specific translations, e.g. of adverse event reports, are in scope. Further, the landscape analysis reflects the overall scope of the document, which focuses on collection, processing, and analysis of adverse event reports. For this reason, research on AI methods to identify covariates for inclusion in propensity score models for epidemiological studies are out of scope. Rather than seeking to provide an exhaustive enumeration, the aim here is to illustrate the range and variety of current applications. Additional examples can be found in recent review articles.[18] The reader is also referred to the many perspectives and commentaries that discuss the use of AI in PV[19,20,21,22] and the cautionary notes that have been provided.[23]

Adverse event reporting and capture

AI solutions have been proposed for a variety of tasks related to natural language processing of social media content to identify references to (personal experiences of) medicine use and adverse events. These tasks include identifying relevant posts,[24,25] identifying relevant parts of such posts,[26] normalizing descriptions of adverse events or medicinal products within such posts to standardized terminologies like MedDRA or ATC,[27] and classifying the relationship between adverse events and drugs mentioned in the same posts.[28] Similarly, screening the scientific literature for adverse events is an investigated AI application.[29]

Individual Case Safety Report Processing

An area of ICSR processing where AI solutions have been in routine use by some organizations since at least the 2010's is duplicate detection, which relates to the identification of multiple unlinked records describing the same adverse event in a particular patient.[30] Duplicate detection methods based on ML and probabilistic record linkage have been implemented for VigiBase,[31] FAERS,[32] and EudraVigilance.[33] The use of natural language processing to improve duplicate detection by extracting and incorporating information from free text has also been explored.[34,35] Rule-based methods are more widely used and would be easier to implement but do not perform as well.[36]

Another area where AI has been used to support ICSR processing is in the encoding of information on adverse events[37,38] or medicinal products[39] in standard terminologies based on verbatim fields and / or free-flowing case narratives. Natural language processing has also been applied to extract relevant information from case narratives and map it to structured fields[40,41,42,43,44,45] and for ICSR translation.[46] A major challenge is the lack of the data homogeneity[47] that could be improved by adhering to existing common data models and standards that have already demonstrated value in the PV domain, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model for longitudinal observational health data[48] and Fast Healthcare Interoperability Resources (FHIR), a standard for health care data exchange, published by HL7®, especially if they are extended to support a broader spectrum of PV cases.[49,50]

Several organisations who process large numbers of case reports have also automated repetitive, labour-intensive tasks using so-called robotic process automation (RPA) technologies.[51] These operate on the user interface of other computer systems like humans would.[52]

Other applications of AI solutions during ICSR case processing include methods have been developed to help support triage incoming reports for human review,[53,54] individual case causality assessment,[55] and automated redaction of person names in case narratives.[56]

Signal detection and analysis

The earliest examples of real-world use of (simple) AI solutions in PV are from the late 1990s. At this point, disproportionality analysis, first conceptualized in the 1970s,[57] began to be implemented as part of triage algorithms to help direct the attention of PV specialists in their analysis of large national and international collections of individual case reports.[58,59,60,61] Since then, various incremental improvements have been introduced and evaluated including automated adjustment for confounding through e.g. regression,[62,63] or propensity scores,[64] extensions to drug-drug interactions,[65,66,67,68,69,70] and other possible risk factors for adverse reactions.[71,72] Methods to detect adverse events associated with the production process or with substandard or counterfeit medicines have also be explored.[73,74,75] In addition, there have been efforts to develop predictive models for statistical signal detection that account for other aspects of a case series, such as its geographic spread and the quality and content of individual reports,[76] the time-to-onset of the reported reactions,[77] or a combination of e.g. Naranjo scores and the proportions of reports on a drug-adverse event combination coming from healthcare professionals and marketing authorization holders, respectively.[78]

Natural language processing has been applied to mine regulatory information,[79] scientific literature, and clinical notes[80,81,82,83] for information on already known/unknown and potentially serious adverse effects. This may support and streamline decision making, especially during early signal assessment and prioritization.

Some published AI-based signal detection exercises provide tantalizing glimpses of how elegant AI solutions may uncover truly novel adverse events.[84] At the same caution is warranted in that highly technical and elegant methods may be associated with overly-optimistic interpretations of, and corresponding messaging about, the results, which may disseminate widely.[85]

Several organizations have developed predictive models for ICSR prioritization to assess causality associations between drugs and adverse events[86] and/or inform a regulatory action.[87] These can be used to prioritize reports for human review during signal assessment and/or case processing. Semantic search has been developed for case narratives to support signal detection and assessment[88,89] and there have been efforts to provide machine learning-based decision support for signal validation[90] and to automatically visualize relevant information on case reports to facilitate human review during signal assessment.[91] Machine learning has been used to help estimate the proportion of patients with a genotype associated with drug toxicity based on the phenotypical manifestations reported in ICSRs.[92]

Applications of unsupervised learning have been developed to support signal detection and analysis, especially seeking to bring together reports describing similar or related adverse events. These include network analyses of adverse events (and to a lesser extent drugs)[93,94,95,96,97] cluster analysis of adverse event reports,[98] and data-driven derivations of semantic representations of adverse events and drugs.[99,100]

Datasets with information about drug side effects and indications such as DrugBank[101] and SIDER,[102] as well as those with information on pharmacology and chemical structures such as Bio2RDF,[103] have been leveraged to enhance PV signal detection and analysis,[104,105] or derive knowledge graphs that can serve as downstream inputs for AI-based predictive signal detection.[106]

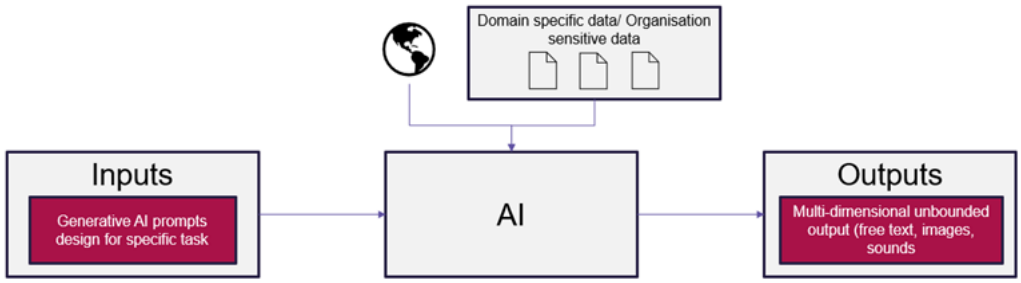
Early applications of generative AI in pharmacovigilance

Early applications of generative AI in PV have started to be explored. Researchers have applied large language models to a variety of PV tasks and reported their experiences. Examples to date include use of LLMs to simplify the patient communication from a regulatory authority,[107] summarization for drug labelling documents,[108] named entity recognition in scientific literature and social media,[109] search of drug safety documentation,[110] Q&A for drug labelling,[111] PV context-aware generation of SQL code[112], and drafting follow-up letters to reporters.[113]

However, the landscape is evolving beyond this framework thanks to emerging technologies like GenAI. These models, developed from deep neural networks with up to hundreds of billions of network parameters provided with vast and opaque large text corpuses, give rise to boundless number of multi-modal inputs and outputs (see Figure 4), therefore deciphering useful from misleading outputs for the end user can become a challenge. The non-deterministic and 'black box' nature of such algorithms as well as the lack of ability to fully understand training can make developing and maintaining trust potentially harder as well as the ease of communicating such to external parties. As noise and biases may infiltrate the results, leading to potentially unreliable and misleading conclusions, necessitating guidance on how these technologies should be used appropriately in PV is needed where societal expectations are greater.

Figure 4: Multi- modal input models

Source: Variation of submitted article to Therapeutic Advances in Drug Safety



Examples of deployed AI solutions

Much of the research and development of AI solutions for PV to date has been experimental, with either no real-world deployment yet or only limited experimental use, for example in the form of pilot studies. However, **Error! Reference source not found.** presents examples of AI solutions that have been adopted for routine use in PV by various PV organizations and are described in the public domain. The deployment of AI solutions by pharmaceutical companies may on the other hand to a large extent be based on software vendor implementations, which are not described in the public domain.

Table 1: Examples of deployed artificial intelligence solutions in pharmacovigilance described in the public domain

Source: CIOMS XIV working group

AI solution	Pharmacovigilance context / database
Automated coding of medicinal products	VigiBase[114]
Duplicate detection	FAERS,[115] VigiBase[116]
Automated triages of individual case reports	Swedish Medical Products Agency[117], pharmaceutical companies[118]
Automated triages for quantitative signal detection	Databases of various regulatory authorities, international organizations, and pharmaceutical companies

Predictive models for quantitative signal detection	VigiBase,[119,120] Netherlands pharmacovigilance centre Lareb[121]
Adverse event cluster analysis for signal detection and assessment	VigiBase[122,123]
Literature surveillance for safety data	EudraVigilance Netherlands pharmacovigilance centre Lareb[124]

Regulatory considerations

Introduction

Since 2017, countries around the world have been developing national AI strategies in order to adapt to technological advancements and their impact on society and the economy (OECD).[125] Countries have developed different regulatory frameworks and guiding principles to ensure the ethical use and trustworthiness of AI systems, and legislation of AI are being implemented (i.e. EU AI Act, AIDA, US Algorithmic Accountability Act and Executive Order: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government) (OECD).[126] In addition, there have been published reflection and discussion papers on the use of AI in medicinal products by the EMA and FDA, as well as a draft guidance on AI use to support regulatory decision making for drug products by the FDA.

Guiding Principles for AI in Pharmacovigilance

There are numerous published guiding principles for safe and responsible use of AI by governments, regulatory bodies and international organisations such as the WHO and OECD, that have been reviewed by the CIOMS Working Group XIV. These publications all define guiding principles and recommend best practices for safe and responsible AI use in regulated fields; however, the majority of these publications were not developed specifically for PV. Furthermore, it should be acknowledged that some discretion was used to establish the guiding principles by the various organizations, as some of the principles were described in conjunction with other principles. Nonetheless, these principles can be applied to the field of PV. The table below provides an overall comparison of the guiding principles, and a non-exhaustive description of the principles is presented in Appendix 2:

Table 2: Comparison of CIOMS Working Group XIV guiding principles for artificial intelligence across regional and country government institutions, and international organizations

Source: CIOMS Working Group XIV

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[127,128]	**Australia**[129]	**Canada**[130]	**Singapore**[131]	**UK**[132]	**US**[133]	**PAHO**[134]	**WHO**[135]	**OECD**[136]
Human Oversight	✓	✓	✓		✓	✓	✓	✓	✓
Validity & Robustness	✓	✓	✓		✓		✓	✓	✓
Data Privacy	✓	✓			✓	✓	✓		
Transparency	✓	✓	✓	✓	✓		✓	✓	✓
Accountability	✓	✓	✓	✓	✓	✓	✓	✓	✓
Societal well-being	✓	✓		✓			✓	✓	✓
Environmental Well-being	✓	✓						✓	✓
Fairness & Equity	✓	✓	✓	✓	✓	✓	✓	✓	✓
Explainability	✓	✓		✓	✓	✓		✓	✓
Safety	✓	✓	✓		✓	✓		✓	✓
Governance	✓				✓			✓	

EMA Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle

On September 9, 2024, the European Medicines Agency finalized its Reflection paper on the use of AI in the medicinal product lifecycle.[137] The reflection paper addresses the use of AI/ML in the safe and effective development, manufacturing and use of medicines.

EMA advocates a risk-based approach for the development, deployment and monitoring of AI and ML tools throughout the system lifecycle. The paper uses the terms 'high patient risk' for systems affecting patient safety and 'high regulatory impact' for cases with a substantial impact on regulatory decision making. It is expected that applicants/marketing authorisation holders (MAHs) and developers of AI and ML systems will perform a regulatory impact and risk analysis. The level of scrutiny of the AI and ML systems will be dependent on the assessment of risk level and regulatory impact.

The paper provides technical and regulatory considerations on the use of AI and ML throughout the lifecycle of medicinal products, from drug discovery and development to post-authorisation settings. Specifically for PV, the paper foresees that AI/ML tools can effectively support activities such as adverse event report management and signal detection, in line with applicable GVP requirements. Applications within PV may allow a more flexible approach to AI/ML modelling and deployment than other domains, for example, to improve severity scoring of adverse event reports and signal detection. It is, however, the responsibility of the MAH to validate, monitor and document model performance and include AI/ML operations in the PV system, to mitigate risks related to all algorithms and models used.

Generally, the applicant or MAH is responsible for ensuring that all elements of the AI and ML applications (i.e. algorithms, models, datasets, and data processing pipelines) are fit for purpose and comply with GxP standards and current EMA scientific guidelines. Member State data protection authorities are responsible for the supervision and monitoring of data protection compliance of AI systems. Applicants or MAHs and developers are recommended to engage with EMA on experimental technology, especially for AI and ML models that may have a high impact on the regulatory decision making.[138]

The EMA is planning to develop further guidance on the use of AI in the medicines lifecycle, including in PV.[139]

FDA Discussion Paper on Using artificial Intelligence & Machine Learning in the Development of Drug & Biological Products

In May 2023, the US FDA published a discussion paper on "Using artificial Intelligence & Machine Learning in the Development of Drug & Biological Products".[140] The FDA acknowledges the increased use of AI/ML in the lifecycle of drug development with novel approaches in data mining, analyzing large multi-omics, PK/PD modeling, real world data, data collection from wearable devices and other datasets (e.g. *in vitro* and *in vivo* studies, mechanistic studies, and multi-organ chip systems. In the post-marketing safety surveillance, the FDA sees the potential to: i) automate the processing and prioritization of individual case safety report (ICSR) using AI/ML, due to the increasing volume of reports and complexity of data sources; ii) classifying ICSRs on the likelihood of causal relationship between the drug and adverse event; iii) determine the seriousness of the outcome of ICSRs; and iv) automate aggregate reports for multiple adverse events for a particular product.

FDA Draft Guidance on Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products

The FDA published a draft guidance titled "Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry and Other Interested Parties" in January 2025. It elaborates a risk-based credibility assessment framework for AI. The scope of the document focuses on the support of regulatory decision making pertaining to the safety, effectiveness, or quality for drugs. Out of scope are drug discovery or scenarios in which AI is deployed for operational efficiencies. The draft guidance is not a regulatory mandate, rather the FDA's current thinking and recommendations on AI use. It considers AI broadly, i.e. not limited to specific subsets of AI such as machine learning. There are three major segments of the draft guidance:

1. Establishing a risk-based credibility assessment framework (see also Chapter on Risk-based approach);
2. Lifecycle credibility maintenance;
3. Options for sponsors for engaging with the agency to discuss AI model development.

The risk-based credibility assessment framework has seven steps as below.

1. Define the question to be addressed by an AI model.
2. Define the "context of use (COU)" defined as "….the specific role and scope of the AI model used to address a question of interest." Importantly this includes whether the questions being answered, and any ensuing classifications or decisions, are based solely on the AI outputs versus the AI being used in conjunction with other information (i.e. "model influence"). This is important because it helps define the associated risk in the subsequent step.
3. Define model risk. This is determined by model influence as defined in COU and decision consequence - i.e. the consequences of an incorrect decision. The risk is highest when

the AI model is operating in a stand-alone capacity and incorrect decisions present a major hazard. The required level of oversight throughout the development and production cycle is positively correlated with the risk.

4. Develop a plan to establish AI model credibility within the COU.
5. Execution of the plan.
6. Document the results of the credibility assessment plan and discuss deviations from the plan.
7. Determine the Adequacy of the AI Model for the COU.[141]

FDA Emerging Drug Safety Technology Program (EDSTP)

The FDA has established the Emerging Drug Safety Technology Program (EDSTP) in June 2024 to engage with industry stakeholders on AI and other emerging novel technologies used in PV and the lifecycle of the drug product. The three goals of the EDSTP include discussion between industry and FDA, knowledge dissemination of emerging AI/ML models or other emerging novel technologies, and to inform potential regulatory or policy development within the context of PV.[142]

Guidance on use of Large Language Models

Since the release of ChatGPT on November 30, 2022, there has been significant work in exploring how generative AI could be adapted to a variety of tasks (such as text and image generation, coding, brainstorming, and research) for productivity gains. Given the potential use and broad applicability of GenAI, regulatory agencies and organizations have developed high level guides and best practices on the safe and responsible use of GenAI by their own staff and broader stakeholder groups, respectively, which aligns with established guiding principles for AI:

- Guiding principles on the use of large language models in regulatory science and for medicines regulatory activities (EMA[143]);
- Guide on the use of generative artificial intelligence (Canada[144]);
- Initial policy considerations for generative artificial intelligence (OECD[145]);
- WHO Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. (WHO[146]).

Guidelines for safe AI

Other related regulatory and international organization (e.g. WHO and OECD) published guidelines for safe AI include:

- Regulatory considerations on artificial intelligence for health, WHO 2023;[147]
- Ethics guidelines for trustworthy AI, European Commission 2019;[148]
- Recommendation of the Council on Artificial Intelligence, OECD 2019, amended 2023;[149]
- Good machine learning practice for medical device development: Guiding Principles, U.S. Food & Drug Administration FDA, Health Canada, Medicines & Healthcare products Regulatory Agency MHRA 2021.[150]

References

[18] *Salas M, Petracek J, Yalamanchili P, Aimer O, Kasthuril D, Dhingra S, Junaid T, Bostic T. The Use of Artificial Intelligence in Pharmacovigilance: A Systematic Review of the Literature. Pharm Med 36, 2022;Oct;36(5):295–306. (Journal abstract) https://doi.org/10.1007/s40290-022-00441-z*

[19] Bate A, Stegmann JU. Artificial intelligence and pharmacovigilance: What is happening, what could happen and what should happen?. Health Policy and Technology. 2023;Jun1;12(2):100743. (Journal full text) https://doi.org/10.1016/j.hlpt.2023.100743

[20] Basile AO, Yahi A, Tatonetti NP. Artificial intelligence for drug toxicity and safety. Trends in Pharmacological Sciences. 2019;Sep1;40(9):624-635. (Journal full text) https://doi:10.1016/j.tips.2019.07.005

[21] Painter JL, Kassekert R, Bate A. Corrigendum: An industry perspective on the use of machine learning in drug and vaccine safety. Frontiers in Drug Safety and Regulation. 2023;Dec4;3:1244115. (Journal full text) https://doi.org/10.3389/fdsfr.2023.1110498

[22] Kassekert R, Grabowski N, Lorenz D, Schaffer C, Kempf D, Roy P, Kjoersvik O, Saldana G, ElShal S. Industry perspective on artificial intelligence/machine learning in pharmacovigilance. Drug Safety. 2022;May;45(5):439-448. (Journal full text) https://doi:10.1007/s40264-022-01164-5

[23] Hauben M, Reynolds R, Caubel P. Deconstructing the Pharmacovigilance Hype Cycle. Clin Ther. 2018;Dec;40(12):1981-1990.e3. (Journal full text) doi:10.1016/j.clinthera.2018.10.021.

[24] Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, Dasgupta N. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Safety. 2014;May;37:343-50. (Journal full text) https://doi.org/10.1007/s40264-014-0155-x

[25] Alvaro N, Conway M, Doan S, Lofi C, Overington J, Collier N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. Journal of Biomedical Informatics. 2015;Dec1;58:280-287. (Journal full text) https://doi.org/10.1016/j.jbi.2015.11.004

[26] 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. Journal of the American Medical Informatics Association. 2015;May1;22(3):671-681. (Journal full text) https://doi.org/10.1093/jamia/ocu041

[27] Karimi S, Metke-Jimenez A, Nguyen A. CADEminer: a system for mining consumer reports on adverse drug side effects. InProceedings of the eighth workshop on exploiting semantic annotations in information retrieval. 2015;Oct22;47-50. (Journal abstract) https://doi.org/10.1145/2810133.2810143

[28] White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the internet. Clinical Pharmacology & Therapeutics. 2014;Aug;96(2):239-246. (Journal full text) https://ascpt.onlinelibrary.wiley.com/doi/full/10.1038/clpt.2014.77

[29] Park J, Djelassi M, Chima D, Hernandez R, Poroshin V, Iliescu AM, Domalik D, Southall N. Validation of a natural language machine learning model for safety literature surveillance. Drug Safety. 2024;Jan;47(1):71-80. (Journal abstract) https://doi.org/10.1007/s40264-023-01367-4

[30] Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Safety. 2014;Apr;37:249-258. (Journal abstract) https://doi.org/10.1007/s40264-014-0146-y

[31] Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Mining and Knowledge Discovery. 2007;Jun;14:305-328. (Journal abstract) https://doi.org/10.1007/s10618-006-0052-8

[32] Kreimeyer K, Menschik D, Winiecki S, Paul W, Barash F, Woo EJ, Alimchandani M, Arya D, Zinderman C, Forshee R, Botsis T. Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems. Drug Safety. 2017;Jul;40:571-582. (Journal abstract)

[33] Personal communication from Tom Paternoster-Howe (March 2025).

[34] Kreimeyer K, Menschik D, Winiecki S, Paul W, Barash F, Woo EJ, Alimchandani M, Arya D, Zinderman C, Forshee R, Botsis T. Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems. Drug Safety. 2017;Jul;40:571-582. (Journal abstract) https://doi.org/10.1007/s40264-017-0523-4

[35] Kreimeyer K, Dang O, Spiker J, Gish P, Weintraub J, Wu E, Ball R, Botsis T. Increased confidence in deduplication of drug safety reports with natural language processing of narratives at the us food and drug administration. Frontiers in Drug Safety and Regulation. 2022;Jun15;2:918897. (Journal full text) https://doi.org/10.3389/fdsfr.2022.918897

[36] Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Safety. 2014;Apr;37:249-258. (Journal abstract) https://doi.org/10.1007/s40264-014-0146-y

[37] Combi C, Zorzi M, Pozzani G, Moretti U, Arzenton E. From narrative descriptions to MedDRA: automagically encoding adverse drug reactions. Journal of Biomedical Informatics. 2018;Aug1;84:184-199. (Journal full text) https://doi.org/10.1016/j.jbi.2018.07.001

[38] Gurulingappa H, Mateen-Rajpu A, Toldo L. Extraction of potential adverse drug events from medical case reports. Journal of Biomedical Semantics. 2012;Dec;3:1-10. (Journal full text) https://doi.org/10.1186/2041-1480-3-15

[39] Meldau EL, Bista S, Rofors E, Gattepaille LM. Automated Drug Coding Using Artificial Intelligence: An Evaluation of WHODrug Koda on Adverse Event Reports. Drug Safety. 2022;May;45(5):549-561. (Journal full text) https://doi.org/10.1007/s40264-022-01162-7

[40] Abatemarco D, Perera S, Bao SH, Desai S, Assuncao B, Tetarenko N. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. Pharmaceut Med. 2018;32 (6):391–401. (Journal full text) https://doi.org/10.1007/s40290-018-0251-9

[41] Wang W, Kreimeyer K, Woo EJ, Ball R, Foster M, Pandey A, Scott J, Botsis T. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. Journal of Biomedical Informatics. 2016;Aug1;62:78-89. (Journal full text) https://doi.org/10.1016/j.jbi.2016.06.006

[42] Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of Biomedical Informatics. 2017;Sep1;73:14-29. (Journal full text) https://doi.org/10.1016/j.jbi.2017.07.012

[43] Pham P, Cheng C, Wu E, Kim I, Zhang R, Ma Y, Kortepeter CM, Muñoz MA. Leveraging case narratives to enhance patient age ascertainment from adverse event reports. Pharmaceutical Medicine. 2021;Sep;35(5):307-316. https://doi.org/10.1007/s40290-021-00398-5

[44] Dang V, Wu E, Kortepeter CM, Phan M, Zhang R, Ma Y, Muñoz MA. Evaluation of a natural language processing tool for extracting gender, weight, ethnicity, and race in the US food and drug administration adverse event reporting system. Frontiers in Drug Safety and Regulation. 2022;Nov14;2:1020943. (Journal full text) https://doi.org/10.3389/fdsfr.2022.1020943

[45] Abatemarco, D., Perera, S., Bao, S.H. et al. Training Augmented Intelligent Capabilities for Pharmacovigilance: Applying Deep-learning Approaches to Individual Case Safety Report Processing. Pharm Med 32, 391–401 (2018). https://doi.org/10.1007/s40290-018-0251-9

[46] Römming H-J, Pushparajan R. AI Translation Assistant for Pharmacovigilance. Poster presented at DIA Europe 2021. (accessed 21 March 2025)

[47] Young IJ, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. International Journal of Medical Informatics. 2019;Dec1;132:103971. (Journal abstract) https://doi.org/10.1016/j.ijmedinf.2019.103971

[48] OMOP Common Data Model for longitudinal observational health data https://www.ohdsi.org/data-standardization/

[49] Hauben M, Aronson JK, Ferner RE. Evidence of misclassification of drug–event associations classified as gold standard 'negative controls' by the observational medical outcomes partnership (OMOP). Drug Safety. 2016;May;39:421-432. (Journal abstract) https://doi.org/10.1007/s40264-016-0392-2

[50] Wang X, Lehmann H, Botsis T. Can FHIR support standardization in post-market safety surveillance?. InPublic Health and Informatics. 2021;(pp. 33-37). IOS Press. (Journal full text) https://doi.org/10.3233/SHTI210115

[51] Ghosh, R., Kempf, D., Pufko, A. et al. Automation Opportunities in Pharmacovigilance: An Industry Survey. Pharm Med 34, 7–18 (2020). https://doi.org/10.1007/s40290-019-00320-0

[52] van der Aalst, W.M.P., Bichler, M. & Heinzl, A. Robotic Process Automation. Bus Inf Syst Eng 60, 269–272 (2018). https://doi.org/10.1007/s12599-018-0542-4

[53] Gosselt HR, Bazelmans EA, Lieber T, van Hunsel FP, Härmark L. Development of a multivariate prediction model to identify individual case safety reports which require clinical review. Pharmacoepidemiology and Drug Safety. 2022;Dec;31(12):1300-1307. (Journal full text) https://doi.org/10.1002/pds.5553

[54] Bergman E, Dürlich L, Arthurson V, Sundström A, Larsson M, Bhuiyan S, Jakobsson A, Westman G. BERT based natural language processing for triage of adverse drug reaction reports shows close to human-level performance. PLOS Digital Health. 2023;Dec 6;2(12):e0000409. (Journal full text) https://doi.org/10.1371/journal.pdig.0000409

[55] Cherkas Y, Ide J, van Stekelenborg J. Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions. Drug Safety. 2022;May;45(5):571-582. (Journal abstract) https://doi.org/10.1007/s40264-022-01163-6

[56] Meldau EL, Bista S, Melgarejo-González C, Norén GN. Automated redaction of names in adverse event reports using transformer-based neural networks. BMC Medical Informatics and Decision Making. 2024;Dec23;24(1):401. (Journal full text) https://doi.org/10.1186/s12911-024-02785-9

[57] Finney DJ. stematic Signalling of Adverse Reactions to Drugs. Methods of Information in Medicine. 1974;13(01):01-10. (Journal full text) https://doi.org/10.1055/s-0038-1636131

[58] Practical Aspects of Signal Detection in Pharmacovigilance. CIOMS Working Group VIII report. Council for International Organizations of Medical Sciences (CIOMS), 2010.

[59] Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. European Journal of Clinical Pharmacology.1998;Jul;54:315-21. (Journal abstract) https://doi.org/10.1007/s002280050466

[60] DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. The American Statistician. 1999;1;53(3):177-190. (Journal full text) https://doi.org/10.1080/00031305.1999.10474456

[61] Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety. 2001;Oct;10(6):483-486. (Journal full text) https://doi.org/10.1002/pds.677

[62] Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2010;Aug;3(4):197-208. (Journal abstract) https://doi.org/10.1002/sam.10078

[63] Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clinical Pharmacology & Therapeutics. 2011;Feb;89(2):243-250.

[64] Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Science Translational Medicine. 2012;Mar14;4(125):125ra31-. (Journal abstract) https://doi: 10.1126/scitranslmed.3003377

[65] Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug–drug interactions in a spontaneous reports database. British Journal of Clinical Pharmacology. 2007;Oct;64(4):489-495. (Journal full text) https://doi.org/10.1111/j.1365-2125.2007.02900.x

[66] Norén GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug–drug interaction surveillance. Statistics in Medicine. 2008;Jul 20;27(16):3057-3070. (Journal full text) https://doi.org/10.1002/sim.3247

[67] Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Science Translational Medicine. 2012;Mar14;4(125):125ra31-. (Journal abstract) https://doi:10.1126/scitranslmed.3003377

[68] Hauben M, Hung E, Chen Y. Potential signals of COVID-19 as an effect modifier of adverse drug reactions. Clinical Therapeutics. 2024;Jan1;46(1):20-29. (Journal full text) https://doi.org/10.1016/j.clinthera.2023.10.002

[69] Hauben M. Artificial intelligence and data mining for the pharmacovigilance of drug–drug interactions. Clinical Therapeutics. 2023;Feb1;45(2):117-133. (Journal full text) https://doi.org/10.1016/j.clinthera.2023.01.002

[70] van Puijenbroek, E., Egberts, A., Heerdink, E. et al. Detecting drug–drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. Eur J Clin Pharmacol 56, 733–738 (2000). https://doi.org/10.1007/s002280000215

[71] Sandberg L, Taavola H, Aoki Y, Chandler R, Norén GN. Risk factor considerations in statistical signal detection: using subgroup disproportionality to uncover risk groups for adverse drug reactions in VigiBase. Drug Safety. 2020;Oct;43(10):999-1009. (Journal full text) https://doi.org/10.1007/s40264-020-00957-w

[72] Hauben M, Hung E, Chen Y. Potential signals of COVID-19 as an effect modifier of adverse drug reactions. Clinical Therapeutics. 2024;Jan1;46(1):20-29. (Journal full text) https://doi.org/10.1016/j.clinthera.2023.10.002

[73] Juhlin, K., Karimi, G., Andér, M. et al. Using VigiBase to Identify Substandard Medicines: Detection Capacity and Key Prerequisites. Drug Saf 38, 373–382 (2015). https://doi.org/10.1007/s40264-015-0271-2

[74] Trippe, Z.A., Brendani, B., Meier, C. et al. Identification of Substandard Medicines via Disproportionality Analysis of Individual Case Safety Reports. Drug Saf 40, 293–303 (2017). https://doi.org/10.1007/s40264-016-0499-5

[75] Olivia Mahaux, Vincent Bauchau, Ziad Zeinoun, Lionel Van Holle, Tree-based scan statistic – Application in manufacturing-related safety signal detection, Vaccine, Volume 37, Issue 1, 2019, Pages 49-55, ISSN 0264-410X, https://doi.org/10.1016/j.vaccine.2018.11.044.

[76] Caster O, Juhlin K, Watson S, Norén GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank: retrospective evaluation against emerging safety signals. Drug Safety. 2014;Aug37:617-628. (Journal full text) https://doi.org/10.1007/s40264-014-0204-5

[77] Van Holle L, Bauchau V. Use of logistic regression to combine two causality criteria for signal detection in vaccine spontaneous report data. Drug Safety. 2014;Dec;37:1047-1057.(Journal full text) https://doi.org/10.1007/s40264-014-0237-9

[78] Scholl JH, van Hunsel FP, Hak E, van Puijenbroek EP. A prediction model-based algorithm for computer-assisted database screening of adverse drug reactions in the Netherlands. Pharmacoepidemiology and drug safety. 2018;Feb;27(2):199-205. (Journal full text) https://doi.org/10.1002/pds.4364

[79] Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, Milward D, Winter A, Lu S, Ball R. Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology. Journal of Biomedical Informatics. 2018;Jul1;83:73-86. (Journal full text) https://doi.org/10.1016/j.jbi.2018.05.019

[80] Avillach P, Dufour JC, Diallo G, Salvo F, Joubert M, Thiessard F, Mougin F, Trifirò G, Fourrier-Réglat A, Pariente A, Fieschi M. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU–ADR project. Journal of the American Medical Informatics Association. 2013;May1;20(3):446-452. (Journal abstract) https://doi.org/10.1136/amiajnl-2012-001083

[81] Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. Journal of the American Medical Informatics Association. 2011;Sep1;18(5):668-674. (Journal abstract) https://doi.org/10.1136/amiajnl-2011-000096

[82] Silverman AL, Sushil M, Bhasuran B, Ludwig D, Buchanan J, Racz R, Parakala M, El-Kamary S, Ahima O, Belov A, Choi L. Algorithmic Identification of Treatment-Emergent Adverse Events From Clinical Notes Using Large Language Models: A Pilot Study in Inflammatory Bowel Disease. Clinical Pharmacology & Therapeutics. 2024;Jun;115(6):1391-1399. (Journal full text) https://doi.org/10.1002/cpt.3226

[83] Wu J, Ruan X, McNeer E, Rossow KM, Choi L. Developing a natural language processing system using transformer-based models for adverse drug event detection in electronic health records. medRxiv. 2024;Jul10:2024-07. (Journal full text) https://doi.org/10.1101/2024.07.09.24310100

[84] Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, Yue P, Tsau PS, Kohane I, Roden DM, Altman RB. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clinical Pharmacology & Therapeutics. 2011 Jul;90(1):133-142. Erratum in: Clin Pharmacol Ther. 2011;Sep;90(3):480. Tsau, P S (corrected to Tsao, P S). (Journal full text)https://doi.org/10.1038/clpt.2011.83

[85] Hauben M, Reynolds R, Caubel P. Deconstructing the pharmacovigilance hype cycle. Clinical Therapeutics. 2018;Dec1;40(12):1981-1990. (Journal full text) https://doi.org/10.1016/j.clinthera.2018.10.021

[86] Kreimeyer K, Dang O, Spiker J, Muñoz MA, Rosner G, Ball R, Botsis T. Feature engineering and machine learning for causality assessment in pharmacovigilance: Lessons learned from application to the FDA Adverse Event Reporting System. Computers in Biology and Medicine. 2021;Aug1;135:104517. (Journal abstract) https://doi.org/10.1016/j.compbiomed.2021.104517

[87] Muñoz MA, Dal Pan GJ, Wei YJ, Delcher C, Xiao H, Kortepeter CM, Winterstein AG. Towards automating adverse event review: a prediction model for case report utility. Drug Safety. 2020;Apr;43:329-338. (Journal abstract) https://doi.org/10.1007/s40264-019-00897-0

[88] Singh LL, Sudarsan SD, Jetley RP, Fitzgerald B, Milanova M. Semantic search tool for adverse event reports of medical devices. InSWWS 2011:proceedings of the 2011 international conference on semantic web & web services. 2011;July18-21:129-135. (Journal full text) https://doi.org/10.13140/RG.2.2.22712.29443

[89] Zekarias A, Meldau EL, Bista S, Félix China J, Sandberg L. Narrative Search Engine for Case Series Assessment Supported by Artificial Intelligence Query Suggestions. Drug Safety. 2025;Mar15:1-3. (Journal full text) https://doi.org/10.1007/s40264-025-01529-6

[90] Imran M, Bhatti A, King DM, Lerch M, Dietrich J, Doron G, Manlik K. Supervised machine learning-based decision support for signal validation classification. Drug Safety. 2022;May;45(5):583-596. (Journal full text) https://doi.org/10.1007/s40264-022-01159-2

[91] Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, Gish P, Lardieri A, Wu E, De S, Naidoo J. Information visualization platform for postmarket surveillance decision support. Drug Safety. 2020;Sep;43:905-915. (Journal abstract) https://doi.org/10.1007/s40264-020-00945-0

[92] Pinheiro LC, Durand J, Dogné JM. An Application of Machine Learning in Pharmacovigilance: Estimating Likely Patient Genotype from Phenotypical Manifestations of Fluoropyrimidine Toxicity. Clin Pharmacol Ther.2020;107(4):944-947. (Journal full text) https://doi.org/10.1002/cpt.1789

[93] Orre R, Bate A, Norén GN, Swahn E, Arnborg S, Edwards IR. A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets. International Journal of Neural Systems. 2005;Jun;15(03):207-222. (Journal abstract) https://doi.org/10.1142/S0129065705000219

[94] Ball R, Botsis T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS?. Clinical Pharmacology & Therapeutics. 2011;Aug;90(2):271-278. (Journal abstract) https://doi.org/10.1038/clpt.2011.119

[95] Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clinical Pharmacology & Therapeutics. 2011;Feb;89(2):243-250. (Journal abstract) https://doi.org/10.1038/clpt.2010.285

[96] 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. Ine-Health–For Continuity of Care. 2014;1178-1182. (Journal full text) https://doi.org/10.3233/978-1-61499-432-9-1178

[97] Fusaroli M, Polizzi S, Menestrina L, Giunchi V, Pellegrini L, Raschi E, Weintraub D, Recanatini M, Castellani G, De Ponti F, Poluzzi E. Unveiling the burden of drug-induced impulsivity: a network analysis of the FDA adverse event reporting system. Drug Safety. 2024;Aug15:1-8. (Journal full text) https://doi.org/10.1007/s40264-024-01471-z

[98] Norén GN, Meldau EL, Chandler RE. Consensus clustering for case series identification and adverse event profiles in pharmacovigilance. Artificial Intelligence in Medicine. 2021;Dec1;122:102199. (Journal full text)HYPERLINK "https://doi.org/10.1016/j.artmed.2021.102199"https://doi.org/10.1016/j.artmed.2021.102199

[99] Simms AM, Kanakia A, Sipra M, Dutta B, Southall N. A patient safety knowledge graph supporting vaccine product development. BMC medical informatics and decision making. 2024;Jan4;24(1):10. (Journal full text) https://doi.org/10.1186/s12911-023-02409-8

[100] Bean DM, Wu H, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, Stewart R, Dobson RJ. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Scientific Reports.2017;Nov27;7(1):16416. (Journal full text) https://doi.org/10.1038/s41598-017-16674-x

[101] Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin NE, Strawbridge SA, Garcia-Patino M. DrugBank 6.0: the DrugBank knowledgebase for 2024. Nucleic Acids Research. 2024;Jan5;52(D1):D1265-1275. (Journal full text) https://doi.org/10.1093/nar/gkad976

[102] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Research. 2016;Jan4;44(D1):D1075-2079. (Journal full text) https://doi.org/10.1093/nar/gkv1075

[103] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics. 2008;Oct1;41(5):706-716. (Journal full text) https://doi.org/10.1016/j.jbi.2008.03.004

[104] Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinformatics. 2015;Dec;16:1-1. (Journal full text) https://doi.org/10.1186/s12859-015-0774-y

[105] Low YS, Caster O, Bergvall T, Fourches D, Zang X, Norén GN, Rusyn I, Edwards R, Tropsha A. Cheminformatics-aided pharmacovigilance: application to Stevens-Johnson Syndrome. Journal of the American Medical Informatics Association. 2016;Sep1;23(5):968-78. (Journal full text) https://doi.org/10.1093/jamia/ocv127

[106] Hauben M, Rafi M, Abdelaziz I, Hassanzadeh O. Knowledge graphs in pharmacovigilance: a scoping review. Clinical therapeutics. 2024;Jul9. (Journal full text) https://doi.org/10.1016/j.clinthera.2024.06.003

[107] Sridharan K, Sivaramakrishnan G. Enhancing readability of USFDA patient communications through large language models: a proof-of-concept study. Expert Review of Clinical Pharmacology. 2024;Aug2;17(8):731-741. (Journal abstract) https://doi.org/10.1080/17512433.2024.23638407

[108] Ying L, Liu Z, Fang H, Kusko R, Wu L, Harris S, Tong W. Text summarization with ChatGPT for drug labeling documents. Drug Discovery Today. 2024;May7:104018. (Journal full text) https://doi.org/10.1016/j.drudis.2024.104018

[109] Dietrich J, Hollstein A. Performance and Reproducibility of Large Language Models in Named Entity Recognition: Considerations for the Use in Controlled Environments. Drug Safety. 2025;Mar;48(3):287-303.(Journal full text) https://doi.org/10.1007/s40264-024-01499-1

[110] Painter JL, Mahaux O, Vanini M, Kara V, Roshan C, Karwowski M, Chalamalasetti VR, Bate A. Enhancing drug safety documentation search capabilities with large language models: a user-centric approach. International Conference on Computational Science and Computational Intelligence (CSCI). 2023;Dec13:49-56. (Journal abstract) http://doi.org/10.1109/CSCI62032.2023.00015

[111] Wu L, Xu J, Thakkar S, Gray M, Qu Y, Li D, Tong W. A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document. Regulatory Toxicology and Pharmacology. 2024;May1;149:105613. (Journal full text) https://doi.org/10.1016/j.yrtph.2024.105613

[112] Painter JL, Chalamalasetti VR, Kassekert R, Bate A. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. JAMIA open. 2025;Feb;8(1):ooaf003. (Journal full text) https://doi.org/10.1093/jamiaopen/ooaf003

[113] Benaïche A, Billaut-Laden I, Randriamihaja H, Bertocchio JP. Assessment of the Efficiency of a ChatGPT-Based Tool, MyGenAssist, in an Industry Pharmacovigilance Department for Case Documentation: Cross-Over Study. J Med Internet Res 2025;27:e65651. doi: 10.2196/65651

[114] Meldau EL, Bista S, Rofors E, Gattepaille LM. Automated Drug Coding Using Artificial Intelligence: An Evaluation of WHODrug Koda on Adverse Event Reports. Drug Safety. 2022;May;45(5):549-561. (Journal full text) https://doi.org/10.1007/s40264-022-01162-7

[115] Kreimeyer K, Dang O, Spiker J, Gish P, Weintraub J, Wu E, Ball R, Botsis T. Increased confidence in deduplication of drug safety reports with natural language processing of narratives at the us food and drug administration. Frontiers in Drug Safety and Regulation. 2022;Jun15;2:918897. (Journal full text) https://doi.org/10.3389/fdsfr.2022.918897

[116] Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Mining and Knowledge Discovery. 2007;Jun;14:305-28. (Journal abstract) https://doi.org/10.1007/s10618-006-0052-8

[117] Bergman E, Dürlich L, Arthurson V, Sundström A, Larsson M, Bhuiyan S, Jakobsson A, Westman G. BERT based natural language processing for triage of adverse drug reaction reports shows close to human-level performance. PLOS Digital Health. 2023;Dec6;2(12):e0000409. (Journal full text) https://doi.org/10.1371/journal.pdig.0000409

[118] Painter J, Haguinet F, Cranfield C,Bate A. MSR20 NLP and Machine Learning to Automate Identification of Suspected Medication Errors from Real World Unstructured Narratives. Value in Health, Volume 26, Issue 6, S281. https://doi.org/10.1016/j.jval.2023.03.1556

[119] *Caster O, Juhlin K, Watson S, Norén GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank: retrospective evaluation against emerging safety signals. Drug Safety. 2014;Aug;37:617-628. (Journal full text) https://doi.org/10.1007/s40264-014-0204-5*

[120] *Caster O, Sandberg L, Bergvall T, Watson S, Norén GN. vigiRank for statistical signal detection in pharmacovigilance: first results from prospective real-world use. Pharmacoepidemiology and Drug Safety. 2017;Aug;26(8):1006-1010. (Journal full text) https://doi.org/10.1002/pds.4247*

[121] *Scholl JH, van Hunsel FP, Hak E, van Puijenbroek EP. A prediction model-based algorithm for computer-assisted database screening of adverse drug reactions in the Netherlands. Pharmacoepidemiology and Drug Safety. 2018;Feb;27(2):199-205. (Journal full text) https://doi.org/10.1002/pds.4364*

[122] *Norén GN, Meldau EL, Chandler RE. Consensus clustering for case series identification and adverse event profiles in pharmacovigilance. Artificial Intelligence in Medicine. 2021;Dec1;122:102199. (Journal full text) https://doi.org/10.1016/j.artmed.2021.102199*

[123] *Rudolph A, Mitchell J, Barrett J, Sköld H, Taavola H, Erlanson N, Melgarejo-González C, Yue QY. Global safety monitoring of COVID-19 vaccines: How pharmacovigilance rose to the challenge. Therapeutic Advances in Drug Safety. 2022;Aug;13:20420986221118972. (Journal full text) https://doi.org/10.1177/20420986221118972*

[124] *Habets PC, van IJzendoorn DG, Vinkers CH, Härmark L, de Vries LC, Otte WM. Development and validation of a machine-learning algorithm to predict the relevance of scientific articles within the field of teratology. Reproductive Toxicology. 2022;Oct1;113:150-154. (Journal full text) https://doi.org/10.1016/j.reprotox.2022.09.001*

[125] *OECD (2023), "The state of implementation of the OECD AI Principles four years on", OECD Artificial Intelligence Papers, No. 3, OECD Publishing, Paris, https://doi.org/10.1787/835641c9-en.*

[126] *OECD (2023), "The state of implementation of the OECD AI Principles four years on", OECD Artificial Intelligence Papers, No. 3, OECD Publishing, Paris, https://doi.org/10.1787/835641c9-en.*

[127] *European Medicines Agency. Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.2024 (Full text accessed 21 March 2025).*

[128] *European Commission: Directorate-General for Communications Networks, Content and Technology, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment, Publications Office, 2020. (Full text accessed 21 March 2025) https://data.europa.eu/doi/10.2759/002360*

[129] *Department of Industry, Science, Energy and Resources Australia. Australia's Artificial Intelligence Ethics Principles. (Webpage accessed 21 March 2025).*

[130] *Government of Canada. Artificial Intelligence and Data Act (AIDA) Companion Document. (Full text accessed 21 March 2025).*

[131] *Ministry of Health Singapore. 1.0 Artificial Intelligence in Healthcare Guidelines (AIHGLE). 2021. (Full text accessed 21 March 2025).*

[132] *Government of the United Kingdom. Department of Science, Innovation & Technology. Implementing the UK AI Regulatory Principles: Guidance for Regulators. 2024. (Full text accessed 21 March 2025).*

[133] *White House Office of Science and Technology Policy (OSTP). Blueprint for an AI Bill of Rights. (Webpage accessed 21 March 2025).*

[134] *Pan American Health Organization. Artificial Intelligence of Public Health. Digital Transformation Toolkit, Knowledge Tools. (Full text accessed 21 March 2025).*

[135] *World Health Organization. Ethics and governance of artificial intelligence for health. 2021. (Full text accessed 21 March 2025).*

[136] *Organisation for Economic Co-operation and Development. OECD AI Principles Overview. 2019, updated in 2024. (Full text accessed 21 March 2025).*

[137] *European Medicines Agency. Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle. 2024. (Full text accessed 21 March 2025).*

[138] *European Medicines Agency. Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle. 2024. (Full text accessed 21 March 2025).*

[139] *Heads of Medicines Agencies - European Medicines Agency Joint Big Data Steering Group. Multi-annual artificial Intelligence Workplan 2023-2028. 2023. (Full text accessed 21 March 2025).*

[140] *U.S. Food and Drug Administration. Using Artificial Intelligence & Machine Learning in the Development of Drug and Biological Products. 2023, revised 2025. (Full text accessed 21 March 2025)*

[141] *U.S. Food and Drug Administration. Center for Drug Evaluation and Research. Artificial Intelligence in Drug Manufacturing. 2023. (Full text accessed 21 March 2025)*

[142] *U.S. Food and Drug Administration. CDER Emerging Drug Safety Technology Program (EDSTP). 2024. (Webpage accessed 21 March 2025)*

[143] *European Medicines Agency. Harnessing AI in medicines regulation: use of large language models (LLMs).2024. (Full text accessed 21 March 2025)*

[144] *Government of Canada. Guide to the use of generative AI. 2023. (Full text accessed 21 March 2025)*

[145] *Organisation for Economic Co-operation and Development. Initial policy considerations for generative artificial intelligence. 2023. (Full text accessed 21 March 2025)*

[146] *World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. (Full text accessed 21 March 2025)*

[147] *World Health Organization (WHO). Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023. (Full text accessed 21 March 2025)*

[148] *European Commission. Ethics guidelines for trustworthy AI. 2019. (accessed 21 March 2025).*

[149] *OECD Legal Instruments, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 2019, amended 2024. (Full text accessed 21 March 2025)*

[150] *U.S. Food and Drug Administration, Health Canada, Medicines & Healthcare products Regulatory Agency. Good machine learning practice for medical device development: Guiding Principles. 2021. (Full text accessed 21 March 2025)*

Principles for integrating and implementing artificial intelligence within pharmacovigilance

This is a summary of the broad framework of principles for integrating and implementing AI within PV that will be addressed in the following chapters in greater detail.

Risk-based approach

A risk-based approach acknowledges the potential hazards that AI systems can pose and recognises that different use cases present varying types and levels of risk. This necessitates a risk assessment that identifies, prioritises, and manages risks that could negatively affect a PV system's behaviour and results, taking into consideration existing process controls. A risk is characterised by both the anticipated impact and the likelihood of negative outcomes.

This approach also supports procedures to identify and reduce errors and biases in a way that is proportionate to their risk. It influences the implementation strategies of AI systems (including documentation, compliance, and record-keeping), which should generally be commensurate with the identified risk.

Human oversight

Human oversight refers to the expected role of humans in the design, implementation, monitoring, and analysis of AI systems in PV. It requires a framework to manage performance and to detect and mitigate potential issues related to the AI system.

Validity

Validity means that a system achieves its intended purpose within acceptable parameters. It requires predefining acceptable performance levels, selecting appropriate data for model training and/or testing, assessing model performance in a realistic setting, and integrating the system into an ongoing quality assessment process.

Robustness

Robustness means that a system reliably achieves its intended objectives (while accounting for variations in data).

Transparency

Transparency regarding AI involves disclosing information between organizations or individuals. This includes sharing relevant documentation of the AI system lifecycle (i.e. design, development, evaluation, deployment, operation, re-training, maintenance and decommission) to facilitate traceability and providing stakeholders with enough information to have a general understanding of the AI system, its use, risks, limitations, and impact on their rights.

Data privacy

Data privacy refers to the fundamental right of an individual to control how their personal information is collected, stored, shared, and used. It is an aspect of the principle of “respect for persons” that is foundational to the conduct of biomedical research. Regulations, legislation and guidance documents provide measures intended to preserve the confidentiality, anonymity, autonomy and control of sensitive and potentially personally identifiable health data in the setting of PV.

Fairness & Equity

Fairness and equity require awareness of and adherence to impartiality, equality, non-discrimination, diversity, justice, and lawfulness. The benefits of AI in PV should be equitable across all relevant

699 populations and groups. Throughout the AI lifecycle, it is important to avoid and mitigate unfair bias,
700 and any discriminatory practices and unjust social wellbeing and environmental impacts.

701 *Governance*

702 Governance refers to the human management and oversight used to control and direct the use of AI
703 in the PV system. An AI governance framework requires implementation of risk management
704 practices and policies to ensure adherence to the AI guiding principles.

705 *Accountability*

706 Accountability applies to clearly defined roles, responsibilities and liability for organisations and/or
707 individuals deploying, operating and managing AI systems. It requires the adoption of appropriate
708 governance measures by relevant stakeholders, including but not limited to regulators, vendors,
709 users, developers, data providers or pharmaceutical companies involved in setting policy, developing,
710 deploying and managing AI systems. This ensures operations remain within expected parameters
711 throughout the AI lifecycle while addressing any unforeseen consequences.

712

Chapter 3: Risk-based approach

Principle

A risk-based approach acknowledges the potential hazards that AI systems can pose and recognises that different use cases present varying types and levels of risk. This necessitates a risk assessment that identifies, prioritises, and manages risks that could negatively affect a PV system's behaviour and results, taking into consideration existing process controls. A risk is characterised by both the anticipated impact and the likelihood of negative outcomes.[151]

This approach also supports procedures to identify and reduce errors and biases in a way that is proportionate to their risk. It influences the implementation strategies of AI systems (including documentation, compliance, and record-keeping), which should generally be commensurate with the identified risk.

Key messages

- Integrating AI into PV processes needs to take into account that the performance of both AI algorithms, and humans, is imperfect.
- The risks potentially associated with the use of AI in PV may affect patient safety, the trust and engagement of PV users, the efficiency of PV processes as well as compliance with regulatory standards and ethical principles.
- By focusing efforts and resources where they most matter, a sound risk-based approach enables organisations to make the most of AI capabilities while ensuring that neither patient safety nor PV stakeholders are adversely affected.
- The risk-based approach applies to the human oversight modalities, the validity and robustness strategy, the level of transparency, and the efforts to uphold fairness and equity, and data privacy.
- The risk assessment should consider the AI system itself, the context of use, and the potential impact and likelihood of risks materialising.
- Comparative performance to current best practice should also be considered.
- Assessment should be end-to-end with an emphasis on end objective.
- A risk-based approach should be reviewed at regular intervals and adapted if needed.

Introduction

Regulatory considerations

Regardless of the integration of AI elements, PV systems are expected to comply with existing regulations and good pharmacovigilance practices (GVP).[152,153] In accordance with GVP, a wide range of PV processes are considered critical for business continuity purposes, including collection and handling of ICSRs, signal management, and PSURs.[154]

Regulatory frameworks generally recommend a risk-based approach in the development, deployment, monitoring, documentation and regulatory oversight of AI systems, to ensure that relevant risks are anticipated, identified and mitigated throughout the system lifecycle.[155,156,157] The European Artificial Intelligence Act (EU AI Act)[158] introduces four risk categories for AI systems: low or minimal risk, limited risk (transparency obligations), high risk, and unacceptable risk (prohibited AI practices). High-risk AI systems, which include e.g. AI-based medical software/devices or AI systems used for staff recruitment, are associated with strict requirements and obligations on providers and deployers, including risk-mitigation systems, high quality data sets for training, validation and testing, logging of activity, detailed documentation, clear user information, human oversight, and a high level of robustness, accuracy, and cybersecurity. While the guiding principles advocated throughout this

report overlap with the EU AI Act's requirements for high-risk AI systems, determining the applicable EU AI Act's risk category of an AI system considered for integration into an organisation's PV process will likely require a careful case-by-case assessment, with legal advice as appropriate. Within the medicines' lifecycle, EMA foresees AI systems with 'high patient risk' in use cases where patient safety is affected and AI systems with 'high regulatory impact' in use cases where impact on the regulatory decision making is substantial.[159] The Artificial Intelligence and Data Act (AIDA) was developed to ensure the development of responsible AI in Canada, with a risk-based approach aligned with international norms, including the EU AI Act, the OECD AI Principles, and the US National Institute of Standards and Technology (NIST) Risk Management Framework (RMF).[160]

During development and other stages of an AI solution's lifecycle as relevant, and depending on the level of risk to individual patients, public health or the regulatory decision making, applicants and developers should consider engaging actively with regulatory authorities and seek suitable scientific advice. Where necessary, technical qualification of the AI technology through appropriate channels should be sought based on legislative or regulatory requirements applicable to medicinal products, medical devices and/or software development.[161,162] Due to its fast-moving nature, the use of AI technology in the medicines' lifecycle including in PV will pose challenges to both regulators, required to adapt and keep abreast of this evolving field,[163] and industry PV stakeholders, required to maintain regulatory compliance (see Chapter on Future Vision).

Motivation and interplay with other guiding principles

A sound, risk-based approach will allow organisations to focus their efforts and resources where they matter most to maximise their AI capabilities while ensuring that guiding principles are upheld, as described earlier.

Rather than a self-standing principle, a risk-based approach is applicable to the other guiding principles presented in this report. Notably, a risk-based approach will inform where, when, how and how much human oversight should be implemented within PV processes involving AI in addition to other risk mitigation activities and conversely, an AI solution may be 'risk-assessed' assessed taking into account the degree and nature of existing human oversight (see Chapter on Human oversight). A risk-based approach should be applied to the testing and verification of AI systems (see Chapter on Validity & Robustness) and the level of documentation and record-keeping (see Chapter on Transparency). A risk-based approach is also relevant to data privacy and fairness and equity. For example, AI systems should be assessed for any risks that specific groups may be under-served or biased against, and those risks should be appropriately mitigated (see Chapter on Fairness & Equity).

Types of risks

This section briefly outlines some of the risks potentially associated with the use of AI solutions in PV.

Risks to patient safety and public health

Inadequate use of AI solutions in PV, or their poor performance, may impede the fulfilment of PV objectives: detection, assessment, understanding and prevention of adverse effects of drugs or vaccines, which may come at the cost of patient safety or public health. Unreliable outputs produced by an AI system, including but not limited to false negatives or false positives, or unfair bias, could negatively impact PV activities with e.g. relevant adverse events not captured, events misclassified during case processing, or signals missed. This could result in safety issues not being identified or being identified with delay, potentially putting patients at risk. In rare scenarios, the late detection of new, unexpected safety signals could have a major public health impact ('Black swan' events).[164] An initially robust AI tool could also start underperforming over time due to e.g. model drift, or become

inoperative due to an IT incident or system failure, which would impede the PV activity that the AI tool is intended to support.

Risks to user trust and engagement

The lack of transparency and interpretability of certain AI algorithms may hinder trust and acceptance by users, including PV professionals (see Chapter on Transparency). Lack of trust from users may also result from poor previous experience with AI systems of insufficient validity and robustness, leading to mistrust of AI solutions in general. In clinic-based PV settings, a more subtle potential source of mistrust is 'uniqueness neglect', in which patients prefer a human clinician over a more accurate computer due to a belief that machines do not fully accommodate their personal human uniqueness.[165] Other possible sources of mistrust include poor performance for certain subpopulations or failure to protect confidentiality of personal data during the development or operation of an AI system. Conversely, some users may put excessive trust in AI systems, leading to automation bias (especially if those have shown robust performance upon validation) and the resulting unconscious bias to accept erroneous outputs. Additionally, integrating AI solutions into existing workflows and systems may pose technical, organisational, and cultural challenges, with a risk of degraded job motivation or satisfaction in the absence of adequate training and change management strategies (see Chapter on Human oversight).

Risks to efficiency

Although the integration of AI in PV processes is generally aimed at increasing efficiency, substandard AI solutions may cause more manual work than they save, if for instance, significant time is required to understand and verify the AI outputs or bring them up to acceptable standards. Uncertainties, such as false positives, in interpretations and actions based on AI outputs might add to inefficiencies or suboptimal use of limited resources.

Legal, ethical and other risks

Other risks may be related to data privacy, cybersecurity threats, intellectual property, liability, or economic and reputational aspects.

This chapter mainly focusses on the impact that the use AI tools in PV processes could have on patient safety. Potential concerns related to user acceptance and other challenges are further discussed in the chapters on Data Privacy, Fairness & Equity, Transparency, Human oversight and Governance & Accountability.

Risk assessment

General considerations

Organisations planning to deploy AI-based tools to support PV processes are expected to perform a thorough risk analysis. This assessment should be performed for each AI system and should form the basis for a risk-proportionate approach applied throughout the AI system's lifecycle from development to routine use.

When determining the level of risk related to the implementation of an AI tool within a PV system, key considerations include the AI technology itself, the context of use, the likelihood of risks materialising and their potential impact.

Artificial intelligence technology

The level of risk may depend on the type of tool used (e.g. static vs dynamic model), the underlying data quality, the novelty of the technology or the maturity of the system (i.e. lifecycle stage).

Particular caution should be exercised with the integration of GenAI models within PV processes. Compared to simpler or more explainable AI approaches, the non-deterministic nature of GenAI and

similar AI models, the opacity of training data and the potential for hallucinations, may make the detection and mitigation of issues more challenging and require consideration of further guardrails.

As the AI landscape continues to evolve, so will AI-related risk areas. New risks may emerge while others may become less prominent; for instance, the current challenges associated with GenAI/LLMs may be solved, making their integration into sensitive PV processes safer.

Context of use and degree of influence

These broadly refer to the place and importance of the AI solution within the overall PV system, including:

- Whether or not the AI solution is used in a critical PV process or high-risk context (e.g. emergency Public Health use, novel substance, clinical trial cases);
- At which stage within a particular process the AI tool intervenes (e.g. automated triage of relevant cases as a preliminary step to signal review) and whether the tool is assistive or directly supports a PV process;
- The extent of human involvement and oversight in the process (see Chapter on Human oversight).

Impact and likelihood

Not all occurrences of system malfunction or suboptimal model performance are as likely, nor will they have the same impact. For instance, a duplicate detection tool applied to a very large database is not expected to detect 100% of duplicates but missed duplicates will have no or limited consequences in terms of patient safety, whereas the late detection of a very serious signal in a context of mass patient exposure happens very rarely but may have dramatic public health consequences (i.e. black swan event).

Examples of structured approaches

Risk-based assessment frameworks have been proposed in various domains and may provide inspiration to organisations wishing to deploy AI solutions within PV systems. Selected examples are briefly described hereafter.

Credibility assessment framework

The FDA proposes a stepwise approach to demonstrate the credibility of AI models to produce information or data intended to support regulatory decision making regarding the safety, effectiveness, or quality of drugs (see also Chapter Landscape analysis).[166] Similar frameworks have been proposed for the use of computational models in medical device submissions[167] or drug development.[168] The preliminary steps of the credibility assessment, as outlined below, help assess the model risk.

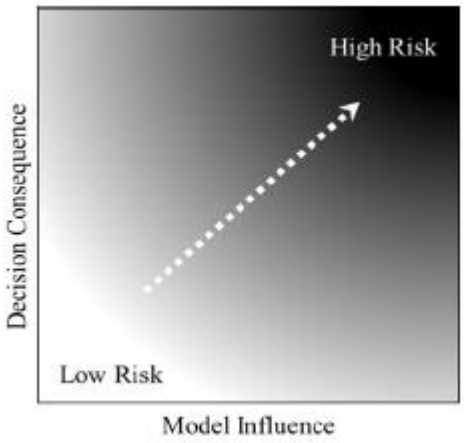
1. *Define the question of interest:* This describes the specific question, decision, or concern to be addressed by the AI model.

2. *Define the context of use:* this is a description of how the model will be used to address the question of interest, i.e. the specific role and scope of the AI model.

3. *Assess the AI model risk:* this is defined by (i) the contribution of the evidence derived from the AI model relative to other contributing evidence used to inform the question of interest, i.e. model influence; and (ii) the significance of an adverse outcome resulting from an incorrect decision concerning the question of interest, i.e. decision consequence. The ratings for decision consequence and model influence are independently determined, but are shaped by the context of use, thus enabling model risk to be case specific. The AI model risk is assessed by subject matter expertise and judgement on the possibility of model output leading to an adverse outcome, rather than

the intrinsic risk of the model itself. As illustrated in Figure 4, the model risk moves from low to high as decision consequence or model influence increases.

Figure 5: Model risk matrix
Source: U.S. Food and Drug Administration.[169]



Algorithmic impact assessment

In Canada, the mandatory algorithmic impact assessment (AIA) tool[170] is designed to help departments and agencies better understand and manage the risks associated with automated decision systems. It is composed of questions in various formats that consider many factors (e.g. system's design, algorithm, decision type, impact, data) within risk and mitigation areas and contribute to a scoring system. The value of each question is weighted based on the level of risk it introduces or mitigates in the automation project. The resulting impact levels (from I: little impact, to IV: very high impact) determine the mitigations required under the Directive on Automated Decision-Making.[171]

Issue detection and risk mitigation

Defining when to mitigate requires knowing how to detect issues based on a pre-defined risk-proportionate testing and verification plan which is laid out during the development of the AI system. Testing and verification are essential steps of Computerized System Validation (CSV), which considers different levels based on AI system maturity. The latest version of the Good Automated Manufacturing Practice 5 (GAMP 5) of the International Society for Pharmaceutical Engineering (ISPE), a framework widely adopted by pharmaceutical companies and health authorities, contains an appendix focusing on AI and ML.[172] Testing should be based on pre-defined key performance indicators and acceptance criteria, considering the human performance, and account for the identified risk areas, e.g. low quality data. Issues can be detected using mechanisms such as 'golden questions', i.e. known truths which are checked each time an AI system undergoes a change.

After the AI system has been proven fit for purpose and deployed, an ongoing process should be in place to monitor its performance and trigger mitigation measures when issues are detected.

A risk-based approach may be very conservative in the initial stages of deployment with additional pre-determined mitigation measures in place, for example, high percentage of human-in-the-loop. As confidence in the routine performance increases over time, based on pre-defined indicators and examination of sample outputs by human experts, a gradual reduction in the frequency, amount (e.g. number of samples) or depth of human controls may be considered.

When issues or performance deviations are detected, risk-based mitigation measures may include:

- *Human-in-the-loop:* Increased or full human review/quality control, indefinitely or until performance levels are back within acceptance criteria, e.g. if a seriousness detection

925 algorithm fails to detect seriousness criteria in some cases, e.g. false negatives, mitigation
926 could involve reviewing all cases classified as non-serious until the issue is understood and
927 addressed;
928 • *AI improvement* strategies such as model re-training or hallucination mitigation strategies;[173]
929 • Articulation of the *level of uncertainty* in AI outputs;
930 • Approaches to *combat automation bias or complacency*,[174] e.g. mock data simulations or
931 injection of simulated false positive outputs for verification/assessment;
932 • *Decommissioning* of the tool when mitigation options appear inefficient or costly, in which
933 case alternative approaches should be considered.

934 Finally, AI components, especially those deployed in critical PV processes, should be included in the
935 organisation's business continuity plan.

936 The above aspects are further developed in the Chapters on Validity & Robustness, Human
937 Oversight, and Governance & Accountability.

938 The risk-based approach should be reviewed at regular pre-determined intervals to adapt oversight
939 measures based on performance data but also as new technical options for risk mitigation emerge.

940 Documentation

941 The key components of the AI-related risk management strategy should be documented, including:
942 • AI system risk assessment;
943 • Testing plan with key performance indicators and acceptance criteria including any
944 comparative assessments;
945 • Planned mitigation measures including human in-the-loop strategy and criteria for more
946 stringent or reduced quality control, and continual monitoring after deployment;
947 • Plans for periodic re-assessment and update of the risk management strategy;
948 • Business continuity plan.

References

[151] *Council for International Organizations of Medical Sciences. Practical Approaches to Risk Minimisation for Medicinal Products. Geneva, Switzerland: Council for International Organizations of Medical Sciences; 2014. (Full text accessed 21 March 2025)*

[152] *European Medicines Agency. Guideline on good pharmacovigilance practices (GVP): Module I – Pharmacovigilance systems and their quality systems. (Full text accessed 21 March 2025)*

[153] *U.S. Food and Drug Administration. Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment. Guidance for Industry. March 2005. (Full text accessed 21 March 2025).*

[154] *European Medicines Agency. Guideline on good pharmacovigilance practices (GVP): Module I – Pharmacovigilance systems and their quality systems. (Full text accessed 21 March 2025)*

[155] *World Health Organization. Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO. (accessed 21 March 2025)*

[156] *European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. (Webpage accessed 21 March 2025)*

[157] *U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry and Other Interested Parties. January 2025. (Full text accessed 21 March 2025)*

[158] *European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. (Webpage accessed 21 March 2025)*

[159] *European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. (Webpage accessed 21 March 2025)*

[160] *Government of Canada. The Artificial Intelligence and Data Act (AIDA) – Companion document.(Webpage accessed 21 March 2025)*

[161] *European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. (Webpage accessed 21 March 2025)*

[162] *U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry and Other Interested Parties. January 2025. (Full text accessed 21 March 2025)*

[163] *Hines PA, Herold R, Pinheiro L, Frias Z, Arlett P. Artificial intelligence in European medicines regulation. Nat Rev Drug Discov. 2023;Feb;22(2):81-82. (Journal full text) https://doi:10.1038/d41573-022-00190-3.*

[164] *Kjoersvik O, Bate A. Black Swan Events and Intelligent Automation for Routine Safety Surveillance. Drug Saf. 2022;May;45(5):419-427. (Journal full text) https://doi:10.1007/s40264-022-01169-0.*

[165] *Hauben M. Artificial intelligence in pharmacovigilance: Do we need explainability? Pharmacoepidemioly Drug Saf. 2022;Dec31(12):1311-1316. (Journal full text) https://doi: 10.1002/pds.5501.*

[166] *U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry and Other Interested Parties. January 2025. (Full text accessed 21 March 2025)*

[167] *U.S. Food and Drug Administration. Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions. 2023. (Full text accessed 21 March 2025)*

[168] *International Conference on Harmonisation (ICH). General Principles for Model-Informed Drug Development. M15. Draft version endorsed on 06 November 2024. (Full text accessed 21 March 2025)*

[169] *U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry and Other Interested Parties. January 2025. (Full text accessed 21 March 2025)*

[170] *Government of Canada. Algorithmic Impact Assessment tool. (Webpage accessed 21 March 2025)*

[171] *Government of Canada. Directive on Automated Decision-Making. (Webpage accessed 21 March 2025)*

[172] *The International Society for Pharmaceutical Engineering (ISPE). GAMP 5 Guide 2nd Edition. 2022. (Abstract accessed 21 March 2025).*

[173] *Hakim JB, Painter JL, Ramcharran D, Kara V, Powell G, Sobczak P, Sato C, Bate A, Beam A. The Need for Guardrails with Large Language Models in Medical Safety-Critical Settings: An Artificial Intelligence Application in the Pharmacovigilance Ecosystem. arXiv preprint arXiv:2407.18322. 2024;Jul1. (Journal full text)*

https://doi.org/10.48550/arXiv.2407.18322

[174] *Adler-Milstein J, Redelmeier DA, Wachter RM. The Limits of Clinician Vigilance as an AI Safety Bulwark. JAMA. 2024;331(14):1173-1174. (Journal full text) https://doi:10.1001/jama.2024.3620.*

Chapter 4: Human oversight

Principle

Human oversight refers to the expected role of humans in the design, implementation, monitoring, and analysis of AI systems in PV. It requires a framework to manage performance and to detect and mitigate potential issues related to the AI system.

Key messages

- Human oversight supports the optimisation of the performance of AI systems deployed in PV and increases trustworthiness and accountability.
- The extent and nature of human oversight for an AI solution should follow a risk-based approach.
- Quality assurance principles should apply to the conduct of the human oversight of AI systems in PV.
- The increased use of automation and AI to support PV processes will require redefining skillsets to integrate AI with human expertise, ensuring robustness and reliability in decision-making processes. This will lead to a transformation of traditional roles and competencies that requires appropriate change management and training strategies.

Introduction

Motivation

Human oversight is required to minimise the risk that an AI system undermines human autonomy or causes other negative or unintended effects.[175] Human agency and oversight are a key requirement of trustworthy AI according to several regulatory frameworks, including the Assessment List for Trustworthy Artificial Intelligence (ALTAI), the EU AI Act, and the Canadian Artificial Intelligence and Data Act (AIDA), for high-risk systems[176,177,178] (see also Chapter on Landscape analysis). Although human review by itself does not guarantee full accuracy of outputs, human oversight is essential to monitor the performance of AI systems and make corrections if needed, thereby increasing trustworthiness and accountability for the AI system, especially in some high-risk applications.

AI systems are often intended to help eliminate manual, labour-intensive or complicated work performed by humans, or to enhance human performance when used as intelligence augmentation tools. However, due to the complexity and sensitivity of certain PV tasks, and the complex and variable nature of PV data, AI components will exhibit increasingly good but imperfect performance. This may require more extensive human intervention during the development, evaluation and deployment of some AI solutions in PV to monitor and mitigate risks.

A key challenge and important starting point for defining an AI quality assurance (QA) approach is to strike a balance between the efficiency boost that an AI system is intended to provide and the level of human intervention that may be required to ensure a high-quality output. In plain words, ideally a human expert should not do work that a machine can do well, and a machine should not do poorly the work that a human expert can do well.[179]

Human oversight is fundamental to a sound risk-based approach (see Chapter on Risk-based approach). The level of monitoring of the performance of AI systems by humans should be proportional to the potential impact on patient safety of an undetected mistake or spurious output by the AI system.

Considerations on human involvement and oversight

Multidisciplinary expertise

The successful integration of AI solutions into PV systems requires that multidisciplinary human expertise is mobilised as appropriate throughout the lifecycle of the tools, from development to routine use. This multidisciplinary expertise is usually obtained through a close collaboration between PV professionals, QA staff, data scientists, statisticians, AI/ML engineers, data engineers, prompt engineers, IT specialists, cybersecurity experts, platform analysts, software engineers, ethics specialists, legal experts, data protection officers, project managers, senior management, etc. (see also Chapter on Governance & Accountability).

PV professionals, i.e. staff performing core tasks in ICSR management, signal detection and analytics or risk management, hold robust 'domain' or 'subject matter' expertise, which is instrumental to effective integration of AI capabilities into PV processes. As such, PV professionals should be engaged in the design, development, pre-deployment and testing/piloting of AI solutions to ensure that they are fit for purpose and widely accepted by the end-users that they - PV professionals - will ultimately be.

Mechanisms of human oversight

Human oversight may serve different objectives and be achieved through governance mechanisms at different stages.[180] There are various possible approaches based on the activity monitored and how much autonomy is granted to an AI system. Depending on the scope, extent and intensity of human intervention, the European Commission's Ethics Guidelines for trustworthy AI describe three main governance mechanisms: human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC). HITL refers to the capability for human intervention in every decision cycle of the AI system. HOTL, which foresees a higher autonomy of the AI system, refers to the capability for human intervention during the design of an AI system and monitoring of its operation. HIC refers to the capability to oversee the overall activity of an AI system, including its broader economic, societal, legal and ethical impact, and the ability to decide when and how to use an AI system. This may include the decision not to use an AI system in a particular situation, to establish levels of human discretion during its use, or to ensure the ability to override a decision made by the system.[181] The delineations of these three terms may vary according to sources[182] and their practical implementation may differ according to individual organisations and use cases. For example, during the lifecycle of a given AI system in PV, human oversight may be exercised at early stages to help define the system's context of use or support the identification or development of reference datasets (HOTL) and, when deployed, to perform quality controls of the system (HOTL) or as part of its execution in case of a semi-automated system (HITL).

As a rule, some level of human oversight is always required and the absence of a human-in/on-the-loop in any major or supporting PV process should be substantiated by a risk assessment, with risk mitigation measures in place.

Monitoring and interacting with deployed artificial intelligence systems

The level, frequency, means and modalities of human intervention required to monitor and interact with AI systems depend on the complexity of the task, the risks associated with suboptimal outputs, the type of AI system, and the performance as assessed during validation (see Chapters on Risk-based approach and Validity & Robustness). As experience with AI evolves, further clarity, guidance, and consistency in assessing these factors are likely to develop. As suggested above, the respective roles of the human and AI components in a particular process could be seen as a continuum, from an

AI tool merely performing preparatory work to support assessment and decision making by a human, to a near-fully automated system merely monitored by a human who performs quality controls. Intermediate approaches may also be envisaged where, for instance, an AI solution flags cases it struggles with to a human specialist.

The metrics and KPIs used to monitor the performance of deployed AI systems should be pre-defined as part of the testing and verification plan (see Chapters on Validity & Robustness and Risk-based approach).

In situations where the standalone performance of an AI system is suboptimal (e.g. if it cannot match the established human performance, or when the associated risks are unacceptably high), one or more manual process steps must be considered, with a human fully in control of the final output. Even when a static AI-based system exceeds human performance upon validation, monitoring after development is still recommended to ensure that the performance does not fall below acceptable levels over time (see Chapter on Validity & Robustness). Each time an AI system undergoes modifications, human oversight should be directed at the change i.e. change-specific samples should be prioritised.

There are different ways the performance of an AI system can be monitored once deployed. In a static AI system, one could perform a one-off retrospective analysis by checking a sample or the totality of generated outputs against expected outputs (see Chapter on Validity & Robustness). This may be followed by post hoc corrections and re-training or re-validation of the model. A more dynamic real-time, in-process interaction can also be envisaged where independent human assessment is applied to confirm or correct the AI output in a decision-support setting. In such a dynamic AI application, the interaction provides an opportunity for immediate feedback to the algorithm to continuously learn and adjust if needed. Running an independent model in parallel to the main AI system may also be an option in a one-off or continuous manner (AI-assisted human oversight).

Caution is required in the monitoring of GenAI/LLM-based systems. Humans-in/on-the-loop should be aware of the inherent variability of outputs, limited explainability and risk of hallucinations, and not overly rely on the AI system's results. Processes must be robust, demonstrated to be effective, and maintain their dependability even in the event of erroneous outputs. Hallucinations, specifically, may lead to seemingly coherent and convincing outputs that may be deceiving for humans-in/on-the-loop. Regardless of the underlying AI technology, PV professionals should be empowered to challenge the system's outputs based on their experience and avoid falling for automation bias. On the other hand, they should be aware of the possibility of confirmation bias and remain open to the possibility that an AI output, albeit unexpected, is correct. AI solutions with high performance may also warrant specific monitoring strategies as humans are more prone to miss very rare[183]

Transformation of traditional roles

As the PV landscape continues to embrace AI capabilities, a reduced dependency on large workforces with PV expertise is expected due to the replacement of some of the activities traditionally performed by PV professionals. Indeed, the increased use of automation and AI within PV processes will unburden PV professionals from certain repetitive, time-consuming, manual activities. This may render certain roles obsolete and thereby reduce the size, diversity and experience of the PV workforce, not unlike the impact on staff observed when organisations offshore activities. This may create legitimate concerns and anxiety about job displacement and employment prospects in the PV space, but also around work culture, motivation and fulfilment. Perceived unfairness may also ensue from the fact that some AI models are trained using historical datasets and documented decisions based on the work originally performed by PV professionals.

On a brighter side, the introduction of AI in PV brings opportunities for growth for PV professionals. With fewer menial time-consuming tasks, PV experts will be able to focus on more scientifically

complex and intellectually stimulating PV activities. In addition, the business needs associated with AI solutions will bring new roles in governance and human oversight. As mentioned earlier, PV professionals will be increasingly involved in the testing, evaluation, implementation, oversight and use of AI models. They will often be best placed to identify those activities in need of automation and suggest AI use cases accordingly. They may have to participate in design and development activities including model training and validation, participate in user acceptance testing, manage the challenges of automating and modifying existing processes, perform monitoring and quality control activities, identify and resolve issues related to inconsistent assessments, and interact with automation experts and vendors.

Contributing to the development, use, and maintenance of AI systems will allow PV professionals to evolve with the changing PV landscape, but this will require that they extend their skillsets beyond core PV competencies.[184] These new skills include specific competencies around the use of the new systems and the critical evaluation of their outputs, as well as more general literacy around data science and AI, including a good understanding of AI capabilities, risks and limitations. Regulatory frameworks such as the EU AI Act impose an obligation on organisations to ensure a sufficient level of AI literacy of staff operating or using deployed AI systems.[185]

Beyond PV professionals, staff working in QA also need to develop an understanding of the organisation's human oversight strategy in addition to some AI literacy, to ensure that human oversight activities are adequate. Likewise, AI experts involved in the design and development of AI in PV solutions will need to develop an understanding of PV processes and the implications of operating in a regulated environment.

Change management and readiness strategies are a key responsibility of organisations, which should put PV staff at the centre of role redefinition and upskilling opportunities. Adequate change management and training plans are a pre-requisite to a seamless, safe and successful integration of AI systems into PV processes, with a wide engagement and adoption by staff and smooth interactions between various roles (see also Chapter on Governance).

Training programs should be carefully crafted, documented and evaluated so that their content and format (including materials and methods) meet the learning needs of the target audience (e.g. PV end-users, QA staff, AI experts). Human training in a decision-support context is an approach that may be drawn on to train staff monitoring and interacting with AI systems. It generally refers to programs designed to educate staff to use specific tools and make informed decisions effectively. This involves not only showing staff how to use the software front-end but also explaining the back-end functionalities and helping them build the skillset for critically evaluating the automated output. Training modalities (e.g. classroom-based vs online, live vs asynchronous) should be adapted to the system's complexity, the supported use case or task, and the specific needs of the organisation or the individual.[186,187,188,189]

References

[175] *European Commission. Ethics guidelines for trustworthy AI. 2019. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. [Accessed 16 August 2024] European Commission. Ethics guidelines for trustworthy AI. 2019. (Webpage accessed 21 March 2025)*

[176] *European Commission. Ethics guidelines for trustworthy AI. 2019. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. [Accessed 16 August 2024] European Commission. Ethics guidelines for trustworthy AI. 2019. (Webpage accessed 21 March 2025)*

[177] *European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. (Webpageaccessed 21 March 2025)*

[178] *Government of Canada. The Artificial Intelligence and Data Act (AIDA) – Companion document. (Webpageaccessed 21 March 2025)*

[179] *Ball R, Dal Pan G. "Artificial Intelligence" for Pharmacovigilance: Ready for Prime Time? Drug Saf.2022;May;45(5):429-438. (Journal full text) https://doi:10.1007/s40264-022-01157-4.*

[180] *Enqvist L. 'Human oversight' in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology. 2023;Jul 3;15(2):508-535. (Journal full text) https://doi.org/10.1080/17579961.2023.2245683*

[181] *European Commission. Ethics guidelines for trustworthy AI. 2019. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. [Accessed 16 August 2024] European Commission. Ethics guidelines for trustworthy AI. 2019. (Webpage accessed 21 March 2025)*

[182] *Enqvist L. 'Human oversight' in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology. 2023;Jul 3;15(2):508-535.*

[183] *Rich AN, Kunar MA, Van Wert MJ, Hidalgo-Sotelo B, Horowitz TS, Wolfe JM. Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. Journal of Vision. 2008;Nov1;8(15):15. (Journal full text)HYPERLINK "https://doi.org/10.1167/8.15.15"https://doi.org/10.1167/8.15.15*

[184] *Danysz K, Cicirello S, Mingle E, Assuncao B, Tetarenko N, Mockute R, et al. Artificial Intelligence and the Future of the Drug Safety Professional. Drug Saf. 2019;Apr;42(4):491-497. (Journal full text)https://doi:10.1007/s40264-018-0746-z.*

[185] *European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. (Webpageaccessed 21 March 2025)*

[186] *Power DJ. A brief history of decision support systems. DSSResources. com. 2007;Mar10;3. (Webpageaccessed 21 March 2025)*

[187] *Wang RY. A product perspective on total data quality management. Communications of the ACM.1998;Feb1;41(2):58-65. (Journal full text) https://doi.org/10.1145/269012.269022*

[188] *Turban E. Decision support and business intelligence systems. Pearson Education India; 2011.*

[189] *Keen PG. Decision support systems: a research perspective. Decision Support Systems: Issues and Challenges, International Institute for Applied Systems Analysis (IIASA) Proceedings Series. 1980 Jun 23;11:23-7. (E-book abstract)*

Chapter 5: Validity & Robustness

Principle

- Validity means that a system achieves its intended purpose within acceptable parameters. It requires predefining acceptable performance levels, selecting appropriate data for model training and/or testing, assessing model performance in a realistic setting, and integrating the system into an ongoing quality assessment process.
- Robustness means that a system reliably achieves its intended objectives (while accounting for variations in data).

Key messages

- PV professionals and decision makers must learn to critically appraise proposed AI solutions whether they acquire them or participate in their development.
- A performance evaluation able to demonstrate acceptable and robust results for the intended use under realistic conditions is crucial. Such an evaluation should cover a wide enough range of relevant examples to interrogate the model's objective and is often based on statistical metrics.
- There should be a focus on looking to ensure sufficient representation of relevant types of data in the test set(s) to detect biases, promote adequate and generalizable performance across the intended deployment domain, assess usability, and identify circumstances where the model may underperform.
- Many PV applications focus on very rare events or patterns (e.g. emerging safety signals, reports of a certain kind – such as related to pregnancy, duplicated reports etc.) and may require enrichment strategies to obtain representative test sets with high enough prevalence of the event of interest. If so, special care should be taken to ensure that performance evaluation results generalize to real-world settings.

Introduction

Ensuring the validity and robustness of AI solutions is central to building trust and achieving the best possible value for end-users. To invest resources optimally, PV professionals and decision makers must learn to critically appraise and evaluate proposed AI solutions regardless of whether they develop them in-house or acquire them from other organizations. This requires familiarity with basic principles for performance evaluation and some of the common pitfalls that may mislead expectations on real-world performance in prospective use.

AI solutions will often be embedded in broader computer systems supporting the PV use case. These should be subjected to general computer system-validation according to standard practices for the organization. In general, this will be considered orthogonal to ensuring the validity and robustness of the core AI solution (and is out of scope for this document). However, some special considerations regarding validation of systems that include dynamic AI models that continually learn from and adapt to incoming data are presented in the Section on Continuous integration and deployment. Our focus will be on key considerations related to establishing the validity and robustness of AI models themselves, including their dependency on underlying data for training and deployment and the need for probabilistic / statistical performance evaluation.

The nature of PV data may in some instances impact our ability and approach to leveraging AI solutions. AI models depend heavily on the quality of the data they are trained on and the data they use for ongoing predictions. PV data suffer from inconsistencies, incomplete entries, and inaccuracies, and may vary substantially depending on the source. For example, the contribution of individual case reports is for the most part voluntary, and reporting practices vary over time,

between organizations and types of reporters. This may impact the types of adverse events that are reported, which information is captured, and how it is encoded. Inconsistencies and inaccuracies can lead to models that are less accurate, and systematic variability can reduce the generalizability of AI solutions to adjacent domains and make them more sensitive to data drift. They may also make it more difficult to ensure consistent performance across regions and organizations (see also Chapter on Fairness & Equity).

Generally, the variable quality and consistency of individual case reports and the complex nature of the studied drug-event relationships may require more extensive human involvement than in other domains to ensure the validity and robustness of AI solutions for PV (see also the Chapter on Human oversight). The practice of PV is subject to medicines regulation, and regulatory expectations regarding validity and robustness may differ from those of the business itself. For example, even if from a business perspective an organization were prepared to adopt an AI solution whose output was not fully reproducible in repeated execution using the same input, the organisation would not consider the acceptability to regulators, as sufficient controls may exist to de-risk the process to which it is applied.

Performance evaluation and testing are crucial considerations in ensuring the validity and robustness of AI solutions. While it will usually be more effective to account for these considerations also during development and training of AI solutions, to do so might not be feasible or required. For example, LLMs can be capable of zero-shot learning, with solid performance on language tasks for which they have not been specifically trained. What is important even then is to demonstrate adequate performance on the relevant tasks in independent testing with conditions reflecting the intended use.

Specification and design

Use case and deployment domain

The intended use case and deployment domain for AI solutions in PV should be clearly defined, and the performance evaluation targeted to these, as far as possible. For example, in evaluating methods for PV signal detection, historical safety signals would typically be a more relevant basis for performance evaluation than well-known, already labelled adverse drug reactions since their reporting patterns differ in important ways[190] Similarly, if an AI solution for recognizing adverse events in free text is intended for broad use, its evaluation should include reports related to various medicinal products and adverse events, from both patients and health professionals, in relevant languages, etc. Ideally, there should be sufficient representation of relevant types of data in the test set to detect biases, promote adequate and generalizable performance across the intended deployment domain, assess usability, and identify circumstances where the model may underperform.

Multi-disciplinary collaboration

Ensuring the validity and robustness of AI solutions often requires collaboration across disciplines, including not only PV decision makers and practitioners, but also for example, data scientists and AI experts, and individuals with experience in computer systems validation. Diverse perspectives and expertise, in-depth understanding of a model's intended integration into the PV system and defined desired benefits and associated risks can help ensure that deployed AI solutions are effective, and address identified needs over their lifecycle.

Applications of AI solutions pertaining to the complex relationships between drugs and adverse events often require a human-in-the-loop, especially in view of the variable quality and provenance of the underlying PV data. AI outputs in such applications need to be interpreted considering the broader clinical context, known pharmacological mechanisms, and possible alternative explanations

which may not be captured in the data at hand and that the AI might not fully account for. Human intervention ensures that the final output is clinically meaningful and scientifically sound. On the other hand, more basic tasks such as redaction of personal data or drug and adverse event encoding may lend themselves to automation with minimal human intervention.

Definition of reference standards

Test sets must be aligned with intended deployment domain and able to demonstrate performance under realistic conditions. Reference standards relevant to the intended use need to be clearly defined and kept up to date. In many PV applications, these may be based on human execution of the task in question. Approaches to mitigate inconsistencies are often required, for example by having multiple human assessors annotate (parts of) the same data. When legacy human annotations are used as the reference standard, efforts should be made to clarify the definitions of relevant categories in the reference standard retrospectively, and to ensure that all included historical annotations adhere to these standards and are relevant for the intended future use. This may require the omission of available annotations that were developed following outdated principles or were based on different types of data. If reference standards are to be developed de novo, an explicit annotation guideline is recommended. This in turn may require a strengthening and clarification of existing processes and guidelines for human decision making on the PV task of interest, sometimes bringing value by harmonizing and making explicit decision processes that may otherwise remain implicit and variable within an organization. In smaller projects with a limited number of participants, special care may be required to ensure that annotations of the test set used for performance evaluation are independent of the development of the AI solution, for example annotations may be performed preferably by individuals with limited insights and vested interest in a specific AI solution to avoid conflicts of interest and confirmation bias. Similarly, if testing human-AI teams, the qualifications of human team member(s) should match those of the intended use case and deployment domain.

Sometimes, boundaries between reference standard categories are not clear, which yields additional sources of possible ambiguity. For example, different organizations may have different requirements on how strong the conviction should be that two individual case reports refer to the same event for them to be classified as suspected duplicates. This may vary even within an organization depending on the intended use case, for example one may cast a wider net in highlighting suspected duplicates if each highlighted pair will be reviewed by a human before action and be more conservative if suspected duplicates will be automatically removed prior to statistical signal detection. Similar ambiguities exist in natural language processing tasks seeking to map free text to standard terminologies such as MedDRA where there may be multiple acceptable terms/codes for a specific verbatim, and it may be inappropriate to treat terms adjacent to the reference standard annotation as false positives.

For unsupervised learning like cluster analysis and representation learning, and for applications of zero-shot learning like text summarization, formal test sets often cannot be obtained and other approaches to performance evaluation must be considered. In some cases, one may rely on human subjective review and assessment of an AI solution's output, but then potential biases must be considered and mitigated. For example, one may present the results of several different AI solutions to a blinded, domain expert and ask which they prefer. There are also performance evaluation approaches specifically designed for unsupervised learning like intruder detection analysis.[191]

A general challenge in PV has been ensuring sustainable and reusable access to reference sets. As AI technologies rapidly advance, the necessity for consistent and frequent testing becomes increasingly important to safeguard against unintended consequences of AI usage. The potential for widespread impact of AI solutions in PV underscores the importance of maintaining up-to-date, accessible reference standards with clarity on how they were developed and related assumptions.

Performance evaluation

Performance evaluation is necessary for critical appraisal of AI solutions. The ability to carry out or assess performance evaluations are crucial skills for those who develop AI solutions and for those to whom AI solutions are proposed.

Many of the metrics relevant to performance evaluation for AI solutions in PV come from information retrieval and apply primarily to use cases that can be viewed as binary classification tasks. In binary classification, we may refer to those instances that we want a method to retrieve as *positive controls* and those that we do not want it to retrieve as *negative controls*. We use this terminology throughout the description below (sometimes replacing positive controls by *target events*), acknowledging that other use cases may require different frameworks of evaluation, for example considering an ordering.

AI solutions, like humans, will not always achieve perfect performance on complex tasks. Therefore, performance is typically assessed statistically for a sample of cases referred to as the test set, considering measures such as precision (= positive predictive value) and recall (= sensitivity) relative to the selected reference standard. The balance between precision and recall (and correspondingly between sensitivity and specificity) should be determined based on the relative costs of different types of errors (and utilities associated with correct decisions). Composite metrics like the F1 score (the harmonic mean of precision and recall) provide single-dimensional measures of predictive accuracy accounting for both precision and recall under some assumptions (for the F1 score that precision and recall are of equal importance and false positives as costly as false negatives). Test sets need to be large, diverse, and representative enough to reflect a sufficient portion of the intended deployment domain and to provide statistically robust estimates of performance. They should include different populations and consider possible scenarios in line with the intended use.[2]

Since the primary interest is the expected performance of an AI solution in prospective use (as part of an overall system), performance evaluation should be independent of any data directly used during its development (this is in addition to any cross-validation or other separation of data for training and validation during development). Various potential sources of dependence between development and evaluation must be considered and eliminated, the most obvious being the risk that the same individual data points are considered in both phases. More subtle forms of dependence, can occur and lead to optimistic performance estimates, for example there may be a disproportional overlap in scope between the training and test sets compared with the deployment domain e.g. if training and test sets cover the same subset of drugs and adverse events, which can be referred to as a specific form of data leakage.[192]

Selection of machine learning models may also account for indirect performance characteristics such as an AI model's susceptibility to overfitting, computational cost and robustness to outliers, especially if test sets are not large and diverse enough to be reliably capture their impact during performance evaluation. When comparing different types of methods, any user-driven design decisions should be fixed and finalized before developers first access test sets. This is especially important for more complex methods with numerous analytical choices regarding model architecture, hyper-parameters, and model initialization.[193] On a related note, complex methods highly dependent on skilful design and deployment by human experts may not readily transfer to adjacent application areas without access to the same expertise. In routine deployment one is less concerned about whether one method is theoretically better than another but rather with which one is likely to perform best for a given purpose, irrespective of what design/analytical choices one made.

[2] For a continually updated inventory, see for example https://oecd.ai/en/catalogue/metrics

Benchmarking

Ideally, performance should be compared against relevant benchmark methods, if available. For example, AI-based signal detection methods may be compared against standard disproportionality measures. In the case of more complex benchmark methods, including those based on AI models, special care must be taken to ensure that the benchmark methods have been appropriately instantiated and fine-tuned to the task at hand to serve as a relevant comparator.

When public benchmark test sets exist, performance may be evaluated against these, ideally as a complement to performance evaluation targeted to the deployment domain of interest. At present, public benchmarks exist only for some specific applications in PV. They include sets of emerging safety signals,[194,195] sets of established adverse drug reactions,[196,197,198,199,200] and clinically relevant drug-drug interactions.[201] However, continual access to benchmark reference sets over time can be a challenge and the degree to which they are maintained and kept up to date varies.

To complement overall performance estimates, subgroup analyses can provide useful information on the strengths and weaknesses of the AI solution for different parts of the deployment domains (See also chapter on Fairness & Equity). Along the same lines, sensitivity analyses can help assess the robustness of the AI solution and its evaluation to variations in specification and design.

Special considerations for low-prevalence settings

Many PV applications focus on rare patterns and events. For example, in a case retrieval task most reports will typically not be relevant for a given topic, such as pregnancy, medication errors, positive rechallenge interventions, or drug-induced liver injury. Similarly, for PV signal detection, most drug-event combinations are not true adverse drug reactions, let alone recently detected safety signals. Managing and analyzing these rare events effectively requires reliable reference datasets, however, existing resources, such as SIDER, are often limited by outdated and static information, underscoring the need for alternative solutions.[202] As an even more extreme example, pairs of duplicate reports are vanishingly rare among all possible pairs of reports in large collections of individual case reports – if 10% of the reports in a database of 1 million reports have a (single) duplicate, the chance that a randomly selected pair would be duplicates is only 1 in 10 million.[3]

This low prevalence of positive controls (i.e. class imbalance) limits our ability to achieve accurate performance evaluation and requires special care and consideration. For example, a balance may need to be struck between the quality of each annotation and the resulting size of the test sets (or the cost/time to develop them), for example related to whether double annotations by multiple assessors are feasible to increase quality or evaluate consistency. The heterogeneity in skills and preferences among different human reviewers poses a significant challenge to achieving consistent quality in annotations. The use of intelligent automation to support decision making can mitigate inconsistency and reduce subjective bias in the evaluation process. Moreover, straight random samples of test cases often contain too few positive controls whereas test sets enriched with positive controls can lead to misleading estimates of precision and recall.

Recall measures how many of the target events are correctly identified (*recalled*) by the AI solution. *Sensitivity* is a synonym. If heuristics are used to increase the proportion of target events in the test set, then recall may be over-estimated since target events which are harder to identify for the AI solution, may less likely be included. This does not mean that rebalancing approaches should necessarily be avoided but if they are used, this should be acknowledged and critically assessed. *Precision* is the proportion of target events among all events highlighted by the AI solution. *Positive predictive value* (PPV) is a synonym. It is highly dependent on the prevalence of target events in the test set, and if test sets have been enriched with target events, naive test set precision estimates will be optimistic. For reliable precision estimates, the prevalence of positive controls in the test sets

[3] $0.10 \cdot 1/10^6$

should match as far as possible that of the intended deployment. For a specific AI solution, this is straightforward to obtain by applying the AI solution to a random sample and annotating all highlighted instances. However, such test sets are tied to the AI solution in question and will need to be developed again, or at least extended, if the solution is modified.

Estimates of precision and recall depend on the selected decision threshold, and performance evaluation should be targeted at decision thresholds relevant to the intended deployment domain, i.e. with a relevant balance between false positives and false negatives.

Beyond summary statistics

Summary statistics as captured by the metrics described in the previous section go only so far in enabling us to assess and understand the performance of an AI solution. Access to and ability to inspect representative, concrete examples of an AI solution's classification of individual instances in a test set is also valuable. Examining false positives and false negatives in an error analysis step can each give useful insights regarding the strengths and limitations of the AI solution and its evaluation. For example, if a false negative in de-identification corresponds to a full name preceded by 'Mr', this may undermine end-users' trust in the solution, even if overall recall is excellent. On the other hand, if the false negative is 'AF' and it is hard to know from the surrounding text if these are initials or an abbreviation for *atrial fibrillation*, then the overall precision metric may be viewed as potentially conservative. Review of correctly classified instances may in turn give insights regarding an AI solution's capacity to solve challenging tasks. Does it correctly classify more difficult cases or just the trivial ones? This may be especially important when there is no baseline comparator, and we may not understand from overall performance metrics the difficulty of the task at hand. When there is a baseline comparator method, one may focus on instances that are differentially classified by the two methods, to better understand the nature of any improved performance of the proposed solution over the comparator.

Reproducibility

Reproducibility for an AI solution requires that it will generate the same output for a given input. Predictive models like support vector machines and decision trees fulfil this requirement. So do certain LLMs and other deep neural networks, once their weights have been fixed at the end of training / fine-tuning, even though their model fitting may include stochastic components so that new weights may result if a model is re-trained on the same data.

GenAI solutions on the other hand include stochastic components also in their execution and will typically generate different outputs for the same prompt, without changes to the underlying models. The same is true for other methods such as mixture model-based cluster analysis and semantic vector representations of adverse events. For such solutions, stability is a key additional performance metric, reflecting how similar the results of fully replicated analyses are. While replicability of results can sometimes be artificially ensured through seeding the pseudo random number generator, this can be non-trivial to do for proprietary models and does not improve the inherent (in)stability of the AI solution, which should be evaluated.

Reproducibility, as described here, pertains to an organization that has full access to a specific AI model and the relevant reference sets. Full reproducibility by, say, the broader scientific community, in addition requires transparency.

Assessing artificial intelligence solutions with human-in-the-loop

Many AI solutions aim for intelligence augmentation, i.e. to support and enhance human decision making. In this context, the relevant focus of performance evaluation would be of the human-AI team. To date, we have limited experience of such studies in PV applications, but at a minimum, they

would need to account for the variability in skills and preferences between different human members of the team. Defining a relevant test set may also present new challenges: for example, for signal detection applications, human domain experts could not be blinded to historical safety signals; and it may be difficult to obtain a reference standard if the aim is for the human-AI team to *exceed* the quality of classification by unassisted human domain experts.

What constitutes acceptable performance may need to account for how the AI solution is integrated with the PV system and whether there is a human in the loop.[203] For example, performance evaluation for an NLP-based system to identify and extract adverse events from source documents might in addition to the overall performance evaluation consider whether errors can be readily spotted in the results and whether the end-to-end hybrid process performs better than a fully manual approach (for an example see Park et al 2023[204]).

Continuous integration and deployment

Deployed models should be monitored in real-world use with a focus on maintained or improved performance. There may be reasons to revise and update performance criteria in production as the business understanding of the task is refined or the conditions for the task itself change due to external factors.

For deployed AI solutions that incorporate ML components, there should be appropriate processes and quality controls for periodical re-training to manage risks of performance degradation or negative impact from dataset drift. In some instances, the retraining may consist of incremental fine-tuning within existing model architectures whereas more substantial changes to the deployment domain may require changes to the architecture of the AI solution. The latter could result from a change in scope from medicines to vaccines, revisions of the underlying medical terminologies or data structures, updated regulation or conventions and more.

Continual performance evaluation can be relevant regardless of whether an AI solution incorporates ML components or not. Its frequency should follow the risk-based approach and depending on the application, may include data-driven safeguards to identify, for example, substantial data drift or performance degradation triggering remedial actions that could include additional evaluation, and possible retraining, stopping use of the algorithm and/or introducing quality control measures to maintain confidence in its results. Documentation of activities and acceptance criteria for re-introducing AI solutions under such circumstances may also be required. As an example using mechanisms such as 'ground truths', i.e. known input/output pairs which are checked each time an AI system undergoes a change, or mechanisms to guard against automation bias.

One of the potential benefits of ML is the ability to improve performance through iterative modifications, including by learning from real-world data. To support this approach, the US FDA, Health Canada, and MHRA described a “Predetermined Change Control Plan” for ML-enabled device software functions (ML-DSF). Their general principles might conceivably be applied to AI solutions in PV. A Predetermined Change Control Plan generally includes: 1) a detailed description of the specific, planned modifications; 2) the associated methodology to develop, validate, and implement those modifications in a manner that ensures the continued acceptable performance of the algorithm; and 3) an Impact Assessment of the benefits and risks of the planned modifications and risk mitigations. The detailed description of the planned modification should include changes to the characteristics and performance of the algorithm resulting from the implementation of the modifications. An example of a modification might include retraining a ML model. A protocol providing the details of the data and methods used to develop, evaluate, and implement such a modification should be created and adhered to. An Impact Assessment of the modification should be carried out and risk mitigation measures developed to ensure that any identified risks will be controlled. This approach should be further incorporated into the quality management system governing the PV process being modified.

There should be straightforward means to report issues or anomalies encountered, and these should be addressed promptly. Ideally, the response would include acknowledging receipt of feedback, providing updates on investigations, and implementing necessary changes to the AI system.

References

[190] Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Safety. 2014;Sep;37:655-659. (Journal abstract) https://doi.org/10.1007/s40264-014-0198-z

[191] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, et al. (2009) Reading Tea Leaves: How Humans Interpret Topic Models. [Unpublished manuscript] (PDF accessed 27 April 2025)

[192] Rudin C, Radin J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition. Harv Data Sci Rev. 2019;1;1-9. (Journal full text) https://doi.org/10.1162/99608f92.5a8a3a3d

[193] Duin RP. A note on comparing classifiers. Pattern Recognition Letters. 1996;May1;17(5):529-536. (Journal abstract) https://doi.org/10.1016/0167-8655(95)00113-1

[194] Harpaz R, Odgers D, Gaskin G, DuMouchel W, Winnenburg R, Bodenreider O, Ripple A, Szarfman A, Sorbello A, Horvitz E, White RW. A time-indexed reference standard of adverse drug reactions. Scientific Data. 2014;Nov 11;1(1):1-0. (Journal full text) https://doi.org/10.1038/sdata.2014.43

[195] Sartori D, Aronson JK, Norén GN, Onakpoya IJ. Signals of adverse drug reactions communicated by pharmacovigilance stakeholders: a scoping review of the global literature. Drug Safety. 2023;Feb;46(2):109-120.) https://doi.org/10.1007/s40264-022-01258-0

[196] Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf.2013;Jan;36(1):13-23. (Journal abstract) https://doi.org/10.1007/s40264-012-0002-x

[197] Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf 2013;Oct;36(Suppl 1):S33-S47. (Journal abstract) https://doi.org/10.1007/s40264-013-0097-8

[198] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Research. 2016;Jan4;44(D1):D1075-9. (Journal full text) https://doi.org/10.1093/nar/gkv1075

[199] Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, Rogers W, Roberts K, Tonning J. A dataset of 200 structured product labels annotated for adverse drug reactions. Scientific Data. 2018;Jan30;5(1):1-8. (Journal full text) https://doi.org/10.1038/sdata.2018.1

[200] Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, Hirschman L, Ball R. ADE eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. Drug Safety. 2021;Jan;44:83-94. (Journal abstract) https://doi.org/10.1007/s40264-020-00996-3

[201] Kontsioti, E., Maskell, S., Dutta, B. et al. A reference set of clinically relevant adverse drug-drug interactions. Sci Data 9, 72 (2022). https://doi.org/10.1038/s41597-022-01159-y

[202] Painter J, Powell G, Bate A. PVLens: Enhancing Pharmacovigilance Through Automated Label Extraction. arXiv:2503.20639v2 [cs.CL] https://doi.org/10.48550/arXiv.2503.20639

[203] The US Food and Drug Administration. Good machine learning practice for medical device development: guiding principles.The US Food and Drug Administration: Silver Spring, MD, USA, 2021. (Webpage accessed 21 March 2025)

[204] Park J, Djelassi M, Chima D, Hernandez R, Poroshin V, Iliescu AM, Domalik D, Southall N. Validation of a natural language machine learning model for safety literature surveillance. Drug Safety. 2024;Jan;47(1):71-80.(Journal abstract) https://doi.org/10.1007/s40264-023-01367-4

Chapter 6: Transparency

Principle

Transparency regarding AI involves disclosing information between organizations or individuals. This includes sharing relevant documentation of the AI system lifecycle (i.e. design, development, evaluation, deployment, operation, re-training, maintenance and decommission) to facilitate traceability and providing stakeholders with enough information to have a general understanding of the AI system, its use, risks, limitations, and impact on their rights.

Key messages

- Declaring when and how AI solutions are used for core PV tasks is critical for building trust among domain experts, decision makers, regulatory authorities, and the public.
- The nature of AI solutions deployed for core PV tasks should be described including their model architectures, expected inputs and outputs, and the level and type of human-computer interaction.
- To give a comprehensive picture of an AI solution's effectiveness and limitations, the presentation of performance evaluation results should describe the scope and nature of the test set(s) used including definitions of their reference standards and sampling strategies. Presented performance metrics should be relevant for the intended deployment domain, compared with relevant benchmarks, and complemented by qualitative review of representative examples of correct and incorrect output.
- If possible, a description of the general principles and logic by which an AI model functions and arrives at its outcomes / predictions should be shared, or the lack of such explainability should be acknowledged and its implications discussed.

Introduction

Transparency provides stakeholders with relevant information regarding the nature and use of an AI solution. It reflects what information is shared with key stakeholders by those who develop or deploy it. The main purposes of transparency are to build trust, to enable individuals and organizations not involved in their development to inspect and scrutinize the design and performance of AI solutions, and to ensure regulatory compliance.

As further elaborated on in the chapter on Governance & Accountability, the primary direction of transparency and disclosure of information varies during the phases of the AI solution lifecycle. For example, during the design phase the organization should be transparent toward developers regarding the specification and requirements for an AI solution, whereas the main direction of transparency is the opposite in the pre-deployment phase. During routine use, the most important form of transparency may be from the organization toward end users (and in some cases regulatory authorities).

Disclosing use of artificial intelligence

It is essential to disclose why, when and how AI is being used in different PV tasks. This is to maintain trust and accountability among stakeholders, including developers, PV professionals and decision makers, regulatory authorities, healthcare professionals, and patients.

Regulatory bodies require disclosure of AI use to assure compliance with applicable laws and regulations. To this end, software vendors and internal development groups need to be transparent toward PV organizations, who in turn need to be transparent toward regulatory authorities. At the same time, those individuals who utilize AI solutions to process or analyze PV data must be informed about the AI's role in their workflows to help them integrate AI into their processes in an informed

manner to support its effective application and ensure that they can identify any issues arising from AI use.

PV professionals should also communicate the provenance of data elements and whether AI solutions contributed to their capture or development. Human interpretation of PV data may depend on how it was ascertained. For example, signal assessors may lend different weight to a case narrative that was auto generated from structured elements compared with one that documents the patients or health professionals' verbatim description of the adverse event. There is also a risk of a vicious circle where AI generated information is used as part of a reference standard in subsequent AI model development, if its provenance is not properly disclosed.

Transparency regarding the artificial intelligence model

Ensuring transparency of the AI models used in PV (to the extent possible), is critical to fostering trust, facilitating informed decision making, and ensuring that these models are applied appropriately. Ideally, transparency should be extended to also capture decisions made by PV professionals resulting from the AI model. Model transparency is not only a technical requirement but also an ethical imperative, ensuring that all parties understand the tools they are working with and can make informed decisions based on their outputs. Below are key aspects of an AI model that should be disclosed to stakeholders. The rationale behind the design choices should also be explained, to help ensure that the model is aligned with its intended use and stakeholder needs.

Table 3: Key aspects of an artificial intelligence model to disclose to stakeholders
Source: CIOMS Working Group XIV

Intended Use	The intended use of each AI model should be clearly defined and communicated. This includes specifying the PV tasks the model is designed to assist with or perform, such as adverse event recognition in free text, signal detection, or case triage.
Human-Computer Interaction	The level and type of interaction between humans and the AI models should be communicated. This includes specifying whether the AI model is executed autonomously, has a human in-the-loop (and what their required competence would be), or aims to provide decision support to down-stream human specialists.
Model architecture	The type of AI model and its general architecture should be disclosed, such as whether it is rule based, uses linear models, or specific types of neural networks, or combines different ML models in an ensemble, etc. Additionally, relevant details about the model's structure, such as the type and depth of a neural network architecture, should be shared.
Model parameters	At a minimum, key predictors or features that drive the decisions of an AI model should be disclosed, if they are known. If feasible, the full set of model weights and parameters can be shared, to enable external replication and external performance evaluation. See further discussion regarding this, at the end of this section. For AI solutions based on GenAI, any predefined prompts should be specified along with any pre- or post-processing steps.
Explainability	If possible, a description of the general principles and logic by which an AI model functions and arrives at its outcomes / predictions should be shared, or the lack of explainability should be acknowledged and its implications discussed. (*See also Section on Explainability).*
Training set	Details about the training set(s) based on which ML components have been developed should be disclosed. This would include their size, scope, and creation date, along with reflections on how well they align with the intended deployment domain.

Standard AI Components	If the AI model incorporates public standard components, such as pre-trained ML models, libraries, or frameworks, or datasets, this should be disclosed, including the specific versions used, date of access, and any custom parameter settings.
Acceptable Inputs	The types of inputs that the AI model expects should be specified. This provides insights regarding the basis for the AI model's outputs and ensures that it is only fed data it is designed to handle, thereby maintaining the accuracy and reliability of its outputs.
Type(s) of Output	The types of output generated by the AI model should be described. Examples may be risk scores, classifications, alerts, or free text.
Known Limitations	Any known limitations regarding the nature of the AI model should be communicated, including e.g. features or types of interactions, which it is unable to account for.

To allow other developers and researchers to fully replicate an AI model and possibly even modify it for further use, an organization might choose to publish its full set of parameters and weights or even share the source code. This level of openness supports peer review and validation by external experts, which can enhance trust in the model's reliability and foster innovation. However, it will not always be feasible due to considerations regarding intellectual property, competitive advantage, or the sheer complexity of large models. Moreover, for many stakeholders, access to raw code and parameters of a complex AI model may not enhance their understanding and will need to be complemented by the other measures for model transparency described above. Understanding the rationale, assumptions, and subjective decisions made in the implementation can be more important for gaining meaningful insights into the model's function and effectiveness. For full scientific reproducibility, developers may also need to share the relevant reference sets, at least those used for performance evaluation. However, depending on the use case and stakeholders involved this may conflict with the data privacy principle.

Explainability

A specific form of transparency relates to disclosure of the general principles and logic by which an AI solution operates and has arrived at a specific output. This may help nurture trust, allow affected individuals to understand and influence outcomes, support down-stream human decision making and facilitate human oversight and regulatory compliance. In this context, *explainability* and *interpretability* are important concepts, which partly overlap.

The set of *Guidelines on the testing of AI-based systems* in the ISO standard for Software testing in Software and systems engineering, characterizes explainability as a *"level of understanding how the AI-based system ... came up with a given result"* and interpretability to a "*level of understanding how the underlying (AI) technology works".*[205]

Similarly, the AI Risk Management Framework of the US National Institutes of Standards and Technology includes the following statement: "Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes".[206]

The Organisation for Economic Co-operation and Development (OECD) Transparency and Explainability Principle 1.3 states:[207]

> "Explainability means enabling people affected by the outcome of an AI system to understand how it was arrived at. This entails providing easy-to-understand information to people affected by an AI system's outcome that can enable those adversely affected to challenge the outcome, notably – to the extent practicable – the factors and logic that led to an outcome."

For the context of this report, we adopt a similar perspective and use *explainability* in a broader sense to reflect the degree to which humans can understand the factors and logic that have led to a specific outcome or that play a role in the general operation of an AI solution.

Concrete examples which illustrate the role of explainability in different PV use cases are provided in the Appendix 4.

Benefits of explainability

Explainability can be beneficial because it may:

- Nurture trust in an AI solution, by enabling stakeholders to make sense of and contextualize an AI solution's output;
- Allow individuals affected by an AI solution's output to challenge and influence the outcome;
- support and speed up human decision-making which builds on or integrates an AI solution output;[208,209]
- Propose scientific hypotheses for consideration by end users - individual or combinations of features such as drugs, diseases, and demographics that are included in the proposed explanation of the findings may provide signals of adverse drug reactions, and adverse drug-disease interactions worthy of evaluation, as well as potential biological mechanisms of adverse drug reactions;[210]
- Enable more complete documentation, audit, and human oversight of AI solutions;
- Contribute to regulatory compliance especially when it is possible to retain and examine the human decision together with the AI output and the explanation upon which the decision was based;
- Facilitate troubleshooting by revealing issues such as possible biases or likely spurious correlations;[211,212]
- Contribute towards model assessment and selection by uncovering what is causing different models trained on the same data to perform differently.

Referring to the definition above, the individuals who could challenge the output of the PV AI system and require explainability are more likely to be stakeholders who are directly involved in the PV process rather than members of the public.[213] They may range from the PV and quality assurance staff who are directly interacting with the AI, the developers who are building or maintaining an AI system to the regulators who are inspecting it. Examples on how different stakeholders in the PV process can benefit from explainability are provided in Appendix 4.

Inherent vs post hoc explainability

AI models of limited complexity may be inherently explainable, allowing the basis for their output to be deduced from direct inspection of their model architectures and parameters.[214] This is also referred to as *ante-hoc* explainability. Examples may include simple decision trees, rule-based classifiers, and regression models.

In contrast, a growing field of research seeks to obtain *post-hoc* explainability (referred to by the acronym xAI) for more opaque AI solutions, including deep neural networks with complex architectures and more parameters than a human can survey or comprehend. With such approaches, a separate layer of methods and techniques are applied top of the AI solution[215] to trace and explain the basis for a specific, already generated output. Some xAI approaches seek to explain the output of complex AI models by estimating relative feature importance and others do so by determining the minimal change in one or more features required to change a given output. There are also xAI methods that provide explainability for a specific output by fitting simpler, inherently interpretable

models to the local context of a specific output. For examples of specific xAI methods in use at the time of writing this report, please see Appendix 4.

When an xAI method is used to gain explainability, it must itself comply to the applicable regulatory requirements for computerised systems. In other words, the xAI method must be validated and proven to be fit for purpose, and processes must be in place to ensure that it continues to be so. Explanations provided by xAI are an 'approximate understanding of the relationship between the data and the predictions' according to Pinheiro et al (2022).[216] The explanations can be imperfect or incomplete and/or provide only a partial explanation.

Challenges related to explainability

Stakeholders are advised to critically consider what type of explainability is required for the intended use case, for whom, and for what purpose, and whether the system they are considering can provide it.[217]

The level of explainability of an AI solution's output should not be the sole determining factor for model selection. Some applications (e.g. machine translation) may not require inherent explainability and may depend on the capabilities and improved performance offered by more complex AI solutions. In such cases, the negative impact of limited explainability may be mitigated by ensuring high transparency regarding other aspects of the AI solution and its performance, coupled with extra care to achieve validity and robustness and human oversight.[218] At the same time, it should not be assumed that explainability necessarily leads to lower performance and that a trade-off between the two needs to be made.[219]

Similarly, while explainability of an AI solution's output can sometimes help identify issues with validity & robustness or fairness & equity, explainability alone does not prove that the system is fit for purpose, nor does it vouch for the trustworthiness of the system.[220] It attempts to clarify what factors led to a specific output but is not indicative of an AI solution's general performance or of its fairness and equity. For example, even if an inherently explainable AI solution does not include age as one of its explicit features, it could bias against an age group, if this bias is mediated by other features. In fact, explanations may make stakeholders more susceptible to overreliance on model outputs, so called automation bias.[221] Also, explainability is no guarantee of transparency – an organization may, for example, choose not to disclose the key features and inner logic of an inherently explainable model such as a decision tree.

Explainability is not the same for all stakeholders. What is understood by model developers could be incomprehensible for other stakeholders[222] and cognitive capabilities must be considered to ensure that explanations are comprehensible to humans.[223] Since humans will need to process and contextualize any explanations provided, they should also be informed about and aware of *their own* possible biases and blind spots which may influence their ability to leverage the explanations. Related to this, it should be noted that, in the worst case, a plausible explanation for an incorrect AI output may increase the likelihood that it is accepted without the appropriate critical review by some end users.

Transparency regarding performance

Transparency regarding an AI solution's assessed performance not only communicates how well an AI model operates in practice but also provides stakeholders with insights into the design, implementation, and decision-making processes behind the model. As such, it provides a bridge between theoretical capability and practical utility. Without a clear view of how an AI solution behaves under realistic conditions, stakeholders cannot fully assess its suitability for use or be confident in its robustness and validity. Performance transparency ensures that all stakeholders, from end-users to regulatory authorities, have a clear understanding of an AI solution's strengths,

limitations, and expected behaviour in the contexts where it will be deployed. This is particularly important in PV, where AI systems support information processing and decision making, with the aim of safeguarding patient safety and public health.

By recording and being able to share detailed performance evaluations with relevant stakeholders, organisations offer clarity on the strengths and limitations of the AI system, as well as the key assumptions made during their design, including quantitative metrics, qualitative examples, and comparisons to benchmarks, and thereby organizations provide the necessary context to build trust and appropriate reliance on AI systems. This transparency allows for informed decision making, ensures that AI systems are used within their intended scope, and helps identify areas where adaptations or special measures may be required. Additionally, it supports continuous improvement by highlighting areas where the model may need further refinement or retraining.

In support of this, there should exist a clear documentation of the data used for performance evaluation, including data acquisition, cleaning and transformation, and processes for managing missing or erroneous data.

Table 4 outlines relevant aspects to disclose to ensure transparency regarding the estimated performance of an AI solution. For further elaboration, see the Chapter on Validity & Robustness.

Table 4: Relevant aspects to disclose to ensure transparency regarding the estimated performance of an artificial intelligence solution

Source: CIOMS Working Group XIV

Scope of evaluation	Describe the nature of the reference sets used for performance evaluation, acknowledging any known deviations from the intended deployment domain (e.g. over- or under-representation of certain drugs, adverse events, patient populations etc.). Relevant information would include the types of data and from where they have been derived.
Sampling	Describe the prevalence of positive and negative controls in the reference set and how this relates to the intended use. If they are different, describe how performance evaluation was adjusted to account for this. Describe any use of data augmentation for performance evaluation.
Reference standard	Disclose the definitions of different categories of classification used in performance evaluation (for example, positive and negative controls in a binary classification task). Share any annotation guidelines used to improve quality and consistency of human annotations in developing the reference standard.
Human input	Describe the qualifications of human assessors contributing to test set development and any use of parallel annotations and evaluations of concordance during this phase. If the AI solution includes a human-in-the-loop during operation, then state the qualifications of those individuals who participated during performance evaluation.
Summary metrics	Present standard performance evaluation metrics when suitable or motivate the use of customized metrics. Place emphasis in this presentation on levels of the decision threshold relevant to the intended use and deployment domain (e.g. with a realistic balance between false positives and false negatives). Complement composite performance metrics with their components (e.g. precision & recall for an F-score).
Benchmarks	Present comparisons against relevant benchmark methods (including human-level performance) and/or standard benchmark reference sets, when available.
Subsets & sensitivity analyses	Present the results of any subset or sensitivity analyses during performance evaluation or acknowledge the lack thereof.
Qualitative review	Provide representative examples of correct classifications and representative examples of incorrect classifications (false positives and false negatives).

1654

References

[205] International Organization for Standardization (ISO). ISO/IEC TR 29119-11:2020. Software and systems engineering — Software testing. Part 11: Guidelines on the testing of AI-based systems. (Abstract accessed 21 March 2025)

[206] National Institute of Standards and Technology. AI RMF Trustworthy & Responsible AI Resource Center. Gaithersburg, MD: National Institute of Standards and Technology. (Webpage accessed 21 March 2025)

[207] Organisation for Economic Co-operation and Development. Transparency and Explainability (Principle 1.3). OECD AI Policy Observatory. (Webpage accessed 21 March 2025)

[208] Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, Gish P, Lardieri A, Wu E, De S, Naidoo J. Information visualization platform for postmarket surveillance decision support. Drug Safety. 2020;Sep;43:905-915. (Journal abstract) https://doi.org/10.1007/s40264-020-00945-0

[209] Botsis T, Dang O, Kreimeyer K, Spiker J, De S, Ball R. A Decision-Support Platform Powered by AI and Humans-in-the-Loop Boosts Efficiency and Assures Quality in FDA's Pharmacovigilance. International Society of Pharmacovigilance, 23rd Annual Meeting 2024, Montreal, Canada. (Webpage accessed 21 March 2025)

[210] Hauben M. Artificial intelligence in pharmacovigilance: Do we need explainability? Pharmacoepidemiol Drug Saf. 2022;Dec31(12):1311-1316. (Journal full text) https://doi: 10.1002/pds.5501.

[211] Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion. 2023;Aug1;96:156-191. (Journal full text) https://doi.org/10.1016/j.inffus.2023.03.008.

[212] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Internet]. arXiv:1602.04938 2016. (Journal full text) https://doi.org/10.48550/arXiv.1602.04938

[213] Hauben M. Artificial intelligence in pharmacovigilance: Do we need explainability?. Pharmacoepidemiology and Drug Safety. 2022;Dec;31(12):1311-1316. (Journal full text) https://doi.org/10.1002/pds.5501

[214] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;Nov1;3(11):e745-50. (Journal full text) https://doi.org/10.1016/S2589-7500(21)00208-9

[215] Reddy S. Explainability and artificial intelligence in medicine. Lancet Digit Health. 2022;Apr;4(4):e214–215. Cited by: Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. npj Digit Med. 2020;Mar;26;3(1):1-5. (Journal full text) https://doi.org/10.1038/s41746-020-0254-2

[216] Pinheiro LC, Kurz X. Artificial intelligence in pharmacovigilance: a regulatory perspective on explainability. Pharmacoepidemiol Drug Saf. 2022;Dec1;31(12):1308-1310. (Journal full text)

https://doi:10.1002/pds.5524

[217] Royal Society. Explainable AI. Royalsociety.org. 2024. (Website accessed 11 August 2024)

[218] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;Nov1;3(11):e745-50. (Journal full text) https://doi.org/10.1016/S2589-7500(21)00208-9

[219] Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review. 2019;Nov22;1(2):1-9. (Journal full text) https://doi.org/10.1162/99608f92.5a8a3a3d

[220] Royal Society. Explainable AI. Royalsociety.org. 2024. (Website accessed 11 August 2024)

[221] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association. 2012;Jan1;19(1):121-127. (Journal full text) https://doi.org/10.1136/amiajnl-2011-000089

[222] Hauben M. Artificial intelligence in pharmacovigilance: Do we need explainability?. Pharmacoepidemiology and Drug Safety. 2022;Dec;31(12):1311-1316. (Journal full text) https://doi.org/10.1002/pds.5501

[223] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Internet]. arXiv:1602.04938 2016. (Journal full text) https://doi.org/10.48550/arXiv.1602.04938

Chapter 7: Data privacy

Principle

Data privacy refers to the fundamental right of an individual to control how their personal information is collected, stored, shared, and used. It is an aspect of the principle of "respect for persons" that is foundational to the conduct of biomedical research. Regulations, legislation and guidance documents provide measures intended to preserve the confidentiality, anonymity, autonomy and control of sensitive and potentially personally identifiable health data in the setting of PV.

Key messages

- The ethical framework to evaluate the use of data privacy protected applications of AI in PV is embedded within the standard principles for research activities involving human subjects.
- The use of certain AI applications in PV requires additional attention to assure data privacy.
- The applications of ethical principles most relevant for the use of AI in routine PV are data privacy, fairness, and equity.
- PV professionals should recognize that existing procedures used to assure regulatory compliance may need to be re-evaluated due to the heightened risks of GenAI to compromise data privacy and for ML to amplify biases.

Introduction

Although data privacy has been recognized as an implicit legal right for well over a century,[224] it was not until the 1970's that this topic began to receive formal international attention. Advances in computer technology began to facilitate the large-scale collection, organization, and evaluation of amounts of data that had previously relied upon paperwork. In the absence of any laws regulating how public bodies could collect, store, or share personal data, the first data privacy law was passed in 1980.[225] In the US, public concerns about the potential misuse of collected data led to the US Privacy Act (1974),[226] which provided boundaries for the collection, integrity, and use of personal data. These same issues raised concerns about transfer of large amounts of personal data across borders, which led to the first international guidelines to protect data privacy in the context of international trade.[227] Similar to this CIOMS Working Group report, the OECD guidelines laid out a set of core principles; however, its intent was to assist governments, business and consumer representatives with the objective of supporting data transfer to facilitate commerce while protecting personal data privacy. Each of these documents were framed as legal responses to technological threats, and did not explicitly highlight the ethical foundations of data privacy. Over subsequent decades, the guidelines have influenced most subsequent data protection regulations/laws, such as the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule 1996 and General Data Protection Regulation (GDPR) 2016, both of which are discussed later in this chapter. As noted in Appendix 2, considerations to protect data privacy are specifically identified in a survey of recent major national and international reports on the use of AI, generally and in pharmaceutical development.

Ethical considerations

While data privacy concerns are widely recognized in the use of AI, at the time of this publication, the authors are unaware of a standard reference focusing specifically upon the ethical considerations in the use of AI applied to PV. Many publications that refer to ethics and AI, such as the WHO Guidance, Ethics and Governance of Artificial Intelligence for Health,[228] emphasize several basic principles that were first elaborated in the Belmont Report (1979).[229] That report was federally commissioned (USA) to determine basic ethical principles that are foundational to biomedical and behavioural research

involving human participants. These principles underlying modern human research protection bear upon clinical research as well as certain post-marketing PV activities globally, recognizing that some PV activities are not technically considered research.

The Belmont Report identified three basic principles that are foundational to interventional and behavioural research involving human participants: respect for persons; beneficence; and justice.

Respect for Persons refers to the obligation for the research subject to enter research knowingly and of their own free will. It also considers that some individuals, such as those who are cognitively impaired, may not be capable of making this decision independently and based on self-determination. From a practical perspective, respect for persons is immediately recognized by the use of informed consent documents, which require that research subjects be advised that they are being invited to participate in a research study, that they are made aware of the purpose of the study, including its potential risks and benefits, and that the information be transmitted so that it is comprehensible to the participant. Participants must be able to freely choose whether to participate, including the option to exit the research project. Participants cannot be unduly influenced or coerced into participation. Informed consent is waived for routine PV activities through statutes recognizing the pre-eminence of societal/public health interest.

The principle of Beneficence describes the need for research to be designed to maximize its potential benefits while minimizing potential harms. Research in which the harms of participation are known to outweigh potential benefits would not be justifiable. This concept also captures the core tenet of the Hippocratic Oath to “do no harm”. For those involved in clinical trials, equipoise is a familiar concept, based on the premise that there is genuine uncertainty about which treatment arm in a clinical trial is best with respect to safety and/or efficacy. It captures the concept the research is not frivolous, including that potential risks have been minimized. PV incorporates beneficence into ongoing benefit/risk assessments. It should be acknowledged that there are applications of AI in PV that do not contain personal information, and these activities are outside the scope of data privacy discussed in this chapter.

Finally, Justice refers to the ethical obligation to see that research is conducted among those who might benefit and that involvement in research is sensitive to ensuring that certain populations, e.g. prisoners, are not preferentially selected for research out of convenience, and that the burdens and benefits of research should be born equally among those who might benefit.

The intent of the Belmont Report was to address considerations in medical research, intended to advance generalizable knowledge, and draw distinctions with the distinct responsibilities of medical practice, intended to support individual patient care. The report did not discuss public health activities, but its key principles are used in public health activities ranging from disease surveillance to PV. Certain public health activities are mandated by statute, including PV. Drug safety includes monitoring safety during clinical trials, post-marketing surveillance, and post-approval safety studies (PASS). Clinical trials are defined as research and are a focus of the Belmont Report. PASS is a form of real-world evidence, often required as a condition of product licensure (RWE is the focus of CIOMS Working Group XIII).

The ethical implications for the use of AI in PV for approved products that are most pertinent include data privacy (derived from the principle of Respect for Persons) and fairness and equity (derived primarily from that of Justice). Respect for persons indicates that every individual has the right to control information about themselves. In the context of PV, Justice captures the concept of fairness, meaning that the benefits of PV knowledge should be equitably distributed among those who may use specific medicinal products, i.e. lack of discrimination. Fairness as applied to PV means that the activity is conducted in a non-discriminatory manner, ideally so that the methods support the representativeness of the population being evaluated and equitable to provide insight into the population that may be exposed to the product. In the context of PV, the principle of Justice is

applied through efforts to assure Fairness and Equity, which is the subject of a separate chapter in this report.

Data privacy regulations

Many countries have established laws to protect the data privacy rights of the individual. These laws share the common principle that personal data requires protection, and that this should be accomplished through mechanisms that mitigate risk to the individual while requiring accountability of the entity using the data. Two of the most frequently cited are the HIPAA, 1996, used in the United States, and the GDPR, 2016, which is employed in the European Union. These examples will be used to illustrate commonalities and differences between data privacy regulations, and their implications for the application of AI to PV.

Example: Health Insurance Portability and Accountability Act

As the name suggests, HIPAA (1996) originally focused on health insurance data[230] and was developed to ensure data privacy as medical information moved from analog to digital. At the same time, the legislation was intended to advance common administrative standards for health care data. In short order, the rapid adoption of digital technologies in health care (e.g. electronic health records) and the interest in using electronic data for research and other purposes led to follow on legislation to support the use of electronic health records according to standards that would ensure administrative efficiency while protecting patient privacy and security (HIPAA Privacy Rule, 2000; Security Rule, 2003). The HIPAA Privacy rule defined individually identifiable health information (Protected Health Information – "PHI") and defined safeguards to protect the privacy of this information. HIPAA also delineates Business Associate Agreements, requiring covered entities that handle PHI to comply with its privacy and security rules.

HIPAA emphasizes the confidentiality, integrity and availability of health data, and includes provisions focused on the "minimum data necessary standard", i.e. limiting data to those necessary for the specific purpose. It specifies patients' rights to access and amend medical records. To protect patient confidentiality, HIPAA recognizes types of data that could be used to identify individuals and specifies 18 unique PHI identifiers. The list underscores the range of common data types that are largely unrelated to health care and which contain identifiable information that could compromise patient identity: name(s), geographic subdivisions smaller than a state, dates (except year, e.g. date of birth), telephone numbers, fax numbers, email addresses, social security numbers, medical record numbers, health plan beneficiary number, account numbers, certificate/license numbers, vehicle identifiers, device identifiers, web URLs, internet protocol (IP) addresses, biometric identifiers (e.g. fingerprints); full face photographs, as well as any other unique identifier that could be used to trace the identify of an individual. Once these identifiers are stripped from a source record, the record can be used without restrictions imposed by HIPAA as the record no longer contains PHI.
Public health often balances societal interest with personal rights. Based on overriding societal needs for the safety, effectiveness, and quality of medicinal products licensed for use in the US, routine PV activities conducted by license holders are typically exempt from some HIPAA requirements for patient authorization to disclose and use PHI. Medicinal products are governed in the US by Food and Drug Administration (FDA) regulations that require monitoring the quality and safety of FDA-regulated products, which is conducted in part through adverse event reporting, product tracking, recalls, and post-marketing surveillance. While some PV activities are exempt from HIPAA, data privacy protections remain, including: use of the minimum necessary data standard (collecting only data essential to fulfil the PV responsibility), de-identification and/or anonymization of data (employed where possible); use of technical, administrative, and physical safeguards to prevent unauthorized access, use, and disclosure); and safeguarding by Business Associate Agreements (where vendors or partners are engaged). Within the US, the FDA and Centers for Disease Control and Prevention (CDC, Atlanta) have complementary responsibilities to support and advance health. Among its responsibilities, the FDA is responsible for "protecting the public health by assuring the

safety, efficacy, and security of human (and veterinary drugs), biological products", and medical devices. The CDC responsibilities include protecting "America from health, safety and security threats, both foreign and in the U.S.", including those associated with domestic and international diseases and chronic or acute diseases. This is accomplished in part by conducting "critical science and providing health information that protects (the US) against expensive and dangerous health threats". To fulfil their responsibilities, public health authorities have a somewhat broader remit than the private sector and are able to conduct their responsibilities and use PHI without patient authorization; however, they implement data privacy safeguards both within their organizations and when collaborating with others.

In contrast to public health authorities and the private sector, academic involvement in PV is generally conducted as a research activity (e.g. PASS), and are subject to different oversight, including use of institutional review boards (IRBs, aka ethical review boards) to assure that studies meet appropriate ethical standards, and are conducted with mitigation for data use and privacy, data use agreements, where applicable with other organizations the use of de-identified and limited data sets, and compliance with both HIPAA and the Common Rule.

Example: General Data Protection Regulation

The GDPR (Regulation [EU] 2016/679) is a comprehensive set of rules overseeing personal data protection in the European Union and succeeds the earlier Data Protection Directive (Directive 95/46/EC), which was issued contemporaneously with HIPPA, at the dawn of the internet age. The scope of the GDPR is much broader than HIPAA as it pertains to the use of personal data affecting all manner of human interaction, including processing by automated means as well, and stems from the 1950 European Convention on Human Rights: "Everyone has the right to respect for his private and family life, his home and his correspondence".[231]

The GDPR incorporates principles such as lawfulness, fairness, transparency, accuracy and integrity, purpose limitation, data minimization, confidentiality, and storage limitation. Compliance is a major feature of the GDPR with organizations such as pharmaceutical companies required to have a Data Protection Officer responsible for overseeing compliance. Penalties for non-compliance are significantly greater than those under HIPAA, with fines up to 4% of global turnover. To offer practical examples, based on the requirements and obligations under the GDPR, four risk-based categories may be differentiated:

1. Prohibited use (e.g. genetic data that might be used for the purpose of drug development would require explicit consent from the patient/participant); it is therefore prohibited without such consent;

2. Restricted use (e.g. the use of biometric data to identify employees for security purposes only, in line with necessity, proportionality and other requirements;

3. Permitted use with safeguards (e.g. use of purchasing history to create targeted advertisements would be typically be permissible - provided that efforts were made to protect customer identity, customers would need to consent and have the opportunity to opt-out, and efforts were made to identify and mitigate and risks to customer privacy); and

4. General permitted use (e.g. processing customer information for online billing purposes, provided that the data are limited to those needed to fulfil the transaction, and that security measures are taking to prevent unauthorized access).

Additionally, several safeguard measures may be used, such as data encryption (preventing access without a decryption key), pseudonymisation (replacing identifiable information with pseudonyms to mask identity), and use of Data Protection Impact Assessments to identify and mitigate risks in data

processing to protect the individual. In contrast to HIPAA, GDPR incorporates a "right to be forgotten", permitting individuals to request deletion of their personal data. In the case of special categories of personal data, such as health data, explicit consent may be required for data processing under GDPR and, where collected, such consent may be revocable.

Similar to HIPAA, the rules of the GDPR allow for pharmaceutical companies to meet their legal obligations to conduct PV activities, monitor and report adverse events without consent in order to ensure oversight of the safety and effectiveness of medicinal products – provided that certain safeguards are in place. These responsibilities may limit data protection rights normally in place under the GDPR, e.g. the "right to be forgotten". Other safeguards include requirements for data minimization as well as administrative, technical and organizational measures to protect personal data.

In fulfilling its responsibilities to assure the safety, effectiveness, and quality of medicinal products authorized for use in the European Union, the European Medicines Agency is empowered to assure PV oversight in a manner that acknowledges that certain data protection rights, such as the right to be forgotten, may be limited for specific PV activities. The EMA emphasizes the principles of data minimization, purpose limitation, lawfulness, fairness and transparency in its data use. In contrast to HIPAA, the GDPR has special provisions for international data transfers, imposing restrictions in exporting data collected for EU citizens (regardless of domicile) outside the European Economic Area and applies safeguards to provide an appropriate level of data protection.

Data privacy expectations for PV research conducted by academia in the EU are analogous to those for the US, with IRB oversight and an emphasis upon adherence to principles of data minimization and purpose limitation. Additionally, international collaborations that involve data transfers outside of the EEA require safeguards that typically include contractual language to assure compliance with GDPR rules.

Other data privacy regulations

Although the FDA and EMA data privacy regulations are currently the most widely followed, it should be noted that there are an increasing number of country-specific differences, which pose particular challenges for the use of multinational AI model development using secondary data. Comparison of regulations in place in Germany, China, and Japan illustrate this point.

Table 5: Data privacy regulations for using secondary data in Germany
Source: CIOMS Working Group XIV

Aspect	Germany
Governing Law	General Data Protection Regulation (GDPR) (EU-wide), Federal Data Protection Act (BDSG) (Germany)
Health Data Classification	"Special category data" (Art. 9 GDPR)
Consent Requirements (Research & PV)	Usually required; exceptions for public interest (e.g. PV, RWE)
Secondary Use of Data (e.g. RWE)	Allowed if legal basis exists (public health, scientific research, etc.) with safeguards
De-identification Standards	Pseudonymization encouraged; full anonymization for broader reuse
Cross-border Data Transfer	Allowed to countries with adequacy or with SCCs/BCRs
Pharmacovigilance Exemptions	Explicitly exempt from consent under public health/legal obligation

Regulator Guidance on Biomedical Use	Extensive EMA and national ethics bodies guidance
Oversight Body	German DPAs + European Data Protection Board (EDPB)
Key References	GDPR (Regulation EU 2016/679); EDPB Guidelines 03/2020; EMA Module VI (GVP); BDSG (Germany)

Table 6: Data privacy regulations for using secondary data in China
Source: CIOMS Working Group XIV

Aspect	**China**
Governing Law	Personal Information Protection Law (PIPL)
Health Data Classification	"Sensitive personal information"
Consent Requirements (Research & PV)	Explicit consent generally required; strict interpretation
Secondary Use of Data (e.g. RWE)	Permitted with new consent or proper anonymization
De-identification Standards	Anonymization required to avoid consent; "irreversible" standard
Cross-border Data Transfer	Strict rules: security assessments, contracts, individual consent; limited adequacy
Pharmacovigilance Exemptions	AE reporting permitted but must minimize identifiable data
Regulator Guidance on Biomedical Use	PIPL + draft health data governance rules; evolving
Oversight Body	Cyberspace Administration of China (CAC)
Key References	PIPL (2021); CAC draft regulations on health data; State Council health data measures (2022)

Table 7: Data privacy regulations for using secondary data in Japan
Source: CIOMS Working Group XIV

Aspect	**Japan**
Governing Law	Act on the Protection of Personal Information (APPI)
Health Data Classification	"Special care-required personal information"
Consent Requirements (Research & PV)	Consent generally required, but pseudonymized data may be used for public interest or research
Secondary Use of Data (e.g. RWE)	Allowed with pseudonymization/anonymization and research purpose declaration
De-identification Standards	Recognizes both anonymized and pseudonymized data; latter still regulated
Cross-border Data Transfer	Permitted to "adequate" countries (EU, UK); otherwise, consent or contracts needed

Pharmacovigilance Exemptions	AE reporting allowed without consent under regulatory mandate
Regulator Guidance on Biomedical Use	MHLW guidance on clinical research and PV under APPI
Oversight Body	Personal Information Protection Commission (PPC)
Key References	APPI (2020 amendment); PPC Guidelines; MHLW guidance on GPSP and human research ethics

Practical considerations to support data privacy

As these examples indicate, regulations have been developed to assure appropriate data privacy within the framework required to conduct routine PV activities. The list of 18 unique identifiers enumerated by HIPAA highlights the breadth of the types of data that can be used to identify individuals. In the years following the introduction of HIPAA and the GDPR, there has been recognition that additional measures may be required to anonymize data.

As a regulated industry, pharmaceutical companies must comply with the data privacy and reporting requirements of all countries in which their products are licensed. As an example, the EMA requires adherence to Good Pharmacovigilance Practices (GVP) and to data protection principles from the GDPR. Ensuring compliance requires attention to evolving country-specific regulations as well as the oversight of vendors that support companies (in some cases conducting certain PV activities for individual companies) as well as business partnerships, e.g. where a combination therapy is co-developed by more than one company. Regulatory authorities may have different requirements for reporting patient information, necessitating some customization and additional oversight to assure adherence to local requirements. For example, Australia requires reporting of ethnicity (to support fairness/equity), while it is prohibited in France out of concerns of discrimination.

To support compliance with global data privacy requirements, contractual arrangements with third parties (e.g. vendors, partners) include privacy-specific provisions and language. In the US, contractual arrangements with vendors/partners require Business Associate Agreements. Under GDPR, binding corporate rules (BCRs) may be implemented to enable multinational companies to move personal data within their companies across borders; BCRs are legally binding and require approval from EU authorities. In the EU, an additional layer of oversight is imposed through the required use of in-house data privacy officers for certain businesses such as pharmaceutical companies. Globally, there are a range of potential consequences for data breaches, from requirements for notification to data protection authorities up to and including significant fines and penalties

In the EU, an additional layer of oversight is imposed through the required use of in-house data protection officers for certain businesses such as pharmaceutical companies. In many countries, including in the European Economic Area, there are penalties for data breaches.

Risks to maintaining data privacy as artificial intelligence is employed in pharmacovigilance

One of the promises of AI is that it will permit more efficient processing of large amounts of routine PV data, e.g. individual case reports. Additionally, large language models (LLMs), whether open or closed, permit nearly instantaneous planned (or unplanned) linking of data sources that would otherwise not have occurred, or have been difficult to accomplish. As discussed below, these risks are substantively greater for open vs closed LLMs. GenAI models are also useful for extrapolation – finding patterns that might otherwise not have been recognized. These attributes raise the question of whether current data privacy tools are sufficient to prevent re-identification of deidentified data.

Adequacy of de-identification measures

In 1990 (six years prior to HIPAA) a US researcher used census data to identify 87% of the US population based on three readily available data elements: five-digit mailing (zip) code, gender and date of birth, illustrating that few data points were needed to uniquely identify individuals.[232] Though mailing codes were subsequently classified as PHI under HIPAA, the point remains that just a few generally accessible data points may be needed to compromise data privacy.

A study using data from a children's hospital in Ontario, Canada, demonstrated that the risk of re-identification of individuals based upon de-identified pharmacy data could be minimized, or even eliminated, by reducing the precision of values in selected data elements, such as replacing the admission and discharge dates with the quarter and year of admission. However, the maximum amount of acceptable generalization in the data element values must be determined by formally examining not only the risk of re-identification and breach of patient privacy but also the intended analysis, which may not be conducted without the appropriate level of precision.[233]

The EMA and Health Canada now require public sharing of clinical trial reports as part of the drug approval process. Standards for data anonymization have been issued. Applying these standards, researchers evaluated the risk of re-identification associated with a clinical study report for a nonsteroidal anti-inflammatory drug, grading suspected cases based on the likelihood of accurate matching.[234] The authors found six suspected matches out of 500 reviewed cases and observed that identifying the matches was time-consuming (24.2 hours per case). Matching was best informed by social media and death records. Based on the 0.09 probability risk threshold of re-identification established by EMA[235] and accepted by Health Canada, the authors concluded that existing anonymization guidance was sufficient to provide an adequate level of data protection and advised review the mechanisms by which re-identification had occurred. With rapid advances in AI, the time required to replicate the reidentification exercise reported in that study (published in 2020) will likely have decreased by the date of this report, and will continue to do so.

Risks of data breaches

The potential consequences of re-identification of de-identified data are amplified by numerous examples of data breaches that have occurred throughout the world. A few examples illustrate the breadth of some recent data breaches, along with their potential consequences:

- A purposeful attack on a US financial firm leading to access of more than 100 million customer accounts and credit card applications;[236]
- An apparently politically motivated international attack by a foreign government on a credit reporting agency in the US resulting in the release of names, birth dates, and social security numbers of nearly half the US population, purportedly with the intent of using AI to compromise US government officials;[237]
- A purposeful domestic data breach intended to embarrass a political opponent, involving a cyber-attack on an Asian health care plan result in 1.5 million patient records;[238]
- An unintentional release of Indian government biometric, and other personal data, in a database containing records of 1.2 billion individuals.[239] In this instance, a criminal group exploited the data breach and offered individual patient records for sale. Approximately 100,000 persons are known to have had their data accessed.

Each example occurred before the widespread use of generative AI, a technology that has been advancing rapidly, and which has the potential to efficiently link publicly available data sources with those obtained maliciously, leading to enhanced risk for re-identification and compromising data privacy.

Individual responsibility to protect personal data

In addition to processes to ensure data privacy to conform to data privacy regulations, individuals play a role in protecting their own data. This responsibility grows in importance with the ever-

increasing number of digital tools (e.g. Smartphones, wearables), apps, and software (e.g. AI-assisted translation tools, GenAI) that provide opportunities for individuals to disclose personal data that may not be sufficiently protected. In many instances, there are legal requirements to support an individual's data privacy (e.g. through the GDPR); however, there remain opportunities for lapses in data privacy, and these are particularly worrisome in the use of GenAI. Common mechanisms to advise persons of data use policies may include terms of use (e.g. End User Licensing Agreements – aka EULA), data privacy notices, and, in some instances, the requirement of explicit consent for use of personal data. Individuals may share personal data (e.g. names, phone numbers), when submitting queries without understanding the consequences of disclosing such information. In many instances, notably those using open GenAI tools, these data may no longer be private. Individuals may also share context-specific information, such as a recent illness (or as in a noted example, a motor vehicle accident) that might be used to identify them.[240] Users may also unintentionally provide personal identifying information by sharing (e.g. in social media) context-specific data outputted by GenAI. These data may subsequently be leveraged to identify the individual even if personal data was not directly entered depending upon the data privacy policies of the respective platform.

Potential risks to data privacy using large language models in pharmacovigilance, and approaches for mitigation

In principle, existing data privacy regulations, (or legislation, such as the GDPR), should provide the basis for protection of data used in AI applications to PV. Model development may require the use of PV data (such as ICSRs) to data scientists and machine learning engineers for training as well as execution. All involved parties, which may include both pharmaceutical companies and vendors, may have access to data that is protected, creating the risk for exposure to larger groups. All parties should be aware of data privacy requirements. The risk is potentially greater with LLMs, as the underlying mechanism of these models provides the potential for some re-identification that would otherwise be unlikely. Those organizations that maintain closed LLMs exercise control of prompts as well as the data contained in the models; in contrast, open LLM models do not have this safeguard and run a greater risk of re-identification, due to the LLM capability of drawing from data sources that may be opaque and not intrinsically within the purview of data privacy regulations (for example, containing the sort of data described above under Individual responsibility to protect personal data). Leaks may occur through prompts bypass data privacy considerations or through models that are trained on personal data. Additionally, as noted above, re-identification can occur even with presumed de-identified data. PV requires review of potentially identifiable and sensitive information that includes basic demographics (including birth date) associated with sensitive data elements including medical or health information (including medicine and vaccine exposure), ethnicity, race, sexual orientation, genetic information, biometric data, physical characteristics, lifestyle information, etc., requiring heightened safeguarding measures.

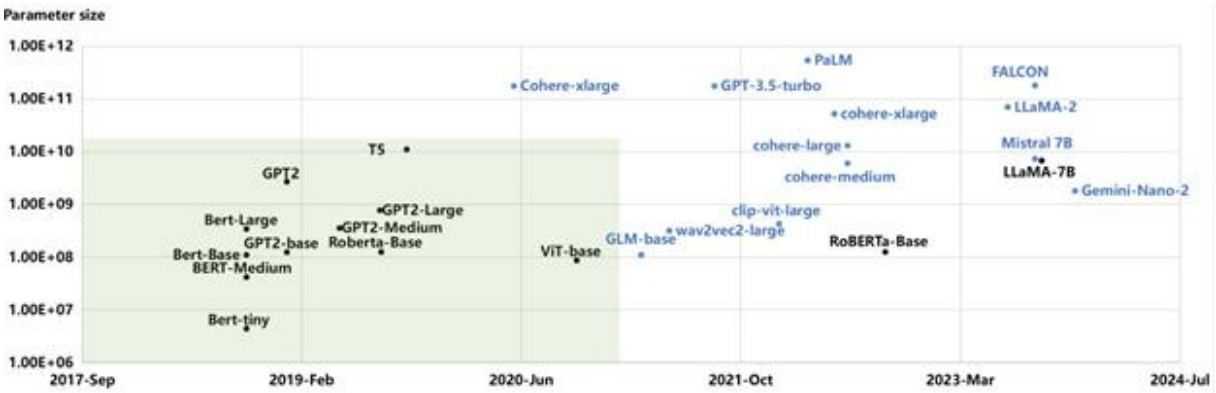
Among the types of challenges posed by GenAI (as well as in some cases ML) for PV are the following.

- Algorithms may be developed within open LLMs, without attentiveness to applicable data privacy requirements, thereby posing a potential privacy risk. If these LLMs are then adopted for use within closed LLMs, there is the potential risk for disclosure of protected information.
- Within a closed LLM, attention should be paid to different sources that may be added to the LLM for unrelated purposes. If genetic data has been collected (with participant consent) for a study and is added to a LLM for a specific analysis, measures would need to be taken to ensure that it is not used for a different purpose outside of the original consent. The accepted practice is to seek consent for additional uses of those data (as the data would now be part of the LLM).
- Generative AI programs can integrate otherwise discrete data sources such as such as census and vital statistics and public health data, which may be linked to a deidentified health record data set (e.g. in the setting of an active PV activity (e.g. a post-authorization safety study) leading to the possibility of reidentification.

These types of risks are amplified in settings where data privacy regulations are lax or poorly enforced. LLMs that are smaller and introduced earlier have tended to have more scrutiny for data privacy than larger and more recent LLMs (see Figure 6). Reasons may include: 1) lack of public availability of newer, larger LLMs; and 2) privacy technologies have struggled to keep up with these newer, larger LLMs.

Figure 6: Relationship of timing of large language model introduction, parameter size and attentiveness to data privacy

Source: [241]



The horizontal axis represents the time of LLMs release, while the vertical axis represents the size of model parameters. Blue dots signify LLM instances not addressed in the literature pertaining to privacy protection, whereas black dots indicate those that have been examined in such literature. The green backdrop delineates the central cluster zone of LLMs with the potential to facilitate privacy protection.

Conclusions

The right to data privacy resides within the well-established framework of basic ethical principles for human research protection articulated in the Belmont Report. National regulatory authorities have provided requirements intended to protect data privacy, indicating the types of data that can be made available, along with safeguards (such as data minimization, anonymization, de-identification and data encryption) along with potential penalties for non-compliance.

Despite data privacy laws and the increasing sophistication of technical measures employed by companies entrusted with personal information, attempted and successful data breaches have been occurring with increasing frequency and often at enormous scale (affecting in some cases >100 million individuals), suggesting both failures in oversight along with technical advances to outwit cybersecurity measures and break into secure data sources.

Ever-increasing computational power, larger linked databases, and the introduction of GenAI, are occurring in parallel with an increasingly globalized PV landscape involving more numerous and complex interdependencies (e.g. business partnerships, international vendors conducting PV activities). The ongoing challenge for PV professionals and regulatory agencies and industry, as well as colleagues and academia will be to assure that within this rapidly evolving data science landscape, data privacy measures are monitored and regularly updated to properly protect personal data.

A potential risk in applying GenAI for PV is patient reidentification, suggesting a need to reconsider the specificity of de-identified data, along with risks associated with open LLMs in which some data sources may be outside the control of the user.

In addition to following existing data privacy regulations, efforts to mitigate risks to data privacy when applying AI to PV may include the following.

- Recognition that the technology is advancing rapidly, requiring ongoing monitoring, e.g. to assure that data de-identification measures are adequate.
- Understanding that open and closed LLMs pose somewhat different challenges to data privacy. Operating closed LLMs in safeguarded environments within institutional firewalls and carefully examining the risks of sharing these models with third parties should be helpful in risk mitigation.
- Audits to evaluate whether only the minimum required personal information is included in reports, that any re-use of data for secondary purposes is consistent with the purposes for which that data was collected and that adequate measures are in place to support compliance with data protection requirements by all entities (e.g. vendors) contributing to PV. Insofar as PV activities may be conducted by a network of collaborating organizations, the organization with the weakest oversight of data privacy may present a risk.
- Oversight regarding access to LLMs for PV practices to assure that use by trained PV professionals is fit for purpose.

References

[224] *Warren and Brandeis. "The Right to Privacy". Harvard Law Review. Vol. IV. December 15, 1890, No. 5. (Webpage accessed 27 April 2025))*

[225] *Gesetz zum Schutz personenbezogener Daten (Hessisches Datenschutzgesetz – HDSG), Gesetz- und Verordnungsblatt für das Land Hessen I, 1970, S. 61.*

[226] *Privacy Act of 1974, Pub. L. No. 93-579, § 552a, 88 Stat. 1896 (1974) (codified as amended at 5 U.S.C. § 552a).*

[227] *OECD, Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, OECD Doc. C(80)58/FINAL (Sept. 23, 1980.)*

[228] *World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. (Full text accessed 25 March 2025)*

[229] *U.S. Department of Health & Human Services. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Washington, D.C.: U.S. Department of Health & Human Services. (Full text accessed 21 March 2025)*

[230] *U.S. Department of Health & Human Services. Health Information Privacy. HIPAA for Professionals.(Webpage accessed 21 March 2025)*

[231] *GDPR.eu. What is GDPR, the EU's new data protection law? (Webpage accessed 21 March 2025)*

[232] *Sweeney L. Simple demographics often identify people uniquely. Carnegie Mellon University, Data Privacy Working Paper. Health (San Francisco). 2000;Jan1;671:1-34. (Full text accessed 21 March 2025)*

[233] *El Emam K, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the risk of re-identification of patients from hospital prescription records. The Canadian Journal of Hospital Pharmacy. 2009;Jul;62(4):307. (Journal full text) https://doi:10.4212/cjhp.v62i4.812*

[234] *Branson J, Good N, Chen JW, Monge W, Probst C, El Emam K. Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations. Trials. 2020;Dec;21:1-9.(Journal full text) https://doi.org/10.1186/s13063-020-4120-y*

[235] *European Medicines Agency, Guidotti T. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. 2018. (Webpage accessed 21 March 2025)*

[236] *Starks T. U.S. charges Chinese military hackers with massive Equifax breach. CNN. (Article accessed 21 March 2025)*

[237] *Geller E. U.S. charges Chinese military hackers with massive Equifax breach. New York Times. 2020. (Article accessed 21 March 2025)*

[238] *The Straits Times. Personal info of 1.5m SingHealth patients, including PM Lee, stolen in Singapore's most serious cyber attack. The Straits Times; 2020. (Article accessed 21 March 2025)*

[239] *Jain M. The Aadhaar Card: Cybersecurity Issues with India's Biometric Experiment. Seattle, WA: Henry M. Jackson School of International Studies, University of Washington; 2019. (Article accessed 21 March 2025)*

[240] *Janmey V, Elkin PL. Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack. AMIA Annu Symp Proc. 2018;Dec5;2018:1329-1337. (Journal full text)*

[241] *Yan B, Li K, Xu M, Dong Y, Zhang Y, Ren Z, Cheng X. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. High-Confidence Computing. 2025;Feb28:100300. (Journal full text) https://doi.org/10.1016/j.hcc.2025.100300*

Chapter 8: Fairness & Equity

Principle

Fairness and equity require awareness of and adherence to impartiality, equality, non-discrimination, diversity, justice, and lawfulness. The benefits of AI in PV should be equitable across all relevant populations and groups. Throughout the AI lifecycle, it is important to avoid and mitigate unfair bias, and any discriminatory practices and unjust social wellbeing and environmental impacts.

Key Messages

- Consider the development and application of AI impacting fairness and equity, whose lack or imbalance may result in discriminatory harm to subpopulations underserved by an AI solution, explicit biases resulting in negative impact, or impact performance by providing inaccurate results.
- Plan and implement mitigation strategies where possible for areas that bias may be introduced reducing potential underperformance; avoid discriminatory harm to underserved populations.
- Equity may be advanced by taking measures (e.g. assess for representative data sets) to assure that AI applications to PV result in outputs (e.g. assessments, aggregated data outputs used for product safety assessments, etc.) that are relevant to populations anticipated to have exposure to the specific medicinal product being evaluated.
- Screening and identifying explicit or potential bias when possible is key to implementing mitigation measures to reduce risk, determining AI applicability and limitations, and establishing expected performance acceptance criteria.
- Scrutinize training and performance evaluation reference data sets for adequate representation and evaluate performance in relevant subgroups when possible. Inadequate reference data is often the cause of inadequate fairness and equity.
- There is limited fairness and equity in ICSRs, with some countries reporting significantly more than others and providing more contextual data available for analysis, such as real-world data (RWD). Consequently, our understanding of routine usage is often limited among underserved populations.

Introduction: general concepts and considerations

In the context of PV, adherence to established laws and regulations such as privacy laws and PV regulations must remain intact with the introduction of AI. What has changed is the increasing awareness of the need for consideration, governance, and mitigation of potential factors that may influence or impact fairness and equity with the use of AI.

Not all fairness and equity concepts, considerations or negative consequences associated with the use of AI will be uniquely specific to PV. That does not negate the need to address these considerations. Biases within AI solutions is a general problem which may impact performance, and not all forms of statistical bias will result in a negative impact on fairness and equity. Within PV, the focus will be regarding unfair bias introduced through data collection, selection, model development and human involvement in the design, development and use of AI that could potentially result in unfairness, discrimination, or inequality.

This chapter will not address the impact of development and use of AI on justice and lawfulness, on individuals' access to essential services, lack of public resources for financing and implementing AI tools and the required ecosystem, and impact on social well-being, because while these are important issues, they are not unique to PV. While not unique to PV, access to AI and the required ecosystem can be a significant barrier for low- and middle-income countries that can result in inequality and underserved populations. Potential workforce implications with introduction of AI in

PV will not be addressed here as it is discussed in the Chapter on Human Oversight under the Section on Transformation of traditional roles. In addition, the rapid acceleration in the use of GenAI is associated with significantly increased energy demand and environmental consequences, both of which are acknowledged as having a broad societal impact, but which are beyond the scope of this report.[242]

Fairness and equity considerations are challenging and can be influenced by cultural differences, historical inequalities, perceptions, and risk factors may appear subjective. To reduce bias, intentional actions are required throughout the AI lifecycle, from design through implementation, and while in production, to reduce discriminatory risk. There are numerous factors that may introduce bias.

Fairness and equity considerations and pharmacovigilance

Fairness and equity principles are fundamental in identifying and addressing discriminatory biases arising from the use of AI systems. In PV, proactive measures are essential to detect, assess, understand and prevent adverse effects to ensure safe and effective use of medicines. When utilizing AI systems in PV, it is essential to implement proactive strategies to mitigate potential harm caused by high-impact AI systems.

Thorough evaluation and ongoing monitoring are required throughout the AI system lifecycle to identify and quantify potential areas of risk and mechanisms through which bias may be introduced as a first step to define strategies to mitigate discrimination biases arising from the use of AI systems in PV. Monitoring for biases is required from conception, development, testing, and following solution deployment. The frequency of monitoring for bias and appropriate modification needs to be defined based on risk assessment, solution results, and potential external factors that may impact model bias and performance.

It is crucial to acknowledge the possibility of bias that may lead to unfair practices or unequal treatment of patients when using AI in activities related to detecting, collecting, assessing, monitoring, understanding, and preventing adverse effects or any issues related to medicinal products.

The PV professional is responsible for ensuring that the AI solution meets the defined business requirements, supports patient safety activities, and does not introduce bias that may inadvertently place patients at risk, a disadvantaged position or at potential for discrimination, e.g. denied the potential benefits of a medicinal product, through exclusion based on race, gender, age, or socio-economic factors. However, we acknowledge that the current system is not perfect; it is essential to ensure that AI does not exacerbate existing issues, but instead, contributes to improving fairness and equity.

Sources of potential threat to fairness and equity

Inherently, humans are biased and can introduce that bias throughout the AI system lifecycle (e.g. requirements gathering, model training, monitoring may not detect poor performance, incorrect results, or missed scenarios). AI experts and developers can have unconscious bias, and potentially if not identified and addressed, the output can have limitations, be discriminatory and may not be recognized as biased. Conversely, the output could be accurate, fair, and equitable; however, results may be rejected by the human with a bias as being poor performance.

Inadequate training and/or testing data set(s)

Bias is primarily introduced in the data used to develop and test AI solutions, which can perpetuate bias and discrimination resulting in harm. Incorrect conclusions can occur when there are data

limitations such as when it is not a complete dataset or does not represent the population where the AI is being applied. The inappropriate or unintended application of AI to populations not represented can occur if data limitations are not transparent or recognized.

Explicit negative bias

Some of the main risk to fairness and equity is introduced from explicit biases (e.g. a case prioritization algorithm down-prioritising reports for men because of patterns represented in the training set).

The lack of robustness and availability of data, e.g. health records not digitally available globally, can lead to underserved populations or underperforming models. When data representation is inadequate, the available data does not correspond to the population and consideration is required to remediate under-represented groups or lack of available organized data, e.g. regions with less developed PV reporting systems, otherwise scenarios will be biased toward groups represented by the training data, and since the data does not represent all groups, e.g. ethnicities, results in bias against these groups, e.g. minorities.

Inadequate data - whether as a result of data not being available or organized in a usable format or structure, lack of data robustness, or inadequate representation of all variables - may result in an under-performing model, or worse, incorrect conclusions as a result of model limitations not being recognized, and this may negatively impact patients' health outcomes. Imbalance of data representation can potentially skew data, amplify imbalances, and it may be difficult to identify and assess bias when reviewing an AI solution's output.

Historically, there have been examples of bias influencing PV activities because of data limitations such as known under-reporting or stimulated reporting of adverse events, with inadequate data or imbalance of data providing artifact that have had negative impacts. Litigation such as class action lawsuits that are pursuing product liability claims can result in stimulating high volume of reported adverse events during the process of legal firms identifying potential plaintiffs that could overshadow unsolicited reports and the imbalance of data could be a threat to fairness and equity considerations if the data imbalance results in incorrect conclusions with groups that are under-represented as a result of skewed data. Local prescribing practices' impact on adverse event reporting and data availability should be considered. These data biases if introduced into an AI solution will potentially magnify the negative impact and remain undetected with difficult identification of underlying bias.

Underserved groups

Under-representation can directly result in underserved population segment(s) and potentially not recognize nuances of subpopulations. Population-specific segmentation can be done by demographics, disease processes, genetic variability, health practices variability, and cultural differences for medical regimens and patient expectations. Such differences can introduce bias with a negative impact if data is exclusive to a specific group, if data is exclusionary, or if nuances of a subgroup are not understood, such as a case prioritization algorithm that underperforms in reports from certain countries in Asia, because they were under-represented in a training data set and differed in important ways from other countries represented.

During clinical trials, such potential harm may be overlooked if subgroups are under-represented in the study population or receive a lesser level of care, e.g. have limited access to medical professionals or facilities. There may be more focus on preventing false negatives to not miss significant information, e.g. the failure of the PV process to detect potential harm restricted to or over-represented in certain subgroups. In the post-marketing period, deployment of an AI solution working less well in certain patient subpopulations could lead to an inability to detect adverse events from these populations. Conversely, false positives may be of greater concern in duplicate detection

where a higher rate of reports falsely flagged as suspected duplicates in a specific country could lead to missed or delayed safety signals there.

Special populations frequently not represented, such as age related (paediatric, geriatric), pregnant women, and infrequent or under-reported events such as rare diseases, and events with social stigmas need to be considered when assessing bias. In the example of an AI solution implemented to support signal detection activities, with limited data from special populations (e.g. pregnancy), the negative impact would be magnified with misinterpreted or missed signals.

Reliance of decision making on data not representative of respective populations (e.g. post-approval risk minimization activities based on data with limited representation of served population) could result in minimization measures not properly addressing safety of patients in the population. If unable to mitigate lack of representation in AI solution, it may require reliance on historical PV approaches and safety measures (e.g. robust monitoring measures for special populations).

Detailed identification of "groups" that could be disfavoured or low volume events proportionate to data set and comprehensive strategies to address data inadequacies can reduce potential bias, discrimination, and underserved populations.

Artificial intelligence solution design

Algorithms should not perpetuate existing bias or discrimination, and the algorithmic design can lead to unintended consequences. When AI was used to develop a model to predict what patients would benefit from proactive intervention in the care of their chronic illness, its results directed more resources to white patients than black patients, because the data set used for training was based on utilization, not need.[243] Given a healthcare system and a universe of healthcare data that is likely to carry country-specific biases, any naïve use of AI will reproduce these biases in its predictions. The likelihood of adverse consequences is more likely because of the apparent opacity of AI, hype about its capabilities, limited understanding of how it works, and unclear pathways to question its conclusions.

Parameters defined by a programmer and how a model processes data could introduce bias or produce inaccurate results by skewing the result. If a developer selects or designs features for an AI solution based on their own conscious or unconscious bias, the resulting output could be suboptimal or even incorrect. In the case of GenAI prompt engineering development, the potential to introduce bias based on the prompt design, lack of specificity, context, or omission of a required prompt could result in an output with a negative bias. Individual preferences influence decisions and subsequently influence data selection and model development. This could occur due to the model developer having an affinity to subgroups like their own profile (e.g. developer is a young person and may select data that does not account for paediatric or geriatric populations).

The development strategy should have a conscious systematic approach to avoid bias and achieve complete and accurate data representation accounting for diverse groups. Documenting how distinct groups are represented in the training and test data may provide insight to limitations, bias, and potential impact supporting implementing mitigation measures. When considering the population of respective groups, confirmation that the data are representative of the global population is needed to ensure balance, demographic parity, and appropriate distribution and allocation.
AI is increasingly employed in the field of medicine to identify patterns and anomalies, such as consistencies, inconsistencies, and outliers in the identification of safety issues and communications. For example, examining sentiment consistency can help flag and mitigate human-induced discrepancies. This proactive approach reduces the risk of unfairness and bias, enhancing the reliability and objectivity.[244]

Risk, impact, and mitigation measures

The consequences of AI on fairness and equity are dependent upon the application of AI within PV, the usability, performance, and the risk of where the AI is being used within the process. When there is discrimination and bias embedded in the AI model through data limitations and/or algorithm development, the negative impact of the resulting biased model is magnified in its application. The model may amplify or skew outcomes resulting in incorrect conclusions, incorrect introduction of an advantage or disadvantage, inequalities, or discrimination of groups or populations.

Evaluating an AI solution pre and post deployment for explicit or potential bias allows for mitigation measures to reduce risk. AI solution explainability may highlight explicit bias and understanding the profile of training data provides a degree of insight into potential areas where bias may be introduced into the solution, determine appropriate use, solution limitations, degree of human oversight required, and expected performance. When evaluating for bias, consideration should be given to post deployment data annotation processes for future retraining activities and mitigate when possible.

Within PV signalling activities, omitted results could cause misrepresentation of a product benefit/risk profile and have a detrimental impact, leading to incorrect human conclusions or decisions impacting patient safety.

Sensitivity analysis of performance across different subgroups can be important to highlight groups or populations underserved by an AI solution. A risk-based approach when selecting subgroups to evaluate performance may be necessary when an exhaustive sensitivity analysis is not feasible and may be dependent upon data limitations for training and test data for subgroups or populations.

Key mitigation strategies

- Evaluate each AI solution for fairness and equity, outlining the assessment method, results, and any measures taken to mitigate.
- Review training and test data sets thoroughly for completeness and adequate group representation.
- Perform sensitivity analysis when possible, evaluating AI model results for equality by changing subgroups/populations to confirm expected results (e.g. modify gender input and evaluate impact to the output) and highlight potentially underserved populations. This is especially important when an AI solution has a lower level of explainability.
- Review AI solution design, parameters, and feature selection for bias when an AI solution is explainable, and the results are not as expected.
- Ensure training data description is transparent highlighting explicit bias and allow clarity on model limitations to reduce inappropriate application or incorrect conclusions.
- Determine level of human involvement required in development and monitoring activities providing required input to ensure accurate performance and fair results.

Identification of potential risk areas is challenging but key to preventing bias, discrimination, and suboptimal model performance. Avoidance of data limitations is not always possible and providing visibility of data characteristics allows appropriate application and opportunity to mitigate risk. It is important to understand the model limitations and communicate to the user community and group monitoring AI performance of limitations and potential bias.

References

[242] Government of Canada. Guide on the use of generative artificial intelligence. Ottawa: Treasury Board of Canada Secretariat; 2023. (Webpage accessed 21 March 2025)

[243] Ziad Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations.Science366,447-453(2019).DOI:10.1126/science.aax2342

[244] Bergman E, Sherwood K, Forslund M, Arlett P, Westman G (2022) A natural language processing approach towards harmonisation of European medicinal product information. PLoS ONE 17(10): e0275386. https://doi.org/10.1371/journal.pone.0275386

Chapter 9: Governance & Accountability

Governance - Principle

Governance refers to the human management and oversight used to control and direct the use of AI in the PV system. An AI governance framework requires implementation of risk management practices and policies to ensure adherence to the AI guiding principles.

Accountability - Principle

Accountability applies to clearly defined roles, responsibilities and liability for organisations and/or individuals deploying, operating and managing AI systems. It requires the adoption of appropriate governance measures by relevant stakeholders, including but not limited to regulators, vendors, users, developers, data providers or pharmaceutical companies involved in setting policy, developing, deploying and managing AI systems. This ensures operations remain within expected parameters throughout the AI lifecycle while addressing any unforeseen consequences.

Key messages

- Governance requires a comprehensive approach across all lifecycle stages of an AI solution as well as the processes it impacts and should therefore be established as early as possible.
- Accountability requires clearly defined roles and responsibilities for stakeholders involved in AI systems for PV; AI systems themselves cannot be held accountable.
- Systems and processes, along with service providers and software vendors, need to be qualified.
- Regular reviews of AI systems and how they adhere to the AI principles are necessary to ensure ongoing regulatory compliance and performance.
- A governance framework grid for an AI system in PV can serve as a structured guide to help relevant parties to document e.g. assessments, actions and references processes, such as Standard Operating Procedures (SOPs) etc. throughout the lifecycle of the AI system.
- Governance and accountability should be independent of the business' utilization and value proposition of the AI system to facilitate unbiased decision making.

Introduction

Previous chapters have discussed in detail the importance of taking a risk-based approach, providing adequate human oversight, demonstrating validity and robustness, and addressing transparency, data privacy, fairness and equity when integrating and implementing AI solutions into the overall PV system. This chapter outlines the guiding principles of governance and accountability in AI-enhanced PV. We will discuss the importance of these two principles, the stages of the AI lifecycle that require specific governance actions, the roles and responsibilities of various stakeholders, regulatory oversight, and the need for ongoing training in the rapidly evolving field of AI technology.

Robust governance and clear accountability are crucial for the success of AI initiatives. These principles help ensure that AI systems are used responsibly and ethically, are compliant with regulations, while fostering trust and transparency among stakeholders. Clearly defined roles and responsibilities enable all stakeholders to understand their obligations and effectively oversee AI systems.

As AI technology evolves, governance and accountability frameworks will need to be adapted. New risks and challenges will emerge, requiring updated principles and practices. Continuous review and adaptation are essential for staying ahead of these changes. This includes the refinement of the proposed governance framework grid for practical use.

Governance framework

A governance framework grid (referred to as "grid") for AI solutions in PV (see Table 2) is a structured guide designed to identify key considerations to address each of the principles throughout the lifecycle of the AI system, including concept, development, deployment, and monitoring phases of the AI model developed for PV use.

In addition to serving as a structured guide for planning and overseeing AI solutions, the grid can also aid in self-assessment. By detailing where each action or process is recommended, the grid helps ensure that the principles such as transparency, accountability, and a risk-based approach are consistently adhered to, facilitating the integration of AI into PV systems. Regular reviews of KPIs by a governance body, aimed at ensuring adherence to the AI guiding principles, can facilitate identification of gaps and drive improvements in the AI solution. While a governance body with expertise and focus on AI's use in PV is needed in early phases, integration of the governance process into the overall PV system oversight mechanisms should be considered when the AI solutions enter routine use phase. If a risk emerges that warrants significant modification to the AI solution, the AI focused governance body may need to be re-engaged.

Consultation with the grid can occur in multiple ways. A unit within a PV organization may have an idea for AI-based automation and specify governance requirements upfront when commissioning a vendor or internal development team. Alternatively, a vendor might present a ready-made AI solution to a PV organization, which then can be evaluated against the AI guiding principles for example by applying this grid. Early consideration of governance principles is crucial for the successful implementation of an AI system. These principles should guide the development or selection of a vendor solution, deployment, and ongoing management. Early planning should be focused on identifying potential risks and determining mitigation strategies. Furthermore, it can stimulate focus on alignment with ethical and regulatory standards of the AI system from the outset, setting the foundation for a robust and compliant AI system.

The grid is composed of five lifecycle phases of the AI solution: an initial requirement specification phase where business units typically provide input, followed by development, pre-deployment, post-deployment, and routine use. These phases are valid for both initial qualification and iterative changes of the AI solution. In each phase, the AI guiding principles should be considered, and in the grid, each principle constitutes a cell for relevant documentation hereof. When the grid is used for a specific AI solution, each cell is intended to provide information about actions, considerations, or references to where these actions are documented, such as SOPs, working instructions, or repositories containing log files, reviewed performance metrics, or names of accountable persons/review bodies. Illustrations of how each guiding principle is applied throughout the lifecycle phases can be found below, and examples of how to put this grid into practice can be found in Appendix 3: Use cases.

Table 8: Governance framework grid
Source: CIOMS Working Group XIV

	Collection of specifications, requirements	***Development & change management***	***Pre-deployment & post-change sign-off***	***Post-deployment & post-change hyper-care***	***Routine use***	***General considerations***
Risk-based approach	Risk assessment (theoretical) - AI model - Context of use - Impact & likelihood of risks	Risk mitigation plan	Risk assessment (empirical) Adjustments to risk mitigation plan based on performance evaluation	Intensive or targeted monitoring for risk assessment (empirical), target high risk areas Mitigation if needed	Routine monitoring (e.g. risk for model drift) Mitigation if needed	Review and refine risk-based approach at regular intervals

Human oversight	Multidisciplinary expertise HOTL (human-on-the-loop; design)	Define HITL (human-in-the-loop; strategy Change management, including staff training plan	Fine-tune HITL strategy Staff training roll-out	Implement HITL strategy Intensive or targeted intervention	Routine HITL activities Adjust and fine-tune(?) HITL strategy as needed	HIC (human-in-command; holistic oversight) HOTL (general monitoring)
Validity & robustness	Specification of use case & deployment domain Specification of reference standard(s) Specification of benchmarks Requirements on reproducibility	Training & validation Development or acquisition of reference standards	Performance evaluation Benchmark comparisons	Performance monitoring	Continuous integration and deployment Periodic performance monitoring	Special considerations for low-prevalence settings Reproducibility Assessing AI solutions with human-in-the-loop
Transparency	[From organization to developer] Model requirements - Intended use - Human-computer interaction - Explainability - Expected outputs Performance evaluation requirements - Scope - Reference standard	[From developer to organization] Model - Architecture - Parameters - Acceptable inputs - Expected outputs - Standard AI components - Training & validation - Known limitations Explainability in support of model development, debugging, and documentation	[From developer to organization] Performance evaluation - Scope - Sampling - Reference standard - Human input - Summary metrics - Benchmarks - Subsets & sensitivity analyses - Qualitative review Explainability in support of assessing Validity & robustness and Fairness & equity	[From developer to organization and from organization to end user] Performance evaluation - Deviations Explainability in support of assessing Validity & robustness and Fairness & equity	[From organization to end user (and regulatory authorities)] Disclosing use of AI Explainability in support of building trust with end users	
Data privacy	Specification of use -Specification of data sources -Identification of data elements that contain identifiers -Jurisdictions / provenance of data Data privacy by design (data minimization)	Training data set selection, algorithm design	Test set (if publishing is intended, e.g. as a public benchmark, need to assure data privacy consistent with local legislation/regulations/guidance)	Adherence to data privacy considerations in running the models with full / accruing data sets Greater attention to deviations in hypercare	Ongoing processes to identify and rectify data privacy issues in routine use	Data privacy legislation/regulations vary by country/region, potentially leading to inconsistencies (emerging risks of new prompts)

	-Data protection impact assessment					
Fairness & equity	Context of use -acceptable application	Training data set selection, algorithm design and cognitive bias Avoid -Explicit or potential unfair bias -inadequate data inclusion	Pre-deployment performance evaluation -Reference data sets inadequate (e.g. unavailable, inadequate representation) -Algorithm design -Human/cognitive bias		Routine Monitoring -poor performance related to model shift, inadequate training data (underrepresented populations, special populations)	Ensure model and training data description is transparent, and limitations highlighted to reduce inappropriate application or incorrect conclusions
Governance & accountability	Consideration on how AI solution fits into existing PV system Key roles (non-exhaustive examples) - PV experts - AI experts	Agreement on KPIs to support implementation Key roles (non-exhaustive examples) - PV experts - Data scientists - AI experts - IT specialists - Ethics specialists	Refinement of KPIs to support implementation, based on performance evaluation Key roles (non-exhaustive examples) - PV experts - Data scientists - AI experts - Senior management	Approval of risk mitigation strategies Key roles (non-exhaustive examples) - PV experts - AI experts - IT specialists - Data protection officers - Cybersecurity experts	Integration into overall PV quality management system including oversight of KPIs Key roles (non-exhaustive examples) - PV experts - AI experts	

Description of each lifecycle phase included in the grid is presented below:

Collection of specifications, requirements: This is the initial phase where the stakeholders are identified and engaged, and the project's objectives, scope and features are defined. The multidisciplinary team of PV professionals, data scientists, AI/ML engineers, software engineers, IT specialists, and other domain experts (also refer to the chapter on Human Oversight), is typically managed by system developers, software vendors, or an internal IT development team. This phase provides a roadmap for developers and end-users, and lays the foundation for the entire development process. Like traditional software, as an AI solution evolves, the requirement specifications may also require iterations, and consequently, the grid may need to be reconsidered accordingly.

Development & Change Management: In this phase, the multidisciplinary team focuses on acquiring, creating or modifying AI systems, ensuring they are built with the necessary functionality and adherence to governance principles. Whether developing an AI system or selecting a vendor solution, these principles will apply throughout.

Pre-Deployment & Post Change "Sign-Off": At this phase, the AI system transitions from the development stage to deployment into the PV process. Before implementation, a thorough validation and approval process is required to ensure the AI system, or any changes hereof, is ready for deployment. Typically, a PV expert becomes accountable for the results produced by the AI solution and for adapting the processes in which the solution will be used. Documentation of this phase may include risk assessments, review of sufficient adherence to principles, sign-off forms, validation reports, and many references to SOPs detailing the sign-off procedure, etc.

Post-Deployment & Post Change "Hypercare": Following deployment, this phase is critical for the immediate monitoring of the AI model's performance or the latest changes' impact. It is a period of intensive observation to promptly identify and resolve any unanticipated issues, as real-life application of the AI system in the PV process might surface issues due to various reasons such as incorrect assumptions, design flaws, unintended bias, in earlier stages. This phase, dominated by hypercare, differs from traditional software hypercare; for AI solutions, immediate fixes may not be feasible and other measures such as human intervention or increase in human oversight might be needed. Documentation is expected and may include incident logs and performance analysis reports specific to the most recent change while under observation.

Routine: This phase signifies the full integration of the AI solutions into the PV process. It involves ongoing monitoring, maintenance, and documentation to ensure full oversight and allows for the identification of trends through the monitoring of pre-defined KPIs. This phase may reference routine reports, logs of ongoing actions, and which SOP or working instruction manages this review process, reflecting the model's full operational status.

Of note, discoveries during post-deployment or routine use phases may necessitate the AI solution being sent back to pre-deployment for enhancements.

The following, non-exhaustive examples illustrate aspects to consider for each guiding principle in relation to the lifecycle phases in the grid:

Transparency: In the Development phase, there is a focus on creating comprehensive documentation of the development activities including reason for changes and data used in model training. In Pre-Deployment, transparency is further enhanced by adding model performance evaluation, and empirical evidence for fairness and equity. Also, the documentation created should ensure consistent understanding of the intended use among different stakeholders. In routine use, the most important transparency is toward the end-users and those responsible for the continual performance evaluation and monitoring.

Accountability: Throughout all phases, there is a consistent need to assign and document responsibility, whether it is to IT, vendors, or to PV experts. This ensures clarity about who is accountable for the AI model's development, change management, deployment, and performance at any time.

Risk-based approach and human oversight: This begins with identifying the level of risks associated with development of the AI solution. When relevant, it may involve the development of clear annotation guidelines for human domain experts to ensure solid method development and performance evaluation. The next step is to propose appropriate mitigation strategies such as defining "human-in-the-loop" within an AI solution and other oversight measures up to eventually creating risk mitigation requirements in the user interface. It continues with redefining human oversight in Pre-Deployment, and further refining these concepts in the Routine phase based on real-life observations. This sequential approach highlights the need for evolving risk management as the AI model advances through its lifecycle. A risk-based approach in general is recommended for all measures taken to adhere to AI guiding principles.

Any changes to the AI system must undergo the same rigorous governance considerations as the initial deployment. This ensures that modifications do not compromise the system's integrity or performance. Documentation and validation are essential to maintain transparency and accountability. Change management processes should be in place to handle updates and modifications effectively and account for a post-deployment phase, that based on “hypercare”, will confirm performance and quality beyond routine monitoring.
As computer system validation requirements need to be met at the same time, it is advisable to de-couple AI model version control from the rest of the software versioning.

Governance body and accountability assignments

To effectively manage the review and agree on actions and risk assessments towards the different principles, it is advisable to nominate a governance body. This group ideally should be a diverse, cross-functional team that has sufficient awareness of the end-to-end process and the extent of automation within it. It should include representatives from all relevant stakeholders and representation from the software vendor may also be considered. This diversity ensures a broad and balanced review of the AI solution. The governance body oversees the development, deployment, and ongoing management of the AI system to ensure that all actions align with guiding principles and regulatory standards. The governance body also determines accountable persons for the respective lifecycle phases, which includes sign-off of the documentation prior to deployment of the AI system into the PV process. Because business cases are often drivers of AI initiatives, the governance body should also include the respective project managers or sponsors to ensure adequate resourcing of governance measures during each phase of the lifecycle.

Unlike traditional software, the governance body of an AI system should review the adherence to the AI use guiding principles in defined intervals, and ad-hoc if needed, to ensure the assessments are still valid. This is due to the rapid evolution of the field and the inherent risks of AI models that changing inputs, rules or other unforeseen issues may disrupt the solution at varying degrees, some significantly. The appropriate frequency and scope of reassessment of a deployed AI solution should be assessed. There should be measures ready to intervene or even disable the AI solution if necessary. Once the AI solution reaches the routine use phase, governance can be handed off to process owner to be integrated in the overall PV system monitoring process. Nevertheless, if a risk emerges that warrants significant modification to the AI solution, the AI focused governance body may need to be re-engaged. The introduction of version control for the governance framework grid should also be considered.

Traceability and version control

Traceability and version control are crucial aspects of managing AI solutions, particularly in a regulated field like PV where errors could impact patient safety or public health. They can enable evaluation and reproducibility of earlier versions of an AI solution and are often required for audit purposes (however, see also the discussion of AI solutions with stochastic components in the Chapter on Validity & Robustness). General best practices from existing version control frameworks can offer orientation for the version control of AI models, which should be documented alongside other relevant systems involved in the end-to-end process. They should include clear change control processes within both a user acceptance testing environment and the production environment.

Documentation of an AI model should comprise its entire lifecycle, and may cover the justification, initial scoping and conception, development, deployment, validation, and post-deployment. It should allow for the retrieval and reproducibility of essential steps and decisions, including justifications and reasoning for deviating from pre-specified plans. As in traditional computer system validation, experiments conducted in, or before, the development environment are not required to be documented step by step. However, when the outcome of such an experiment or analysis impacts how an AI solution is evaluated or deployed, the justification for such decisions should be documented. If a decision is based on certain results or insights from the development stage, this should be documented.

During the development phase, AI models undergo continual experimentation and iterative improvement. Transparency between the development team and the PV organization is crucial to ensure efficiency and that the solution is fit-for-purpose. Developers may create multiple versions of a model, test various features, and experiment with different training sets. In this context, focus should be on maintaining clear records of significant milestones – such as major changes in model

architecture, the introduction of new datasets, or significant shifts in performance metrics. This allows developers to track the evolution of the model and understand the implications of key changes without being overwhelmed by the sheer volume of minor tweaks and experiments.

Once an AI model moves from development to routine use in a production environment, the need for rigorous traceability and version control increases substantially. Deployed versions of the model should be documented in detail. In addition to the source code for each version, its underlying model architecture, training and test sets, and performance evaluation results should also be documented. From a regulatory perspective, the appropriate place to declare this would be in a document such as the PV System Master File (PSMF), in the EU.

The continual improvement and adaptation of AI models post-deployment should also be documented. It may be triggered by human domain experts or built into the deployment of the AI solution itself including pre-specified monitoring of deterioration of performance or model drift. Some challenges related to this for Software as Medical Device have been described by FDA [65].

When integrating external AI components (such as pre-trained models or libraries), it is important to document the versions of these components, particularly if they play a critical role in the model's performance. However, it may be sufficient to document these components at the time of significant milestones rather than during every iteration. As an example, for AI-based static systems, previous work proposes a specific documentation approach with proposed considerations for documentation within the different stages of the AI solution lifecycle.

Roles and responsibilities in artificial intelligence-enhanced pharmacovigilance systems

Organizations are accountable for the quality processes associated with their PV system, including the oversight of the AI components by the system owner. Oversight activities may be executed by a third party under appropriate supervision. Regulations, e.g. EU AI Act, may require organizations to establish specific roles, such as those to promote AI literacy, and facilitate fairness and equity.
AI systems themselves cannot be held accountable. Human oversight is essential for ensuring the safe and responsible use of AI. Clear roles and responsibilities must be defined for all stakeholders involved in AI initiatives.
The roles of PV experts are evolving with the introduction of AI. Already now, AI introduces new tasks, such as overseeing AI systems and interpreting their outputs. PV experts must adapt to these changes and develop new skills and competencies (see chapter on Human Oversight) to fulfil their obligations. This is especially relevant for members of the governance body and persons nominated as accountable for a lifecycle phase. The governance framework grid allows stakeholders to assess whether certain new activities will become relevant at specific steps, highlighting training needs early.

Just like with traditional software providers, the collaboration between vendors of AI solutions and PV experts is crucial. This collaboration can facilitate that AI systems meet PV requirements and governance principles. Regular audits of vendors and AI systems are essential for maintaining compliance and ensuring development standards. Effective collaboration and audits foster transparency and accountability. This can ensure AI systems that are reliable, meet regulatory standards and are inspection ready.

Regulatory authorities also play a role in monitoring AI in PV. They oversee that AI systems comply with regulatory standards and governance principles through inspections. Regulatory authorities are also developing guidance on the use of AI in the drug lifecycle, including PV (see Chapter on Landscape analysis). Integration of AI tools into the PV system must include appropriate regulatory documentation, such as in the Pharmacovigilance System Master File (PSMF) (see Chapter on Transparency).

PV inspections are likely to increasingly focus on AI systems, with inspectors reviewing AI-related documentation, performance metrics, and governance practices. Inspectors will need adequate

competencies to evaluate these systems effectively. This includes technical knowledge of AI and data science. As a result, continuous development and training are needed for inspectors to fulfil their role in new and fast-evolving areas.

Balancing innovation with regulatory compliance and adherence to guiding principles is important for the success of AI initiatives. This involves fostering a culture of responsible innovation. These goals can be achieved by establishing effective governance processes that include regular reviews of AI solution KPIs.

Chapter 10: Future considerations for development and deployment of artificial intelligence in pharmacovigilance

The evolution and future of artificial intelligence in pharmacovigilance

The chapter explores the continuing transformative impact of AI on PV from the current application to a vision of how AI might impact PV in the future. The CIOMS Working Group XIV's discussion in the earlier chapters of this report is grounded in common principles. Use cases (presented in Appendix 3) detail various AI systems under evaluation, various stages of deployment and assessment of their effectiveness within the discipline. To try to predict into the future, it is essential to recognize that the trajectory of AI is dynamic and highly unpredictable. Indeed, the only truly predictable elements are that AI will be ubiquitously deployed and is set to revolutionize many aspects, arguably all, of drug development and medical practice, from bench to bedside – as well as PV. For this reason, this chapter is grounded on the further developments of AI in PV described in earlier chapters of this report and anticipates how applications based on the principles might need to evolve as AI use in PV becomes more prevalent and sophisticated.

The chapter provides considerations for PV stakeholders, including regulators and healthcare professionals and other industry stakeholders to ensure AI's safe and equitable deployment in PV. The skillsets needed by PV professionals today will likely differ from those required in the future necessitating involvement in the design, development, deployment, and routine use of AI in PV. The examples illustrate the direction and immense potential of AI adoption in PV; however, these examples are speculative to a certain extent and are not meant to be exhaustive. AI is set not only to potentially revolutionize PV, dissolving traditional boundaries of PV, but also expand its footprint far more broadly across medical sciences.

The current decade represents a nascent phase for AI adoption in PV, and it is worthwhile acknowledging that the broad field of AI, particularly GenAI, is currently advancing rapidly. Further and more extensive deployment of AI may necessitate changes in how we think or approach PV strategies in the years ahead, driving the discipline of PV beyond its traditional frameworks and transforming it into self-detecting, real-time monitoring of safety data that aligns with the evolution of AI-driven medical science; for example, with the ability to rapidly analyse and extract vast quantities of safety data for case reporting and signal detection purposes. By leveraging this capability of AI, PV will evolve from a reactive focus on reporting and assessment to a forward-looking approach centred on proactive prediction and prevention and real / near real time learning systems.

The initial phase of evolution has started to impact PV's core activities including case management and safety surveillance, as it continues to move from reporting and assessment towards prevention, enabled by advancements in AI-enhanced healthcare and into radically new areas of medicine. These technologies have the potential to reduce manual workload and the burden on PV professionals. For example, by accelerating response times for priority events, increased capabilities to sift through large and varied sources of safety data including literature, automatically creating case reports, and performing signal detection.[245,246,247,248,249]

Transformative role of pharmacovigilance long-term and beyond: from detection to prevention

Advanced AI systems are poised to take PV beyond current boundaries and into automated or augmented decision making.[250] AI, with capabilities for approximate reasoning, could handle ambiguity and partial truth values, for instance, in assessing safety data from social media entries or

from fragmented safety-relevant data across different systems. Such AI systems may enable PV professionals to make nuanced decisions in case classification (e.g. assigning causality) or other PV situations requiring medical decision making. This would be particularly useful for cases with incomplete or conflicting data, where the "gray area" requires sophisticated, context-aware analysis and/or medical judgment.[251,252]

In the future, an expert AI system, designed specifically for PV, may emulate the judgement and decision-making processes of seasoned professionals or organizations with deep expertise in the field. These systems may not supplant human experts but augment their capabilities, enabling more nuanced and efficient decision making. An expert PV AI system would ideally be tailored to incorporate advanced analytical preciseness specific to therapeutic areas such as oncology, immunology, vaccines and medical devices as well as very different therapeutic options such as digital therapies, ensuring they adapt to the complexities of specific therapies, diseases, and patient populations, within those therapeutic areas, while performing or supporting PV work. While the development of such systems requires significant investment, their potential to drive the next generation of targeted PV solutions positions them as a critical innovation in advancing patient safety.

By mid-century, it is possible that traditional PV will have transitioned from primarily detecting and processing adverse effects to a frontline technology-driven discipline that is engineering technologies that can detect, evaluate and share the information with "self" (human or organ: heart, kidney, liver, lungs etc.).[253,254,255,256]

This may then allow HCPs and patients to take a more active role in vigilance and prevention by taking corrective actions before adverse symptoms arise. As a discipline, PV leveraging AI is likely to evolve into a function that develops technologies enabled by AI to perform a proactive assessment of anomalies, self-report and self-learn on how to prevent the presence of such anomalies in the future and continue to promote patient wellness and safety. This will include "true" AI-enabled proactive self-regulated vigilance and risk mitigation.

Future development and deployment of AI and the guiding principles

The CIOMS Working Group XIV members made the careful decision to structure the report around common principles for the use of AI in PV, based in part upon the recognition that this transformative technology is in a period of exponential growth. A report that was prescriptive and overly reliant upon current examples would quickly become outdated, especially if AI technologies from other healthcare domains are leveraged. The authors expect that common principles for the use of AI in PV will be durable for the foreseeable future. What is less certain is how the guiding principles may be applied. Although the principles are robust and are expected to endure, it is likely that they will evolve in parallel with the technical AI advances and their use in detection, prevention and decision making by the individual human or subject going under medical treatment. The potential implications are discussed for each of the guiding principles below.

Risk-based approach

Chapter 3 discusses risk-based approaches including risk mitigation, and also considers the regulatory framework required.

The proliferation and advancement of AI may lead to continuous self-learning and potentially autonomous AI systems, with potentially great advancement in PV and benefit to patients and HCPs.

Nevertheless, such systems come with potential concerns and risks. For example, a significant concern is the potential for AI to distort our understanding of a medicine's benefit-risk profile in real-world settings. Traditionally, these profiles are evaluated through carefully designed frameworks involving spontaneous reporting systems and planned surveillance studies. However, AI-driven tools

may inadvertently restrict prescribing practices; for instance, by limiting access to AI-enhanced PV systems for high-risk patients or preventing off-label use.

Further complicating matters, the adoption and availability of such tools may vary across healthcare systems and regions, introducing inconsistencies in data patterns that are challenging to interpret. This fragmented landscape can obscure the true influence that AI systems exert on prescribing decisions, making it difficult to assess their actual impact on patient outcomes. In addition, incorrect interpretation and poor utilization of AI is likely to significantly hamper patient safety. The principles of human factors and ergonomics (HFE) can assist in simplifying AI design and consecutively optimize human performance ensuring better understanding of AI outcome.[257]

The oversight and risk mitigation of such advanced AI systems demand a dynamic risk assessment framework; one that integrates near-real-time monitoring and adaptive evaluation processes. Ensuring effective communication of these evolving risks to all stakeholders, including patients, will be crucial. As part of risk mitigation, healthcare leaders must embrace flexible governance models that account for AI's evolving nature, ensuring that transparency, accountability, and equitable access remain at the forefront.

Human oversight

Chapter 4 covers human oversight including the changing and transformation of traditional roles in PV as AI use becomes increasingly embedded and ubiquitous.

As AI systems become increasingly pervasive and autonomous, the role of human oversight will inevitably shift. While maintaining a "human-in-the-loop" approach will likely remain essential, this may prove insufficient for highly complex or higher-risk applications — including aspects of PV. Conversely, in some scenarios, human oversight may substantially change and become less relevant, as AI systems surpass human capabilities in reviewing data and regulating their own processes.[258,259,260]

This evolving landscape will require PV professionals to develop new skillsets and undergo specialized training to effectively oversee AI-driven systems. The focus must extend beyond traditional oversight methods to include competencies in understanding, interpreting, and guiding AI behaviours. By cultivating these skills, PV professionals can ensure that human oversight remains meaningful and effective in safeguarding patient safety and public health.

Validity & Robustness

Chapter 5 discusses validity and oversight and considers multi-disciplinary collaborations required as well as reference standards and performance evaluation that might be needed to ensure robust and valid AI systems.

As AI becomes more embedded and sophisticated, the challenge is to develop appropriate methods and systems that validate and ensure data integrity in tandem with the developments. For example, with the potential for generating vast amounts of data in real time or near real time, there is a need for more appropriate validation methods to avoid the risk of false signals. This may require PV individuals to develop new skill sets or even new specific scientific disciplines. AI use with some advanced technologies would need the creation of new standards and validation methods for the outputs and real-time / near-real-time safety data generated, e.g. neurotechnology such as implantable chips, smart organs, nanotechnology and smart organs.

Transparency

Chapter 6 covers transparency and explainability of AI systems and related challenges.

As AI becomes increasingly pervasive, our ability to track its deployment and understand its decision-making processes may diminish, posing significant challenges to explainability and transparency. AI systems may mirror complex statistical processes and advance programming or AI-coded programs. Consequently, the necessity, and even practicality, of full transparency may face new challenges. Expectations of transparency may need to evolve as trust in AI systems strengthens and meets predefined confidence thresholds.

Much like AI's role in data analysis, statistics, and signal detection today, tracing AI's precise influence on downstream decisions may become increasingly difficult. Just as the complexities of prior distributions in Empirical Bayes Geometric Mean (EBGM) disproportionality models are widely accepted yet rarely scrutinized, established trust in AI-generated outputs may drive a shift in focus, with the expectation that errors or miscalculations will still prompt corrective actions to ensure sound decision making.

In parallel, as trust in AI solidifies, the emphasis on explainability may similarly evolve. While transparency will remain important, its most critical value may emerge during incidents or errors. Much like the role of flight data recorders in aviation, explainability may become vital for understanding failures and enhancing system improvements rather than serving as a constant requirement.

This shift may significantly influence PV decision making, emphasizing timely interventions and near-real-time root cause analysis. Looking ahead, organizations may need to balance the benefits of enhanced-AI performance against the degree of transparency required, carefully weighing improved efficiency with the need for interpretability in high-stakes decisions.

Data privacy

The right to control one's personal data is durable and has been widely adopted internationally. What is likely to occur in the coming years is that preserving data privacy will become more challenging. As noted in Chapter 7, leaks of personal data have been increasing in frequency, with some at enormous scale.[261] The increasing use of online platforms for communications and services has been accompanied (in some countries) by a common lack of understanding into how collected data are used along with an acquiescence to the risk of data breaches. Breaches have occurred for reasons ranging from neglect to criminal intent. In the case of health care data, the release of personal data contrary to individual approval carries risks for emotional well-being, stigmatization, and discriminatory treatment.

The pressures to amass and link large health care data sources are compelling, both on account of operational efficiencies (assuring consistencies in clinical care as well as medical care costs) and the advancement of scientific knowledge. At this time, the use of GenAI is in its infancy, and the only certainty is that it will both improve in quality and accelerate in use, as it is applied to many areas of biomedical research and clinical practice, and indeed in our daily lives. The use of open LLMs carries particular risks for the unintended disclosure of personal data, a topic that is likely to receive attention in coming years as the risk becomes clearer.

Societies will need to balance the pressures for the commoditization of data to maximize learning and therefore better outcomes for patients with AI, with protections against unintended disclosure. One possibility is that data sharing will be automated, but that systems have built-in checks and an obligation to maximise the demonstrable value of the data for the patients and/or patients' carers. Security measures to support anonymization might incorporate blockchain or similar technology to make complete anonymization possible without a patient key to allow all care-relevant data to be safely shared with complete confidence and assurance that the Individual's data are anonymised. Without the appropriate regulatory checks and balances, it is also easy to see that these data could easily be misappropriated or abused.

AI's evolution may usher in an era where access to underlying safety data becomes instantaneous, enhancing real-time insights and facilitating seamless data sharing. These advancements could significantly improve the timeliness and accuracy of safety assessments. However, an opposing scenario is equally plausible, one in which data sharing becomes increasingly restricted due to proprietary concerns, legal complexities, or public mistrust. As awareness grows regarding data's value as a commercial asset, particularly in insurance and other industries, heightened caution may further constrain data flow.

Balancing these dynamics will be critical. Establishing transparent frameworks that foster trust, ensure data integrity, and promote responsible data sharing will be essential to fully realize AI's potential while safeguarding public confidence.

Fairness & Equity

The fairness and equity Chapter 8 considers how and what type of discriminatory biases might be identified, addressed and/or prevented arising from the use of AI systems.

Fairness and equity should mean that patients and health care professionals should have equal access to all the new and advanced AI technologies.

It is important, as described in the data privacy section above, to ensure that PV with ubiquitous AI use is deployed equitably, and that data sharing does not put individuals at risk for e.g. higher costs associated with more advance monitoring, genetic profiling and/ or personalized risk / remediation, e.g. avoiding the risk of discrimination for insurance or treatment purposes.

AI should help to ensure equal understanding of safety data and its relevance to all patients, irrespective of social circumstances and background, and the understanding of benefits and risks to specific individuals or subgroups of the population.

This presents a unique challenge for vigilance, particularly in identifying rare, unexpected anomalies, the so-called Black Swan incidents.[262] While current PV systems are well-equipped to anticipate, assess, and manage common safety risks, they must also adapt in detecting these outlier events, particularly where advanced AI systems are deployed.

Governance & Accountability

Chapter 9 of this report covers Governance & Accountability including a governance framework grid for the lifecycle phases of AI solutions in PV.

The accelerated integration of AI underscores the need for dynamic, risk-based governance frameworks capable of near-real-time interventions.

This is especially true as AI systems become more autonomous and self-determining, for example, with automated patient or HCP alerts, which will self-monitor their function and output and take preventative measures based on self-detected alerts. Such advancements raise critical questions: how will governance, accountability, and human oversight of PV of these new technologies evolve in tandem with these capabilities?

Ideally, regulatory authorities and industry leaders in PV will establish robust oversight mechanisms to ensure that AI systems in PV are developed and deployed responsibly. Safeguards must be in place to protect against data misuse, uphold privacy standards, and ensure these technologies ultimately enhance outcomes for patients.

The growing autonomy of AI in PV further emphasizes the need for adaptable regulatory frameworks. Continuous surveillance, proactive auditing, and rigorous inspection protocols will be essential to mitigate risks, uphold patient safety, and protect public health. Achieving this will require

a shift toward governance models that are as agile and responsive as the technologies they seek to manage.

Conclusions to the future considerations for development and deployment of artificial intelligence in pharmacovigilance

Proliferation and deployment of AI and its integration into PV is set to give a paradigm shift in this discipline, which is likely to be focused on rapid or real-time data collection, assessment and reporting; for example, with the ability to analyse and extract vast quantities of safety data for case reporting and signal detection purposes at a rapid pace. This could fundamentally change the way we work to take advantage of these technological advances, for example, streamlining processes and causing changes in the wider healthcare environment and beyond, including patient privacy.

Along with the enormous potential for AI in PV, there are many challenges which warrant future consideration, particularly around oversight of autonomous AI systems, and how AI may impact data privacy and ethical frameworks. It is critical that the guiding principles outlined in this report remain as core considerations, but with the understanding that they will need to evolve and adapt with advancements and application of AI in PV and medicine in general. This is to ensure AI use in PV remains unbiased, transparent, and secure to prevent misuse or accidental harm. The appropriate human oversight, including regulatory and ethical safeguards, will be as crucial as the technological advancements being applied.

References

[245] *Kjoersvik O, Bate A. Black Swan Events and Intelligent Automation for Routine Safety Surveillance. Drug Saf. 2022 May;45(5):419-427. doi: 10.1007/s40264-022-01169-0. (Article accessed 28 April 2025)*

[246] *Ventola CL. The nanomedicine revolution: part 1: emerging concepts. P T. 2012 Sep;37(9):512-25. (Abstract accessed 28 April 2025)*

[247] *Ventola CL. The nanomedicine revolution: part 2: current and future clinical applications. P T. 2012 Oct;37(10):582-91. PMID: 23115468; PMCID: PMC3474440. (Abstract accessed 28 April 2025)*

[248] *Ventola CL. The nanomedicine revolution: part 3: regulatory and safety challenges. P T. 2012 Nov;37(11):631-9. (Abstract accessed 28 April 2025)*

[249] *Ventola CL. Big Data and Pharmacovigilance: Data Mining for Adverse Drug Events and Interactions. P T. 2018 Jun;43(6):340-351. (Abstract accessed 28 April 2025)*

[250] *Shneiderman, Ben, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, arXiv.org (February 23, 2020). https://arxiv.org/abs/2002.04087v1 (Extract from forthcoming book by the same title) Copyright Ben Shneiderman 2020 (Webpage accessed 28 April 2025)*

[251] *Eliminating Bias in AI-Assisted Decisions: Chouldechova, A., & Roth, A. (2020). The Frontiers of Fairness in Machine Learning, Communications of the ACM, discusses the ways AI can counteract cognitive biases in human decision-making. (Website accessed 28 April 2025)*

[252] *Hauben, Manfred. Artificial Intelligence and Data Mining for the Pharmacovigilance of Drug–Drug Interactions, Clinical Therapeutics, Volume 45, Issue 2, 117 - 133. DOI: 10.1016/j.clinthera.2023.01.002 (Article accessed 28 April 2025)*

[253] *Xiaolu Zhu, Zheng Wang, Fang Teng, A review of regulated self-organizing approaches for tissue regeneration, Progress in Biophysics and Molecular Biology, Volume 167, 2021, Pages 63-78, ISSN 0079-6107, https://doi.org/10.1016/j.pbiomolbio.2021.07.006. (Article accessed 28 April 2025)*

[254] *Wang C, He T, Zhou H, Zhang Z, Lee C. Artificial intelligence enhanced sensors - enabling technologies to next-generation healthcare and biomedical platform. Bioelectron Med. 2023 Aug 2;9(1):17. doi: 10.1186/s42234-023-00118-1. (Article accessed 28 April 2025)*

[255] *Mir, M., Permyakova, A., Das, R., & Ünal, A. B. (2018). Smart biomaterials for tissue engineering: Functional and antibacterial nanostructures. Biomaterials Science. Volume: 6, Issue: 9, Pages: 2382–2395.*

[256] *Discussion on advanced sensors for healthcare 2022-2024: "The Emergence of Smart Organs: Science and Fiction?" at National Institutes of Health, U.S. Department of Health and Human Services. This provided an overview of developing "smart" organ and tissue technology.*

[257] *Choudhury, A and Asan, O, Human Factors: Bridging Artificial Intelligence and Patient Safety (October 5, 2020). Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care. 2020;9(1):211-215. doi:10.1177/2327857920091007 (Article accessed 28 April 2025))*

[258] *Kurzweil, Ray (2005). The Singularity is Near: When Humans Transcend Biology. Viking Press.*

[259] *Bonabeau, E, Marco D, and Guy T, Swarm Intelligence: From Natural to Artificial Systems (New York, 1999; online edn, Oxford Academic, 12 Nov. 2020), https://doi.org/10.1093/oso/9780195131581.001.0001 (Abstract accessed 28 April 2025)*

[260] *Imagine you are Living in the Age of Singularity: Karakas, F (2021), Creative Adventures; 75 (Webpage accessed 28 April 2025)*

[261] *Cybersecurity Stats: Facts And Figures You Should Know. Forbes. (Webpage accessed 28 April 2025)*

[262] *Kjoersvik O, Bate A. Black Swan Events and Intelligent Automation for Routine Safety Surveillance. Drug Saf. 2022 May;45(5):419-427. doi: 10.1007/s40264-022-01169-0. (Article accessed 28 April 2025)*

2786

2787

APPENDIX 1: Glossary

This glossary provides definitions specific to terms within the context of Artificial Intelligence use in pharmacovigilance. Refer to the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) compiled by CIOMS in the Glossary of ICH Terms and Definitions and all other relevant glossaries available for any additional terms not described within this glossary.

Accountability

Accountability applies to clearly defined roles, responsibilities and liability for organizations and/or individuals deploying, operating and managing artificial intelligence systems. It requires the adoption of appropriate governance measures by relevant stakeholders (including but not limited to Regulators, Vendors, Users, Developers, Data Providers or Pharmaceutical Company) involved in setting policy, developing, deploying, maintaining and managing artificial intelligence systems. This ensures operation within expected parameters throughout the artificial intelligence lifecycle as well as managing any unforeseen consequences.

Proposed by CIOMS Working Group XIV.

Adverse event

Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment. An adverse event (AE) can therefore be any unfavourable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product.

Adopted from: Council for International Organizations of Medical Sciences (CIOMS) Glossary of ICH terms and definitions. Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text accessed 4 April 2025)

Adverse reaction

A response to a medicinal product that is noxious and unintended, meaning a causal relationship between the product and the event is at least a reasonable possibility.

Adopted from: Council for International Organizations of Medical Sciences (CIOMS) Glossary of ICH terms and definitions. Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text accessed 4 April 2025)

Artificial intelligence

An artificial intelligence (AI) system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

Adopted from: OECD (2024), "Explanatory memorandum on the updated OECD definition of an AI system", OECD Artificial Intelligence Papers, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/623da898-en.

Note: In the context of pharmacovigilance, the use of AI systems and activities is aimed at enhancing drug safety monitoring, patient safety and regulatory compliance.

Artificial intelligence literacy

Having the essential abilities needed to understand, learn and work in a digital world through AI-driven technologies.

Adopted from: Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, Maggie Shen Qiao, Conceptualizing AI literacy: An exploratory review, Computers and Education: Artificial Intelligence, Volume 2, 2021, 100041, ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2021.100041.

Augmented intelligence / Intelligence augmentation

Augmented intelligence is a conceptualization of artificial intelligence that focuses on artificial intelligence's assistive role. It emphasizes the use of artificial intelligence for enhancing, i.e. augmenting or amplifying human intelligence, rather than replacing it. Inherent in this view is the recognition that artificial intelligence and humans work together in a human-centered partnership, where each one can perform certain tasks better than either could alone.

Combined from:

- Madni AM. Augmented intelligence: A human productivity and performance amplifier in systems engineering and engineered human–machine systems. Systems engineering for the digital age: practitioner perspectives. 2023;Oct8:375-391. (Chapter abstract) https://doi/10.1002/9781394203314.ch17

- World Medical Association (WMA). *WMA Statement on augmented intelligence in medical care. 2019.* (Webpage accessed 3 April 2025)

Automation bias or automation complacency

Automation bias and automation complacency are overlapping manifestations of automation-induced phenomena, where human attention plays a central role. Both refer to the human tendency to favour suggestions from automated decision-making systems over non-automated contradictory information even when it is correct. They can involve attentional bias directed toward the automated output, or insufficient attention and monitoring of the automated output, especially in context of multi-tasking where manual tasks compete with the human expert's attention.

Combined from:

- Parasuraman R, Manzey DH. Complacency and bias in human use of automation: An attentional integration. *Human factors*. 2010Jun;52(3):381-410. (Journal full text) https://doi.org/10.1177/0018720810376055

- Cummings ML. Automation bias in intelligent time critical decision support systems. In Decision making in aviation. 2017;Jul5;289-294. Routledge. (Chapter abstract accessed 4 April 2025)

Bias

The tendency of a measurement process to over- or under-estimate the value of a population parameter.

Adopted from: Council for International Organizations of Medical Sciences (CIOMS). Glossary of ICH terms and definitions. Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text accessed 4 April 2025)

In AI, bias may be systematic difference in treatment of certain objects, people, or groups in comparison to others [18]. Bias can be introduced into study design, conduct or analysis. Sources of bias include selection bias (of study sample), operational bias, and analyses that do not account for missing data.

Adopted from: International Medical Device Regulators Forum (IMDRF). Machine Learning-enabled Medical Devices—A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions. 2021. (Full text accessed 3 April 2025)

In the context of artificial intelligence, bias can occur when the artificial intelligence data or algorithms reflect or perpetuate existing social inequalities, leading to discriminatory or unfair artificial intelligence outputs.

Adopted from: University of Saskatchewan. Generative Artificial Intelligence: Glossary of AI Related Terms. (Webpage accessed 4 April 2025)

Black-Box model

An analytics model that provides results based on received data but the logic used to provide those results cannot be determined or inferred on how it achieved those results.

Proposed by CIOMS Working Group XIV.

Black Swan event

Event of extreme impact that, although outside the realm of regular expectations (i.e. prospectively unpredictable), prompts humans to concoct explanations for its occurrence after the fact, making it seemingly explainable and predictable (i.e. retrospectively distorted).

Combined from:

- Kjoersvik O, Bate A. Black swan events and intelligent automation for routine safety surveillance. *Drug Safety.*2022;May;45(5):419-427. (Journal full text)

- Taleb NN. Black swans and the domains of statistics. The American statistician. 2007;Aug 1;61(3):198-200. (Journal abstract) https://doi.org/10.1198/000313007X219996

Business continuity plan

Set of provisions and systems for the prevention of / recovery from events that could severely impact on an organisation's staff and infrastructure in general or on the structures and processes for pharmacovigilance in particular, including the urgent exchange of information within an organisation, amongst organisations sharing pharmacovigilance tasks as well as between marketing authorisation holders and competent authorities.

Adopted from: Heads of Medicines Agencies (HMA). European Medicines Agency (EMA). Guideline on good pharmacovigilance practices (GVP): Module I – Pharmacovigilance systems and their quality systems. 2012. (Full text accessed 3 April 2025)

Change management

Change Management describes processes, methods and techniques designed and used to plan, implement and control changes to organizational structures and/or business processes. Methodologies span around people, process and culture.

Typically Change Management includes following components: Leadership alignment, Stakeholder engagement, Communication, Training, Impact Assessment, Continuous improvement.

Adopted from: International Organization for Standardization (ISO). *What is change management*: a Quick guide. (Webpage accessed 3 April 2025)

Class imbalance

Imbalance between categories in classification tasks. This affects model performance metrics, e.g. by the fact that a model always predicting the same outcome will be 99% accurate if 99% of test cases belong to the corresponding class.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Cluster analysis

An unsupervised machine learning method that groups data elements based on similarities to identify patterns that are not immediately evident.

Proposed by CIOMS Working Group XIV.

Computerized system validation

Process of establishing and documenting that the specified requirements of a computerized system are fulfilled consistently from design until decommissioning of the system and/or transition to a new system. The approach to validation should focus on a risk assessment that takes into consideration the intended use of the system and the potential of the system to affect human subject protection and reliability of trial results.

Adopted from: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *Integrated Addendum to ICH E6(R1): Guideline for good clinical practice E6(R2).* 2016. (Full text accessed 3 April 2025)

Confirmation bias

Confirmation bias is the tendency to give greater weight to data that support preliminary assumptive results, while failing to seek or dismissing contradictory evidence.

Adopted from: Elston DM. Confirmation bias in medical decision-making. *Journal of the American Academy of Dermatology*. 2020;Mar1;82(3):572. (Journal full text) https://doi:10.1016/j.jaad.2019.06.1286

Cross-validation

Resampling method used to assess the generalisation ability of a machine learning model and prevent overfitting.

Adopted from: D. Cross-Validation. *Preprint submitted to Encyclopedia of Bioinformatics and Computational Biology, 2nd edition (Elsevier).* 2019;542-545. (Full text accessed 3 April 2025).

Note: This is an alternative to maintaining separate training and validation data sets to provide a more efficient use of data during development.

Data anonymization

Anonymisation of personal data is the process whereby both direct and indirect personal identifiers are removed, and technical safeguards are used to ensure zero risk of re-identification.

Adopted from: World Health Organization (WHO). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. Geneva: World Health Organization. 2024. (Webpage accessed 3 April 2025)

Data drift

Change in the input data distribution a deployed model receives over time, which can cause the model's performance to degrade. This occurs when the properties of the underlying data change. Data drift can affect the accuracy and reliability of predictive models.

Adopted from: U.S. Food and Drug Administration. *FDA Digital Health and Artificial Intelligence Glossary–Educational Resource*. 2024. (Webpage accessed 3 April 2025)

Data privacy

Data privacy refers to measures taken to protect the fundamental right of individuals to the protection of their personal information. In the setting of PV, these measures emphasize the protection of sensitive and personal data (including health data).

Proposed by CIOMS Working Group XIV.

Decision tree

A model that uses a tree-like structure where data is progressed through various pre-defined tests or attributes to reach a final decision or prediction.

Proposed by CIOMS Working Group XIV.

Deep learning

A variant of machine learning involving neural networks with multiple layers of processing units known as artificial neurons, or 'perceptrons' (nodes), which together facilitate extraction of higher features of unstructured input data (for example, images, video and text).

Adopted from: Thirunavukkarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nature medicine. 2023;Aug;29(8):1930-1940. (Journal full text) https://doi.org/10.1038/s41591-023-02448-8

Approach to creating rich hierarchical representations through the training of neural networks with many hidden layers.

Adopted from: European Medicines Agency (EMA). Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. 2024. (Full text accessed 3 April 2025)

Explainability

The degree to which humans can understand the factors and logic that have led to a specific outcome or that play a role in the general operation of an AI system.

Proposed by CIOMS Working Group XIV.

Fairness and equity

Fairness and Equity requires awareness and adherence to the ideas of impartiality, equality, non-discrimination, diversity, justice and lawfulness. Avoidance and mitigation of unfair bias, discriminatory or unjust social wellbeing and environmental impacts and/or outcomes should be considered throughout the whole artificial intelligence lifecycle.

Proposed by CIOMS Working Group XIV.

False negative

The determination of a data point not belonging to a class of interest when the reference or test set states that does belong.

Proposed by CIOMS Working Group XIV.

False positive

The determination of a data point belonging to a class of interest when the reference or test set states that does not belong.

Proposed by CIOMS Working Group XIV.

Feature

A measurable property or characteristic of the data or engineered through data processing or transformation of the data that is used to train a model.

Proposed by CIOMS Working Group XIV.

Error! Bookmark not defined.

Few-shot learning

AI developed to complete tasks with exposure to only a few initial examples of the task, with accurate generalization to unseen examples.

Adopted from: Thirunavukkarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nature medicine*. 2023;Aug;29(8):1930-1940. (Journal full text) https://doi.org/10.1038/s41591-023-02448-8

Fuzzy logic

An approach to variable processing that allows for multiple possible truth values to be processed through the same variable. Fuzzy logic attempts to solve problems with an open, imprecise spectrum of data and heuristics that makes it possible to obtain an array of accurate conclusions.

Proposed by CIOMS Working Group XIV.

Generative artificial intelligence

Category of artificial intelligence techniques in which algorithms are trained on data sets that can be used to generate new content, such as text, images or video.

Proposed by CIOMS Working Group XIV.

Governance

A governance framework requires implementation of risk management practices and policies, responsible use, security, openness, fairness, and ethical practices to ensure adherence to the Artificial Intelligence Guiding Principles in the report of the CIOMS Working Group XIV.

Proposed by CIOMS Working Group XIV.

Hallucination

In natural language generation tasks, hallucinations are generated content that is either nonsensical or unfaithful to the provided source content.

Adopted from: Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232. 2023 Nov 9.

Human agency

Human agency is the capacity for human beings to make choices out of their own volition and to follow those choices to action.

Proposed by CIOMS Working Group XIV.

Human-in-command

The capability of a human to oversee the overall activity of an artificial intelligence system, including its broader economic, societal, legal and ethical impact, and the ability to decide if, when, and how to use an artificial intelligence system.

Adopted from: European Commission. *Ethics guidelines for trustworthy AI.* 2019. (Webpage accessed 3 April 2025)

Human-in-the-loop

The capability for human intervention in every decision cycle of the artificial intelligence system.

Adopted from: European Commission. *Ethics guidelines for trustworthy AI.* 2019. (Webpage accessed 3 April 2025)

Human-on-the-loop

The capability for human intervention during the design of an artificial intelligence system and monitoring of its operation.

Adopted from: European Commission. *Ethics guidelines for trustworthy AI.* 2019. (Webpage accessed 3 April 2025)

Human oversight

Human oversight refers to the expected role of humans in the design, implementation, monitoring, and analysis of AI in PV. It requires a framework to manage performance and to detect and mitigate potential issues related to the AI system.

Proposed by CIOMS Working Group XIV.

Individual Case Safety Report

A report containing information about a suspected adverse drug reaction related to the administration of one or more medicinal products to a specific patient at a specific time.

Adopted from: Council for International Organizations of Medical Sciences (CIOMS) Glossary of ICH terms and definitions. Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text accessed 4 April 2025)

Interpretability (*See also Explainability*)

A description of the general principles and logic by which an AI model functions and arrives at its outcomes / predictions should be shared, or the lack of explainability should be acknowledged and its implications discussed.

Proposed by CIOMS Working Group XIV.

Error! Bookmark not defined.

Knowledge graph

A heterogeneous knowledge base consisting of triples (facts) each comprised of object pairs and connecting relationships modelled through graphs and ontologies (a standardized machine readable semantic framework for representing all objects, and their properties and relationships in a domain of knowledge), which extract new insights from existing data sets via their integration.

Adopted from: Manfred Hauben, Mazin Rafi, Knowledge Graphs in Pharmacovigilance: A Step-By-Step Guide, Clinical Therapeutics, Volume 46, Issue 7, 2024, Pages 538-543.

Large language model

A type of artificial intelligence model using deep neural networks to learn the relationships between words in natural language, using large datasets of text to train, these include those with or without decoders.

LLMs can be classified according to their availability and accessibility into two main categories: open-source models (freely accessible to the public) and closed-source models (developed as commercial products and often necessitating licenses or subscriptions for use).

Adopted from: Heads of Medicines Agencies (HMA). European Medicines Agency (EMA). *Guiding principles on the use of large language models in regulatory science and for medicines regulatory activities*. 2024. (Full text accessed 4 April 2025)

Local Interpretable Model-Agnostic Explanations (LIME)

A technique that approximates a black box machine learning model with a local, interpretable model to explain each individual prediction.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Machine learning

Computational process of optimising the parameters of a model from data, which is a mathematical construct generating an output based on input data. Machine learning approaches include, for instance, supervised, unsupervised and reinforcement learning, using a variety of methods including deep learning with neural networks.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

(AI) Model

Mathematical or computational method with parameters (weights) arranged in an architecture that allows learning of patterns (features) from training data.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Model drift

A process where the model performance changes overtime either in a positive or negative performance outcome.

Adopted from: S. Wang, S. Schlobach, M. Klein, Concept drift and how to identify it J Web Semant: Sci Serv Agents World Wide Web, 9 (2011), 10.1016/j.websem.2011.05.003

Natural language processing

Field of artificial intelligence focusing on the interaction between computers and human language.

Adopted from: Thirunavukkarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nature medicine*. 2023;Aug;29(8):1930-1940. (Journal full text) https://doi.org/10.1038/s41591-023-02448-8

Negative controls

A real-world data point sampled as not belonging to the class of interest or deliberately created to not trigger a positive response from an artificial intelligence model.

Proposed by CIOMS Working Group XIV.

Neural network

Computing system inspired by biological neural networks, comprising (nodes), edges/weights, activation functions usually arranged in layers, communicating with one another and performing transformations upon input data.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Overfitting

Learning details from training data that cannot be generalised to new data.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Parameter, hyper- parameter

Variable within a machine learning model that is updated — usually automatically — during training to maximize performance. In deep learning, parameters are the 'weights' or data transforming functions comprising neural network nodes.

Adopted from: Thirunavukkarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nature medicine*. 2023;Aug;29(8):1930-1940. (Journal full text) https://doi.org/10.1038/s41591-023-02448-8

Hyper-parameters are parameters that are used to configure a model. Unlike model parameters, they cannot be directly estimated from data learning and must be set before training a machine learning model. Hyper-parameter tuning is a step often required to build effective ML models.

Adopted from: Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing. 2020 Nov 20;415:295-316.

Performance degradation

When results from an artificial intelligence system either fail or diminish in their ability to achieve the expected or required results as achieved earlier.

Proposed by CIOMS Working Group XIV.

Personal data

'Personal data' means any information relating to an identified or identifiable natural person ('data subject'). Information such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person are examples of personal data. Sensitive (personal) data refers to special categories of personal data.

Adopted from: European Parliament, Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Art. 4(1). *Official Journal of the European Union.* 2016; L 119. (Webpage accessed 4 April 2025)

Pharmacovigilance

The science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug related problem.

Adopted from: CIOMS Cumulative Glossary, with a focus on Pharmacovigilance (Version 2.1). Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text) https://doi.org/10.56759/ocef1297

Pharmacovigilance system

System used by an organisation to fulfil its legal tasks and responsibilities in relation to pharmacovigilance and designed to monitor the safety of authorised medicinal products and detect any change to their risk-benefit balance.

Adopted from: Heads of Medicines Agencies (HMA). European Medicines Agency (EMA). *Guideline on good pharmacovigilance practices (GVP) Module I – Pharmacovigilance systems and their quality systems.* 2012. (Full text accessed 3 April 2025)

Precision

Proportion of retrieved samples which are annotated as positive controls in the reference set, calculated as the ratio between correctly classified positive controls and all samples assigned to that class. Precision is also known as positive predictive value (PPV).

Adopted from: Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports.* 2022;Apr8;12(1):5979. (Journal full text)https://doi.org/10.1038/s41598-022-09954-8

Positive controls

A real-world data point sampled as belonging to the class of interest or deliberately created to trigger a positive response from an artificial intelligence model.

Proposed by CIOMS Working Group XIV.

Predictive models

A machine learning algorithm that analyzes data to identify patterns and trends, allowing it to make predictions about future outcomes or events based on input data.

Adopted from: De Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, Aardoom JJ, Debray TP, Schuit E, van Smeden M, Reitsma JB. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ digital medicine.* 2022;Jan10;5(1):2. (Journal full text) https://doi.org/10.1038/s41746-021-00549-7

Quality management system

Part of the pharmacovigilance system and consists of its own structures and processes. It shall cover organisational structure, responsibilities, procedures, processes and resources of the pharmacovigilance system as well as appropriate resource management, compliance management and record management.

Adopted from: D. Cross-Validation. *Preprint submitted to Encyclopedia of Bioinformatics and Computational Biology, 2nd edition (Elsevier).* 2019;542-545. (Full text accessed 3 April 2025).

Real-world data

Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. Examples of RWD include data derived from electronic health records (EHRs); medical claims and billing data; data from product and disease registries; patient-generated data, including from mobile devices and wearables; and data gathered from other sources that can inform on health status (e.g. genetic and other biomolecular phenotyping data collected in specific health systems).

Adopted from: Council for International Organizations of Medical Sciences (CIOMS) Glossary of ICH terms and definitions. Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text accessed 4 April 2025)

Recall

Proportion of positive controls correctly classified as such, calculated as the ratio between correctly classified positive controls and all positive controls. Also known as sensitivity or true positive rate (TPR).

Adopted from: Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports.* 2022;Apr8;12(1):5979. (Journal full text)https://doi.org/10.1038/s41598-022-09954-8

Reproducibility

The ability to achieve consistent results when analysis is repeated under the same conditions. Data and computer codes are used to regenerate the results.

Adopted from: National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Board on Research Data and Information; Division on Engineering and Physical Sciences; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Analytics; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Board on Behavioral, Cognitive, and Sensory Sciences; Committee on Reproducibility and Replicability in Science. Reproducibility and Replicability in Science. *Washington (DC): National Academies Press (US)*; 2019; May7. Chapter 3, Understanding Reproducibility and Replicability. (Chapter full text accessed 4 April 2025)

Risk-based approach

A risk-based approach acknowledges the potential hazards that artificial intelligence systems can pose and recognises that different use cases present varying types and levels of risk. This necessitates a risk assessment that identifies, prioritises, and manages potential risks that could negatively impact a pharmacovigilance system's behaviour and results, taking into consideration existing process controls. A risk is characterised by both the anticipated impact and the likelihood of negative outcomes.

This approach also supports procedures to identify and reduce errors and biases in a way that is proportionate to their risk. It influences the implementation strategies of AI systems (including documentation, compliance, and record-keeping), which should generally be commensurate with the identified risk.

Proposed by CIOMS Working Group XIV.

Robustness

A system reliably performs to its intended objectives or requirements accounting for known variances in data.

Proposed by CIOMS Working Group XIV.

Semantic vector

A mathematical representation of a word, phrase, or document as an identifier, where the identifier's position in the high-dimensional space captures the meaning or relationship of that word/phrase, allowing artificial intelligence systems to understand the context and similarity between different pieces of text based on their meaning.

Adopted from: Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. *Journal of biomedical informatics.* 2009;Apr1;42(2):390-405. (Journal full text) https://doi.org/10.1016/j.jbi.2009.02.002

Sensitivity analysis

An assessment technique used to evaluate how changes in input data or model parameters affect the output of an artificial intelligence model.

Proposed by CIOMS Working Group XIV.

SHapley Additive exPlanations (SHAP)

Explainable artificial intelligence framework that can provide model-agnostic local explainability for tabular, image, and text datasets.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Note: It is derived from cooperative game theory of payouts to groups of cooperating individuals.

Signal

Information that arises from one or multiple sources (including observations and experiments), that suggests a new potentially causal association, or a new aspect of a known association, between an intervention and an event or set of related events, either adverse or beneficial, that is judged to be of sufficient likelihood to justify further action to verify.

For the purposes of pharmacovigilance a signal is Information on a new or known side effect that may be caused by a medicine and is typically generated from more than a single report of a suspected side effect. It is important to note that a signal does not indicate a direct causal relationship between a side effect and a medicine, but is essentially only a hypothesis that, together with data and arguments, justifies the need for further assessment.

Adopted from: Council for International Organizations of Medical Sciences (CIOMS) Glossary of ICH terms and definitions. Geneva: Council for International Organizations of Medical Sciences. 2024. (Full text accessed 4 April 2025)

(Bioartificial) Smart organ technology

A series of enabling techniques that can be used to produce human organs based on bionic principles.

Proposed by CIOMS Working Group XIV.

Supervised learning

Machine learning that makes use of labelled data during training. (ISO/IEC DIS 22989).

Adopted from: International Medical Device Regulators Forum (IMDRF). Machine Learning-enabled Medical Devices—A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions. 2021. (Full text accessed 3 April 2025)

Target levels

A numerical value that serves as a goal or benchmark for artificial intelligence systems to achieve or surpass during their performance evaluation.

Proposed by CIOMS Working Group XIV.

Test dataset

A subset of the data that is never shown to the machine learning model during training, used to verify what the model has learned. (Modified from ISO/IEC DIS 22989).

Adopted from: International Medical Device Regulators Forum (IMDRF). Machine Learning-enabled Medical Devices—A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions. 2021. (Full text accessed 3 April 2025)

Data used to evaluate the performance of the AI system, before its deployment. It is expected to be similar to production data, and proper evaluation needs test data to be disjointed from any data used during development.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Traceability (AI)

The ability to track and document the data and processes used to create an artificial intelligence model.

Proposed by CIOMS Working Group XIV.

Training

Process intended to establish or to improve the parameters of a machine learning model, based on a machine learning algorithm, by using training data. (Modified from ISO/IEC DIS 22989).

Adopted from: International Medical Device Regulators Forum (IMDRF). Machine Learning-enabled Medical Devices—A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions. 2021. (Full text accessed 3 April 2025)

Training dataset

Data used specifically in the context of machine learning: it serves as the raw material from which the machine learning algorithm extracts its model to address the given task.

Adopted from: European Medicines Agency (EMA). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* 2024. (Full text accessed 3 April 2025)

Transparency

Transparency regarding AI involves disclosing information between organizations or individuals. This includes sharing relevant documentation of the AI system lifecycle (i.e. design, development, evaluation, deployment, operation, re-training, maintenance and decommission) to facilitate traceability and providing stakeholders with enough information to have a general understanding of the AI system, its use, risks, limitations, and impact on their rights.

Proposed by CIOMS Working Group XIV.

Unsupervised learning

Machine learning that makes use of unlabelled data during training. (ISO/IEC DIS 22989)

Adopted from: International Medical Device Regulators Forum (IMDRF). Machine Learning-enabled Medical Devices—A subset of Artificial Intelligence-enabled Medical Devices: Key Terms and Definitions. 2021. (Full text accessed 3 April 2025)

Validity

Validity means that a system achieves its intended purpose within acceptable parameters. It requires predefining acceptable performance levels, selecting appropriate data for model training and/or testing, assessing model performance in a realistic setting and integrating the system into an ongoing quality assessment process.

Proposed by CIOMS Working Group XIV.

Validation dataset

Data used to tune hyperparameters or to validate some algorithmic choices (rule design, etc.).

Adopted from: International Organization for Standardization (ISO). ISO/IEC DIS 22989. Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. 2022. (Webpage accessed 4 April 2025)

Zero-shot learning

Artificial intelligence developed to complete tasks without exposure to any previous examples of the task.

Adopted from: Thirunavukkarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nature medicine*. 2023;Aug;29(8):1930-1940. (Journal full text) https://doi.org/10.1038/s41591-023-02448-8

APPENDIX 2: Comparison table of guiding principles

Table 9: Comparison of CIOMS Working Group XIV guiding principles for artificial intelligence across regional and country government institutions, and international organizations – Extracted description of principles

Source: CIOMS Working Group XIV

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
Human Oversight	AI systems should support human agency and human decision making, as prescribed by the principle of respect for human autonomy.	When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.	Human Oversight means that high-impact AI systems must be designed and developed in such a way as to enable people managing the operations of the system to exercise meaningful oversight. This includes a level of interpretability appropriate to the context.		AI systems should function in a robust, secure and safe way throughout the AI life cycle, and risks should be continually identified, assessed and managed. Where appropriate, users, impacted third parties and actors in the AI lifecycle should be able to contest an AI decision or outcome that is harmful or creates.	Automated systems…"in design and development, pre-development and on-going disparity testing and mitigation, and clear organizational oversight"	Formal processes for human control and review of automated decisions are mandatory.	The principle of autonomy requires that any extension of machine autonomy not undermine human autonomy. In the context of health care, this means that humans should remain in full control of health-care systems and medical decisions. Human oversight may depend on the risks associated with an AI system but should always be meaningful	Mechanisms should be in place, as appropriate, to ensure that if AI systems risk causing undue harm or exhibit undesired behaviour, they can be overridden, repaired, and/or decommissioned safely as needed.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
								and should thus include effective, transparent monitoring of human values and moral considerations.	
Validity & Robustness	Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact	AI systems should reliably operate in accordance with their intended purpose.	Validity means a high-impact AI system performs consistently with intended objectives. Robustness means a high-impact AI system is stable and resilient in a variety of circumstances.		Consider how the associated actors on the AI supply chain can regularly test or carry out due diligence on the functioning, resilience and security of a system. Provide tools and guidance for undertaking AI-related safety risk assessments and implementing appropriate mitigations.		AI interventions should follow scientific best practice including being reliable, reproducible, fair, honest, and accountable.	All algorithms should be tested rigorously in the settings in which the technology will be used in order to ensure that it meets standards of safety and efficacy. The examination and validation should include the assumptions, operational protocols, data properties and output decisions of the AI technology. There should be robust, independent	AI systems must function in a robust, secure and safe way throughout their lifetimes, and potential risks should be continually assessed and managed.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
	with the AI system in an adversarial manner.							oversight of such tests and evaluation to ensure that they are conducted safely and effectively.	
Data Privacy	Principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that	AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.			Encourage AI developers and deployers (within their remit) to mitigate and build resilience to cybersecurity related risks throughout the AI life cycle. Encourage AI developers and deployers to consider and mitigate where possible potential malicious or criminal use of AI products and services.	"Data privacy…protections are included by default, including ensuring that data collection conforms to reasonable expectations and that only data strictly necessary for the specific context is collected".	Privacy, confidentiality, and security of data use must be foundational to every AI development.	Data protection laws are "rights-based approaches" that provide standards for regulating data processing that both protect the rights of individuals and establish obligations for data controllers and processors.	

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
	protects privacy.								
Transparency	…transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system.	There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.	Transparency means providing the public with appropriate information about how high-impact AI systems are being used. The information provided should be sufficient to allow the public to understand the capabilities, limitations, and potential impacts of the systems.	End-users of AI-MD (e.g. medical practitioners, patients) should be informed that they are interacting with an AI-MD.	Encourage AI developers and deployers (within their remit) to implement appropriate transparency and explainability measures.		Transparent approaches must always be used and communicated when developing AI algorithms. Everything must be as open and sharable as possible. Tools and underlying concept of Openness must be a feature and a critical success factor of any AI development.	AI should be intelligible or understandable to developers, users and regulators. Two broad approaches to ensuring intelligibility are improving the transparency and explainability of AI technology.	This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.
Accountability	The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment	People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the	Accountability means that organizations must put in place governance mechanisms needed to ensure compliance with all legal obligations of high-impact AI systems in the context in which they will be used.	While developers should be responsible for the proper design of algorithms used in the AIMD, organisations using AI-MD to deliver care will be responsible for the decision to implement the AI-	See Governance.	"…should have access to timely human consideration and remedy by a fallback and escalation process if an automated system fails, it produces an error, or you would like to	See Validity & Robustness.	Although AI technologies perform specific tasks, it is the responsibility of human stakeholders to ensure that they can perform those tasks and that	Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
	and/or use of AI systems.	outcomes of the AI systems, and human oversight of AI systems should be enabled.	This includes the proactive documentation of policies, processes, and measures implemented.	MD and the clinical outcomes arising from the use of AI-MD in ensuring that safe care is delivered. Similar to the implementation of any other MD, the use of AI-MD does not change the liability of the implementing institution or the individual medical professional in their provision of appropriate and safe care.		appeal or contest its impacts on you."		they are used under appropriate conditions. Institutions have not only legal liability but also a duty to assume responsibility for decisions made by the algorithms they use, even if it is not feasible to explain in detail how the algorithms produce their results.	
Societal well-being	…the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle. Ubiquitous exposure to social AI systems in all	AI systems should benefit individuals, society and the environment.		Safeguards in the design, development, and implementation of AI-MD should be put in place to ensure that patients' interests, including their safety and well-being, are protected.			Actions and solutions must be people centred and not be used solely by itself. As one of many technologies to aid public health AI should respect the rights of the individual.	AI technologies should not harm people. They should satisfy regulatory requirements for safety, accuracy and efficacy before deployment, and measures should be in place to ensure quality control	This Principle highlights the potential for trustworthy AI to contribute to overall growth and prosperity for all – individuals, society, and planet – and advance global development objectives.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
	areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or negatively impact our social relationships and attachment.							and quality improvement. Thus, funders, developers and users have a continuous duty to measure and monitor the performance of AI algorithms to ensure that AI technologies work as designed and to assess whether they have any detrimental impact on individual patients or groups.	
Environmental well-being	See Societal well-being.	See Societal well-being.						AI systems should be designed to minimize their ecological footprints and increase energy efficiency, so that use of AI is consistent with society's efforts to reduce the impact of human beings on the earth's	See Societal well-being.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
								environment, ecosystems and climate.	
Fairness & Equity	…enable inclusion and diversity throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models.	AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.	Fairness and Equity means building high-impact AI systems with an awareness of the potential for discriminatory outcomes. Appropriate actions must be taken to mitigate discriminatory outcomes for individuals and groups.	The development and implementation of AI-MD should not result in discriminatory or unjust clinical impact on patients across different demographic lines (e.g. race, gender, etc.).	AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes. Actors involved in all stages of the AI life cycle should consider descriptions of fairness that are appropriate to a system's use, outcomes and the application of relevant law.	"Algorithmic discrimination protections…should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities…".	Fairness, equality and inclusiveness in impact and design should always form the foundation of any AI initiative for Public Health. Discussions, developments, and implementation must be grounded in the globally-agreed ethical principles of human dignity, beneficence, nonmaleficence and justice.	Inclusiveness requires that AI used in health care is designed to encourage the widest possible appropriate, equitable use and access, irrespective of age, gender, income, ability or other characteristics. AI developers should be aware of the possible biases in their design, implementation and use and the potential harm that biases can cause to individuals and society.	AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and should include appropriate safeguards to ensure a fair and just society.
Explainability	…the ability to explain both the technical processes of the AI system and	See Transparency		The decisions or recommendations from an AI-MD should endeavour to be explainable	See Transparency	"Automated systems should provide explanations that are technically		See Transparency	See Transparency

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
	the reasoning behind the decisions or predictions that the AI system makes.			and reproducible. The level of explainability is dependent on the varying expectations of the end user and the risks of the AI-MD. End-users should be consulted during the development or adoption of the AI-MD to ensure the explainability meets their expectations.		valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context… in plain language and assessments of the clarity and quality of the notice and explanations should be made public whenever possible."			
Safety	See Validity & Robustness.	See Validity & Robustness.	Safety means that high-impact AI systems must be proactively assessed to identify harms that could result from use of the system, including through reasonably foreseeable misuse. Measures must be taken to mitigate the risk of harm.		Enable AI deployers (within their remit) and end users to make informed decisions about the safety of AI products and services. Communicate the level of safety related risk in their remit by appropriately identifying, monitoring, communicating	"Automated systems…should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts…".		Preventing harm requires that use of AI technologies does not result in any mental or physical harm.	See Validity & Robustness.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	**EU**[263,264]	**Australia**[265]	**Canada**[266]	**Singapore**[267]	**UK**[268]	**US**[269]	**PAHO**[270]	**WHO**[271]	**OECD**[272]
					and acting upon risks.				
Governance	See Data Privacy.				Governance measures could be put in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle.			Human rights standards, data protection laws and ethical principles are all necessary to guide, regulate and manage the use of AI for health by developers, governments, providers and patients.	

3413

References

[263] Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), 9 September 2024. (PDF accessed 27 April 2025)

[264] European Commission: Directorate-General for Communications Networks, Content and Technology, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment, Publications Office, 2020, https://data.europa.eu/doi/10.2759/002360 (PDF accessed 27 April 2025)

References

[265] Australian Government, Department of Industry, Science and Resources, Australia's Artificial Intelligence Ethics Principles (Webpage accessed 27 April 2025)

[266] Government of Canada, The Artificial Intelligence and Data Act (AIDA) – Companion document (Webpage accessed 27 April 2025)

[267] Artificial intelligence in healthcare guidelines. Ministry of Health, Singapore Government Agency. (PDF accessed 27 April 2025)

[268] Implementing the UK's AI Regulatory Principles Initial Guidance for Regulators. Department for Science, Innovation & Technology. UK Government. February 2024. (PDF accessed 27 April 2025)

[269] Blueprint for an AI Bill of Rights, Making automated systems work for the American people. The White House. (Website accessed 27 April 2025)

[270] Artificial Intelligence in Public Health, Digital Transformation Toolkit. Pan American Health Organization, 2021. PAHO/EIH/IS/21-011. (PDF accessed 27 April 2025)

[271] Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO. (Website accessed 27 April 2025)

[272] OECD AI Principles overview © 2025 OECD. (Website accessed 27 April 2025)

APPENDIX 3: Use cases

Use Case A: Large Language Models data extraction for case processing

Source:[273]

Area of PV: ICSR Processing

A1. Business rational and challenges

As per Good Pharmacovigilance Practices (GVP), pharmaceutical companies must act on potential adverse reactions to drugs. With significant increases in the number of case reports in recent years, case intake/processing operations face complex challenges beyond the number of cases, such as handling very diverse data sources including unstructured texts and scanned documents or managing sudden peak inflows with a finite workforce. With the complexity of the relevant data points ranging from simple demographics to more complex lab values, simpler technology approaches like Named Entity Recognition were unsuccessful in consistently improving case intake/processing operations under real-world circumstances. The use of Large Language Models (LLMs) in case intake/processing provides potential to advance processes without compromising quality.

A2. Solution

A pharmaceutical company executed a proof-of-concept study to assess the feasibility as well as the quantitative and qualitative business impact of utilizing LLMs for case intake purposes. Specifically, LLMs were applied for data extraction from source documents for case intake and processing while covering regulatory and compliance aspects.

To process the selected source documents and extract pre-defined pieces of information, a 3-step semi-automatic processing pipeline was set up. The pipeline consisted of (1) pre-processing steps to unify the input for the LLM (OpenAI´s GPT-4), (2) a JSON-formatted extraction template that guided the LLM in structuring the information as well as providing hints regarding the location of the information in the source data, and (3) post-processing steps to match the model output with fields where predefined values were applicable. Original source documents were augmented by references and highlighting of extracted key terms.

For the assessment of the business impact of using LLMs for case intake, a selection representative cases was identified. A graphical user interface (GUI) was designed for the purpose of comparing the processing performance of (a) the fully manual process vs (b) the manual process augmented by fields pre-filled by the results of the LLM extraction pipeline. Four experienced professionals were randomly assigned to either process version (a) or (b). The processing times were tracked for each source document to derive the overall processing time regarding extraction of the representative set of fields.

A3. Results

In this study, two key results were derived from the implementation of LLMs in the case intake and processing operations:

The first result focused on the performance of the LLM model, measured through the match scores of all extracted fields and averaged across cases of a category for the full number of source documents in scope of this study. The statistical evaluation revealed that the model achieved match scores, ranging from 85% to 100% for clinical studies, and 60% to 100% for patient support programs (PSP) cases. For literature cases, while the sample size precludes a robust statistical evaluation, model performance ranges from 67 to 100%, suggesting qualitative results that align with the other types.

The high match scores achieved by the model demonstrate its capability to extract accurate and relevant information from unstructured sources. This can be translated into tangible efficiency gains for business operations.

The second result highlighted the efficiency gains identified in the business impact assessment. The implementation of LLM in case intake led to an estimated efficiency gain of 39%, translating to time savings of approximately 20 minutes per case. Specifically, the study found that the average number of data points extracted per case was 69.4, with only 2.4 data points requiring manual correction.

Implementing LLMs is not just a technical enhancement; it represents a strategic move towards improving operational efficiency and ensuring high-quality outcomes in PV practices.

A4. Compliance with the governance framework

Table 10: Use case A: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	A risk-based approach has been followed thoroughly and a 100% QC by human interaction has been applied.	A	A	N/A	N/A	N/A
Human oversight	Implementation of dedicated features to support human oversight, including user-friendly interfaces and references to the source data. The 100% human QC ensures robustness of all extraction outputs.	A	A	N/A	N/A	N/A
Validity and robustness	No continuous learning is applied; rather, the model is used in a locked state. Releases of new versions are quality assured on a sufficiently broad test set to derive.	A	A	N/A	N/A	N/A
Transparency	Model performance has been measured with match score. The correction of the failures can be used as feedback in regular intervals to improve the prompting strategy.	A	A	N/A	N/A	N/A
Data Privacy	The service is established on a private cloud. Access is provided only to project team members. Personally identifiable information is redacted prior to the actual data extraction step.	A	A	N/A	N/A	N/A
Fairness and Equity	Not applicable. The application is not providing any data consolidation or decision support. The 1:1 match of the data extraction is verified by the human QC.	N/A	N/A	N/A	N/A	N/A
Governance & Accountability	The LLM model is using tailored prompting strategy maintained on vendor domain to test the data extraction. The model is provided by Open AI, and is powered by a selection of large language models (“LLMs"). The case intake and processing team takes over the accountability and perform the 100% human QC process. The ultimate accountability remains with the pharma company.	A	A	N/A	N/A	N/A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

References

[273] *Römming, H.-J. el. How LLMs can advance safety case intake – points to consider and insights from a Proof of Concept. Submitted to "Therapeutic Advances in Drug Safety" as editorial article in Mar-2025.*

3478

Use Case B: Case deduplication

Source:[274]

Area of PV: ICSR Processing

B1. Business rational and challenges

Adverse event reporting systems (AERS) are essential in PV as they support the identification and evaluation of safety signals related to the use of medical products. Expert review in safety monitoring involves several steps, such as data mining and case series analysis, which are significantly affected by the AERS data quality. A representative example of quality issues is duplication, where more than one report describes the same patient case and the same adverse event experience for the same product. Duplicate reports may result in false or missed safety signals and increase the workload for safety evaluators by misinterpreting the actual number of true adverse events and making a product-event relationship look weaker or stronger.

B2. Solution

A regulatory agency that maintains an AERS for drugs and biologics with >28 million historical reports and an average of 8,000 new submissions daily sought an efficient solution to deduplicate all historical and incoming adverse event reports. The regulatory agency collaborated with an academic partner to address this issue by developing a **deduplication pipeline** relying on modern technologies (mainly, natural language processing, network analysis, and cloud computing) and utilizing structured data and free-text narratives. The pipeline executes an initial pass to filter down the pairs of reports by placing minimum requirements on similarity based on demographic data and other features. Subsequently, a pairwise streamlined worker implementing a **duplicate detection algorithm** performs a probabilistic comparison of all qualifying report pairs and calculates two scores, a probabilistic weight score and a second component score value, that together rate how similar the two reports are. In the third step, the pairs exceeding a preselected validated threshold that was specified in a dedicated analysis are merged into networks (a.k.a. groups) of potentially duplicate reports and split into tightly linked communities (a.k.a groups) of actual duplicates. Finally, a reference case selection component identifies the most representative report in each duplicate group based on several parameters and the remaining reports in the group are flagged as duplicates and they are excluded from subsequent data mining calculations. An existing decision-support tool developed to support the case series analysis allows for evaluating the groups of duplicate reports and verifying the reference case, keeping medical reviewers in the loop.

B3. Results

In an early research study, the **duplicate detection algorithm** was applied to two datasets of post-market reports, one including vaccine product reports and one containing reports for biologics, identifying 77% of and 13% of known duplicate pairs, respectively, with (nearly) perfect precision in both cases (95% and 100%, respectively) (https://pubmed.ncbi.nlm.nih.gov/28293864/). This algorithm was refined in subsequent steps to reach acceptable levels of performance that, in some cases and based on new evaluations using drug adverse event reports, supported the detection of duplicate pairs with an F-measure >0.9. The medical reviewers who participated in this new evaluation round felt confident about the algorithm and expressed their interest in using it, as discussed in the corresponding publication (https://www.frontiersin.org/articles/10.3389/fdsfr.2022.918897/full). Subsequently, the medical reviewers generated a gold standard of 2300 reports with labelled duplicates in a systematic process to support the validation of the recently built **deduplication pipeline**, which was then compared with existing deduplication approaches used at the regulatory agency. The deduplication pipeline outperformed these approaches (results unpublished) and was approved for processing all historical reports and incoming live data in an ETL process. As of December 20, 2024, the pipeline, installed on

the AWS environment and tightly integrated with the agency's AERS, has screened >30 million historical reports and continues deduplicating an average of 8,000 new submissions daily.

B4. Compliance with the governance framework

Table 11: Use case B: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	A risk-based approach has been discussed extensively, especially regarding missed or false positive duplicate reports. It has been determined that implementing the pipeline in the decision-support system, with humans-in-command, eliminates any risks for the case series analyses. What remains to be done is acknowledging any risks for data mining calculations and potential noise in signal detection; this part has not yet been fully developed and mostly affects the routine use of deduplication for data mining calculations and not its use in case series analyses that is currently fully implemented.	A	A	A	A	A
Human oversight	Human experts actively provided feedback to the software engineers during the development stage and evaluated the deduplication output to refine and validate the pipeline. Human experts can confirm or modify the reference case selection using an existing decision-support tool while conducting their case series analyses in the routine use setting. On the other hand, data mining calculations incorporate deduplication output without humans being involved.	A	A	A	A	A
Validity & Robustness	The deduplication pipeline has been evaluated and validated to ensure it meets expectations and serves its intended purpose. The effect of deduplicated data on data mining calculations and the discovery of potential safety signals, which is one of the major uses of deduplication output, has not yet been investigated.	A	A	A	A	A
Transparency	Several publications, technical reports, and other documentation describe the pipeline and results of all evaluations conducted with safety reviewers' assistance.	A	A	A	A	A
Data Privacy	Fully complying with the principle as all processing occurs in a secure cloud environment.	A	A	A	A	A
Fairness & Equity	The deduplication pipeline has been evaluated and validated in several rounds and is closely monitored in the post-deployment phase. The pipeline is fully migrated to the production environment to be routinely used at the time of writing this report; it is therefore marked as partially aligned since this process has not been completed yet.	A	A	A	A	A
Governance & Accountability	System administrators have full control and continuously monitor the deduplication pipeline as well as the use of its output in the decision-support tool. A plan has also been developed to incorporate the deduplication output in the data mining calculations.	A	A	A	A	A

	Clearly defined roles were specified in the development, pre-deployment, and post-deployment stages, where the Contractor led the pipeline's construction and incorporation into the decision-support tool and the existing environment at the regulator's site, assisted by the end users and other stakeholders. Roles have not yet been fully assigned in the routine use setting.					

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

References

[274] Kreimeyer K, Spiker J, Dang O, De S, et al. Deduplicating the FDA adverse event reporting system with a novel application of network-based grouping. J Biomed Inform. 2025 May;165:104824. doi: 10.1016/j.jbi.2025.104824. (PubMed accessed 29 April 2025)

Use Case C: Artificial intelligence translation assistant

Source:[275]

Area of PV: ICSR reporting

C1. Business rational and challenges

The pharma company had engaged with a vendor to consolidate and streamline the global case intake and translation process. The vendor had established 2 hubs in Europe and Asia to cover 16 languages across 32 countries replacing a distributed network of multiple local country organizations and local vendors. To further increase productivity, the vendor had been requested to automate the translation process.

C2. Solution

While processing foreign language adverse event reports, about half of the effort was required for accurate translation of source documents from local languages to English, enabling centralized case management in English and subsequent submission to authorities. The pharma company and the vendor formed a common project team consisting of experts on ML and PV associates to pilot an AI-powered translation assistant based on commercially available technology. The team had set up a private cloud environment to store learning data (source texts and human-edited translations) and developed a user interface to input original text and retrieve and (if necessary) edit the result. The system automatically stores and analyses any modifications done by the users to enable further learning iteration and improvement of the first-time quality of the AI translation assistant. A 100% QC (Quality Control) by a human translator of all the translations was established to always verify the accuracy of the translation. The solution facilitates continuous learning through the automated integration of the manual edits into the translation model in defined regular intervals. With each model update the relevant quality measures (BLEU scores, see below) are re-calculated.

C3. Results

The translation's quality was assessed by BLEU scores. BLEU (Bilingual Evaluation Understudy) is a metric for evaluating machine-translated text. The BLEU score is a number between zero and one that measures the similarity of the machine-translated text compared to a set of high-quality reference translations. Within 6 months the AI translation assistant mimicked the quality of a human translator (i.e. BLEU equal or greater than 0.6).[276]

The results of the AI Translation Assistant pilot for the first language (Portuguese) were leading to a reduction of translation efforts by ca. 30%. Hence the solution was extended to 5 further languages (Chinese, Spanish, French, German, and Dutch). Pharma company and vendor teams are jointly and continuously evaluating the BLEU score to monitor the quality of the solution.

Improving the AI model is a function of case volume as every revised sample translation provided by the QC team helps to improve the model. More samples make better models, and better models finally reduce the effort for the team, allowing them to work through more cases, faster, and with greater consistency.

Since the initial setup of the AI Assistant for Translation in 2021 the technology has further matured. The translations obtained by the assistant are generally of high quality and less than 10% corrections are still necessary by the human translators. An analysis is ongoing to move from 100% human QC to a sampling approach as a continuous monitoring measure. Depending on the detailed results per language, the sample size may be adjusted within reasonable borders (up to 100%) to ensure high quality translations continuously.

C4. Compliance with the governance framework

Table 12: Use case C: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	A risk-based approach has been followed thoroughly and a 100% QC by human translators has been applied. With further maturing of the system and under close monitoring of the overall quality a reduction of human QC for individual translations should be possible.	A	A	A	A	A
Human oversight	To ensure human oversight a 100% human QC of the translated text by the vendor translators was established from the beginning. The BLEU scores are regularly measured for each language to identify changes in the overall performance.	A	A	A	A	A
Validity and robustness	The system has been implemented following the vendors standard validation approach. The 100% human QC ensures validation of all translation outputs. Any failure of the translation assistant would be immediately detected and corrected.	A	A	A	A	A
Transparency	Transparency of the translation performance is obtained as all translations are tracked by the system as well as any edits by the human translator. These edits are used in regular intervals to improve the model.	A	A	A	A	A
Data Privacy	The service is established on a private cloud. Access is provided only to project team members. Personally identifiable information is redacted prior to the actual translation process. The original source document remains available only for the local team who received the initial information and who may have to follow-up with the initial reporter.	A	A	A	A	A
Fairness and Equity	Not applicable. The application is not providing any data consolidation or decision support. The 1:1 match of the translation is verified by the human QC.	NA	NA	NA	NA	NA
Governance & Accountability	The translation assistant is a standalone tool owned by the vendor. Hence, the regular life-cycle governance is executed by the vendor and available on request to the pharma company. It concerns, e.g. the update of the model based on learning progress. While the responsibility for the execution of the translation lies with the vendor the ultimate accountability remains with the pharma company. Hence, in addition to the 100% human QC process by the vendor, the pharma company is doing a defined sample QC of the overall case intake results, including the translation.	A	A	A	A	A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

References

[275] *Römming H-J, Pushparajan R. AI Translation Assistant for Pharmacovigilance. Poster presented at DIA Europe 2021. (Full text accessed 21 March 2025)*

[276] *Google Cloud. Evaluate AutoML Translation models. Google Cloud. (Webpage accessed 21 March 2025)*

Use case D: Large language models for context-aware Structured Query Language

Source Article:[277]

Area of PV: Safety analysis

D1. Business rational and challenges

Safety scientists are often reliant on technical teams for safety query formulation and extraction of data from safety databases using SQL, which can introduce delays in assessment. The aim therefore was to enhance the accuracy of information retrieval from PV databases by employing LLMs to convert natural language queries (NLQs) into Structured Query Language (SQL) queries, leveraging a business context document.

D2. Solution

A sandboxed version of OpenAI's GPT-4 model was utilized within a retrieval-augmented generation (RAG) framework, enriched with a business context document, to transform NLQs into executable SQL queries. The study was conducted in three phases, varying query complexity, and assessing the LLM's performance both with and without the business context document.

D3. Results

The integration of a business context document markedly improved the LLM's ability to generate accurate SQL queries (i.e. both executable and returning semantically appropriate results), increasing from 8.3% with the database schema alone to 78.3% with the business context document. This enhancement was consistent across low, medium, and high complexity queries.

The method is an assistive method to enable non-technical users to perform complex data queries, potentially enhancing timeliness of PV data analysis and reporting.

D4. Compliance with the governance framework

Table 13: Use case D: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	Within this study, the intent is to demonstrate use of natural language to generate SQL queries to retrieve data from a safety database. The risk based approach should consider the feasibility of implementation and controls and processes needed to ensure its appropriate and trusted use.	A	A	N/A	N/A	N/A
Human oversight	Within the PoC human experts have reviewed the relevance of the outputs against a reference standard, within a product setting consideration will need to be given to how human oversight will be provided to ensure robustness of the outputs, including monitoring performance over time and at defined intervals (e.g. change in model version). Human oversight could be applied using a risk-based approach e.g. taking into consideration the intended use of the output.	A	A	N/A	N/A	N/A
Validity & Robustness	The tool has been evaluated against a curated reference standard, beyond the PoC consideration should be given to generalisability in production use.	A	A	N/A	N/A	N/A

Transparency	Whilst there is transparency of the GPT model, the use of RAG and context specific documentation provides transparency of the pipeline and how data is processed to achieve the output.	A	A	N/A	N/A	N/A
Data Privacy	This is an assistive tool not using individual patient data to generate SQL outputs.	N/A	N/A	N/A	N/A	N/A
Fairness & Equity	This is an assistive tool not using individual patient data to generate SQL outputs.	N/A	N/A	N/A	N/A	N/A
Governance & Accountability	During the proof of concept (PoC), the accountability of the methodology remains with the developer. However, if the methodology is integrated into a production setting, accountability would transition to the human subject matter expert. Governance within a PoC ensures that scientific integrity principles are adhered to, while future product use governance should cover how the tool fits into the overall PV system and quality management system (QMS).	A	A	N/A	N/A	N/A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

References

[277] *Painter JL, Chalamalasetti VR, Kassekert R, Bate A. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. JAMIA open. 2025;Feb;8(1):ooaf003. (Journal full text) https://doi.org/10.1093/jamiaopen/ooaf003*

Use Case E: Causality assessment of adverse drug reactions

Source:[278]

Area of PV: Causality Assessment

E1. Business rational and challenges

Assessing the causal relationship between an adverse event and the patient's exposure to a drug is a critical part of the PV process. Causality assessment is a time-consuming process requiring manual review by medical experts who evaluate data in the case with data from external sources (e.g. drug labels, scientific publications, drug mechanism of action, and disease symptoms). As the volume of adverse events to be reviewed increases an opportunity exists to create solutions that leverage ML to support the medical experts by predicting causality assessments.

E2. Solution

The authors of this paper created a modelling feature set comprising of various data attributes from solicited cases from the pharmaceutical company's safety database relevant to causality assessment of drug-event combinations. This was supplemented by engineered data features comprising external data and data from other internal sources. The resulting training data schema (shown below) was selected as it provides a comprehensive set of features relevant to the causality assessment process.

Table 14: Use case E: Modeling Data
Source:[279]

Modeling Data		
Case Level Data		**External Sourced Data**
Causality Label	Medical History Exclusions	Disproportionality
Rechallenge	Drug Exclusions	Anatomical Therapeutic Class & System Organ Class
Labeledness		Temporal Relationship
Reporter Causality		Temporal Compatible

In parallel, a separate decision support tool (CASCADE) was developed and validated through consultation with experienced drug safety physicians. A decision tree structure was adopted due to its increased transparency and interpretability when compared to other causality assessment algorithms. This increased transparency and interpretability allow a clear statement of the rationale for the assessment to be written (e.g. "The case is deemed causally related as it is (a) Labelled for the event (b) The event has a plausible temporal relationship, etc.").

The work on the decision tree provided a basis for the subsequent predictive model, informing contributing factors and the topology of the resulting Bayesian Network model. The authors rationale for selecting this type of model include: the ability to combine multiple sources of information with expert knowledge, transparency and interpretability, and their capability to model complex frameworks with causal dependencies where a lot of uncertainty exists. Model training utilized an annotated dataset of 50k cases, with a separate test dataset of 20k cases. Both the training and test dataset represented a broad range of drug classes and event categories. All cases had been previously assessed by medical experts and were taken from a period where the causality assessment practices were consistent.

E3. Results

The model demonstrated high performance (sensitivity was 0.900, with Positive Predictive Value of 0.778) in predicting the causality assessment of drug–event pairs compared with clinical judgment using global introspection. The authors also explored a learned topology Bayesian Network model

with the same training data. The learned topology model was found to have inferior performance compared to their CASCADE-based model.

E4. Compliance with the governance framework

Table 15: Use case E: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based Approach	A risk-based approach was used to limit the scope of the model to solicited, post-marketing cases as automating the causality of assessment of these cases was determined to have a lower risk/ impact to the PV system.	A	A	N/A	N/A	N/A
Human Oversight	Drug safety physicians and SMEs were involved in the data review and model development activities, ensuring the applicability of the model to its intended purpose. As this was a POC/study, there was no discussion about the creation of a quality management framework to support human oversight for future production use.	A	A	N/A	N/A	N/A
Validity & Robustness	The use case and deployment domain are described in the paper. The data selection, model training and testing activities used in model development are discussed in detail, as is the approach used for performance assessment. The authors consider areas for investigation that may be used to further demonstrate the model's validity and improve robustness.	A	A	N/A	N/A	N/A
Transparency	The paper provides information about intended use of the model and its design, including the decision tree (CASCADE). Data, results, areas for further investigation, and how the model could be applied in a PV system are discussed.	A	A	N/A	N/A	N/A
Data Privacy	In alignment with the principles in this book, the data used for the development, training, and validation of the model is from the company's internal post-marketing safety database suggesting it was obtained with the patient's/reporter's consent and in compliance with relevant privacy laws and regulations.	A	A	N/A	N/A	N/A
Fairness & Equity	Based on the article, it is not possible to comment on whether model development aligns with this guiding principle	N/A	N/A	N/A	N/A	N/A
Governance & Accountability	There is no discussion of governance and accountability activities, as defined in this guidance, in the paper. The authors do acknowledge the need for models to remain compliant with regulatory frameworks and guidelines. Further, the CASCADE decision tree created is referenced as a causality assessment support tool implying accountability for the final causality assessment decision remains with the drug safety SME.	N/A	N/A	N/A	N/A	N/A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management

PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

References

[278] *Cherkas, Y., Ide, J. & van Stekelenborg, J. Leveraging Machine Learning to Facilitate Individual Case Causality Assessment of Adverse Drug Reactions. Drug Saf 2022;45;571-582. (Journal abstract) https://doi.org/10.1007/s40264-022-01163-6*

[279] *Modified from Cherkas, Y., Ide, J. & van Stekelenborg, J. Leveraging Machine Learning to Facilitate Individual Case Causality Assessment of Adverse Drug Reactions. Drug Saf 45, 571–582 (2022). https://doi.org/10.1007/s40264-022-01163-6*

Use Case F: Process efficiencies supporting signal detection

Source Article:[280]

Area of PV: Signal Detection

F1. Business rational and challenges

One of the most time- and resource-demanding procedures for dismissing safety signals is the identification of alternative causes for the reported adverse events (AEs) in ICSRs after signals of disproportionate reporting have been identified. This includes the screening of co-reported drugs to identify alternative potential causes for the newly identified drug–event pair.

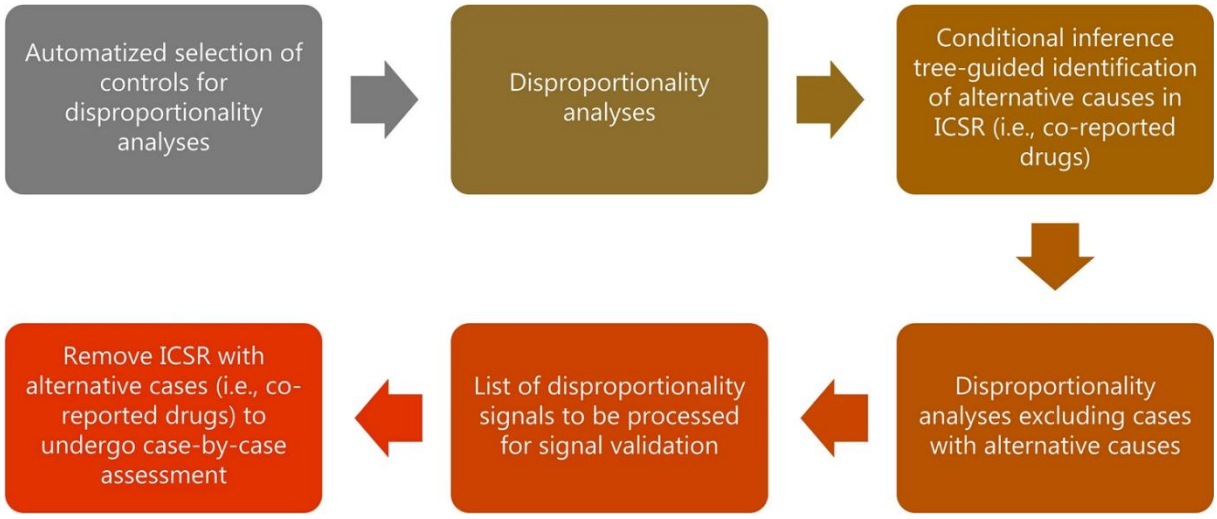
F2. Solution

This study aimed to develop an AI-based framework to automate (1) the selection of control groups in disproportionality analyses and (2) the identification of co-reported drugs serving as alternative causes, to look to dismiss false-positive disproportionality signals.

The implementation of automatic selection of controls and dismissal of false positive signals using a conditional inference tree is summarized in the flowchart below.

Figure 8: Flowchart summarizing the implementation of the automatic selection of controls and the dismissal of false positive signals when using a conditional inference tree

Source:[281]



A dual approach combining the Anatomical Therapeutic Chemical (ATC) classification system code and the approved therapeutic indication in the US Summary of Product Characteristics (SmPC) of galcanezumab was used for automatizing the selection of controls for disproportionality analysis when using FAERs. All active ingredients with the same therapeutic target (i.e. CGRP antagonists) as galcanezumab were identified using the 4th level of the ATC code, or rather the chemical subgroup. DrugBank was used to identify controls with the same approved therapeutic indication but with active ingredient outside the chemical subgroup of galcanezumab, aiming to avoid masking due to drug class effect and confounding by indication.

Disproportionality signals were further analyzed by using conditional inference trees to identify alternative cause co-reported drugs. The SmPC of disproportionally co-reported drugs was screened to identify those drugs that listed in the SmPC the AE in disproportionality signal mimicking procedures performed during signal validation. The disproportionality analysis was conducted again by removing cases with co-reported drugs for which the AE under investigation was listed in the SmPC as these cases had alternative causes for the AE.

F3. Results

AI could significantly ease some of the most time-consuming and labour-intensive steps of signal detection and validation. The AI-based approach showed promising results, however, future work is needed to validate the framework.

F4. Compliance with the governance framework

Table 16: Use case F: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	When developing tools to support signal detection developers should consider the overall impact the tool may have to the PV system within a QMS and consider the need for mitigations that maybe required to support broader deployment.	A	A	N/A	N/A	N/A
Human oversight	Human oversight within the proof-of-concept setting would ensure that the outputs of the conditional inference tree are robust for use in augmenting the quantitative signal detection processes. In a production setting it would include monitoring mechanisms and sampling approaches to ensure ongoing validity.	A	A	N/A	N/A	N/A
Validity & Robustness	The outputs of the conditional inference trees can be subject to existing validation procedures to compare inputs to known outputs.	A	A	N/A	N/A	N/A
Transparency	Transparency would involve clear communication and documentation of the AI system's use in the process of signal detection and validation from ICSRs. Within this proof of concept this includes detailed information about the business case, methodology and training datasets, should the tool move into a production setting further documentation and information relating to how the tool is integrated into the wider PV system and it's QMS would be required.	A	A	N/A	N/A	N/A
Data Privacy	The method uses publicly available information on labelling	N/A	N/A	N/A	N/A	N/A
Fairness & Equity	The method is not using individual patient data.	N/A	N/A	N/A	N/A	N/A
Governance & Accountability	During the proof of concept (PoC), the accountability of the methodology remains with the developer. However, if the methodology is integrated into a production setting, accountability would transition to the human subject matter expert. Governance within a PoC ensures that scientific integrity principles are adhered to, while future product use governance should cover how the tool fits into the overall PV system and quality management system (QMS).	A	A	N/A	N/A	N/A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable

3736 NA: Not Applicable

References

[280] *Al-Azzawi F, Mahmoud I, Haguinet F, Bate A, Sessa M. Developing an Artificial Intelligence-Guided Signal Detection in the Food and Drug Administration Adverse Event Reporting System (FAERS): A Proof-of-Concept Study Using Galcanezumab and Simulated Data. Drug Saf. 2023;Aug;46(8):743-751 (Journal full text) https://doi.org/10.1007/s40264-023-01317-0*

[281] *Al-Azzawi F, Mahmoud I, Haguinet F, Bate A, Sessa M. Developing an Artificial Intelligence-Guided Signal Detection in the Food and Drug Administration Adverse Event Reporting System (FAERS): A Proof-of-Concept Study Using Galcanezumab and Simulated Data. Drug Saf. 2023;Aug;46(8):743-751 (Journal full text) https://doi.org/10.1007/s40264-023-01317-0*

Use Case G: Generative artificial intelligence for enhanced and intelligently structured outputs from large pharmacovigilance document libraries

Source: Internal to CIOMS Working Group member organization

Area of PV: PV document retrieval

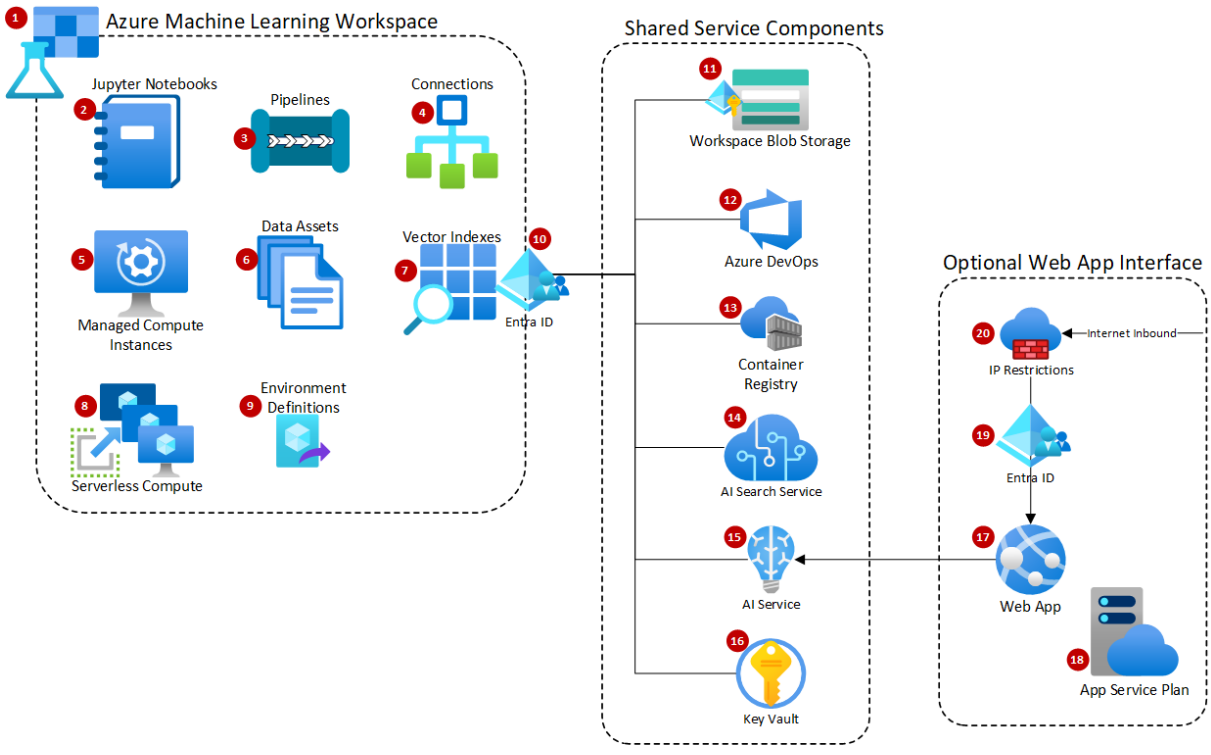
G1. Business rational and challenges

Use of AI represents a burgeoning field that has gained significant traction across various industries including the pharmaceutical industry. However, due to its relative novelty, concrete business use cases remain somewhat scarce, leaving many organizations searching for innovative ways to leverage AI technology effectively. LLMs, such as the Generative Pre-trained Transformer (GPT) models, offer a potentially valuable resource for businesses seeking historical references and streamlined document management solutions, with capabilities to summarize & synthesize large sources of data.

G2. Solution

Harnesses the power of LLMs to optimize our document filing structure. With hundreds of thousands of documents pertaining to patient safety and PV activities, in this use case an AI tool was employed to function as a powerful system for searching, extracting and generating information in an effective and structured manner specific to the parameters and requirements of the (human) user. By automating the process of retrieving relevant information, subject matter experts (SMEs) can redirect their time towards value-added endeavours rather than manual data sifting.

Figure 9: An outline of our initial artificial intelligence architecture
Source: Internal to CIOMS Working Group member organization



The implementation of LLMs not only aids day-to-day tasks but can also enhance efficiency in many different vigilance activities across all different verticals of vigilance. Currently this tool has the ability to help support audit and inspection requests, performative AI researcher for PV projects that require detailed and structured retrieval from safety document libraries, e.g. searching for regulatory / safety correspondence.

Development and deployment of a meticulously crafted vector index of relevant company file structures, allows us to leverage an LLM to easily and efficiently navigate documents, files and data available within those files. By transitioning to newer models as they release, the organization has observed improved accuracy and utility in responses, validated through a manual verification of processes.

Regular feedback sessions are conducted to refine the AI tool's performance and uncover additional use cases. This iterative approach ensures continual improvement and minimizes errors. Guidelines were also developed for framing questions to the model effectively, further enhancing the tool's usability.

Looking ahead, there are plans to expand this application of GenAI, focusing on areas such as summarization of safety literature for example signal detection purposes and summary outputs for case reporting. By training additional models on relevant datasets, including Individual Case Safety Reports (ICSR), the organization aims to automate safety summary generation and aggregate analysis, trending, any other ad hoc safety questions that can be answered from the data from or within the global safety database (GSD) thereby streamlining decision-making processes.

G3. Results

The generative AI tool in this use case has demonstrated potential as an AI 'research assistant' enabling PV workers to quickly and efficiently search hundreds and thousands of safety documents to provide structured and intelligent outputs. The adoption of AI technology like LLMs could empower organizations using this technology to optimize operational efficiency and prioritize tasks that impact product benefit-risk assessments. By leveraging AI-driven document management solutions, subject matter experts can devote more time to critical medical and scientific evaluations, ultimately enhancing product safety and efficacy. In addition, the planned expansion of use of LLMs in GenAI capabilities to evaluate transactional data and provide scientific analysis and summary has the potential to be a game changer in evaluating product or patient level safety risk in real time, simply by "asking a question" with well referenced and documented summaries/conclusions.

G4. Compliance with the governance framework

During the development and pre-deployment phases, the GenAI project is carefully developed and managed with a limited scope, ensuring full alignment with most guiding principles. As the project transitions into production and its exposure and scope expand, careful consideration will be given to maintaining close alignment with these principles. Therefore, compliance is indicated as closely aligned, laying a foundation of trust in the solution's ability to perform vigilance tasks with adaptive and growth capabilities.

Table 17: Use case G: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	GenAI use is a closed environment used for training and testing during development and pre-development. However, there is communication of potential inaccuracies and pitfalls during these phases. Currently there is no anticipated (patient risk) for post-deployment or routine use. As GenAI achieves more general and expanded risks will be regularly reassessed. Therefore, all phases are considered partially aligned with this guiding principle.	A	A	A	A	A
Human oversight	Fully aligned in development phase. Although there is still human oversight from the user,	A	A	A	A	A

	there is probably partial alignment during post deployment and routine use as GenAI would not be routinely scrutinized by human oversight and management					
Validity & Robustness	Validation and testing is extensive during development and pre-deployment and therefore in full alignment with the guiding principle (based on smaller sample set during these stages. Once in post deployment and routine use the data sets are very large and it is not possible to 100% fully test and validate all use cases (therefore in partial alignment). However, if users are noting inaccuracies in information retrieval, feedback will be provided to the human developers of the GenAI to refine and update the system.	A	A	A	A	A
Transparency	In full alignment during development and pre-development. As GenAI system expands in scope and complexity during post-deployment and routine use, further realignment is anticipated. Transparency in relation to the public is not applicable as this is a closed system	A	A	A	A	A
Data Privacy	Fully alignment during all phases. All of data remains internal; therefore, no transfer of information and external vendors/individuals.	A	A	A	A	A
Fairness & Equity	Fully aligned with this guiding principle at all phases. No relevant groups or individuals are being excluded or disadvantaged.	A	A	A	A	A
Governance & Accountability	Accountability from system usage and implementation during development and pre-deployment, e.g. if system is clearly not useful then it will be discontinued / upgraded. Ultimately regulatory accountability resides with subject matter expert / user as they are responsible to review and verify content. Therefore, partial alignment is anticipated from post-deployment onwards	A	A	A	A	A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

Use Case H: Artificial intelligence to support diagnosis and prediction of (hydroxy)chloroquine retinopathy

Source:[282,283]

Area of PV: Pharmacovigilance in The Clinic

H1. Business rational and challenges

PV in the clinic is concerned with the prevention and treatment of adverse drug reactions in individuals. Prevention may be primary, which can be achieved through identifying potential complex or non-obvious combinations of patient characteristics that are predictive of adverse drug reactions to guide optimum medication selection. (i.e. precision medicine) It also encompasses secondary and tertiary prevention (i.e. early diagnosis of adverse drug reactions, and ensuing interventions) to mitigate the impacts of ADRs. Examples follow.

Chloroquine and hydroxychloroquine are important drugs in rheumatology. Although relatively well tolerated compared to some other therapeutic options, retinal toxicity, is a risk which can result in serious visual impairment if not detected early so that the drug may be discontinued in a timely manner. Even so, by the time of retinopathy diagnosis there may be irreversible retinal damage. Conversely, if predictive AI can provide sufficient leading indicators or progression, therapy duration and attendant therapeutic benefits might be maximized. Historically the gold standard for screening and detection has been fundus photography and automated perimetry. More recently, multifocal electroretinography (mfERG) and optical coherence tomography (OCT) have been added to the diagnostic armamentarium. Each of these are routinely assessed by human readers, ideally retinal specialists, but subtle changes, including temporal patterns, can be missed, and not all locales have the necessary instrumentation or available retinal specialists. It would be ideal to augment human visual assessors to identify early functional changes indicative of retinopathy prior to onset of irreversibility or better predict progression. AI has shown potential in detecting or predicting various ocular diseases based on retinal images/fundus photography, such AMD, DR. More AI has been retrospectively developed and tested to diagnose or predict (hydroxy) chloroquine retinopathy.

H2. Solution

AI has been applied to colour fundus photographs , OCT and multifocal electroretinographic tracings for diagnosing hydroxychloroquine retinopathy, Fan et al studied hyperspectral imaging (HIS) of 176 fundus photographs from retinopathy positive (25) versus retinopathy negative (66) patients at a referral clinic using four deep learning models for the detection of retinopathy Kulyabin et al compared deep learning-based classification of raw mfERGs versus models based on conventional readout parameters of the mfERG for classification, and for prediction (regression) of visual field sensitivities from 53 predominantly female patients (35 retinopathy negative, 9 minimal retinopathy, and 9 manifest retinopathy) monitored with mfERGs and perimetry for period of 0.7-20,9 years. Kalra et al used random forests for automated diagnosis and prediction of disease progression using clinical features and features based on spectral domain OCT (SD-OCT) obtained from 388 eyes / 368 patients, a majority being female. Habib et al trained support vector machines (SVM) to identify hydroxychloroquine retinopathy in 1463 eligible eyes (748 predominantly female patients), of which 95 eyes (48 patients) were eligible for inclusion as controls.

H3. Results

The best performing deep learning models in the study of Fan et al achieved accuracy, precision, recall, specificity, and F1-scores of ≥0.95., with superior performance using hyperspectral images versus the original retinal images. Habib et al's SVM returned a specificity of 84.0% with sensitivity of 90.9%. Performance could be calibrated to place a premium on sensitivity for screening or specificity for diagnosis. Kalra reported a mean AUC of 0.97, a sensitivity 95% and specificity of 91% for detection, and mean AUC=0,89, recall of 90% ad specificity of 80% for progression prediction.

Kulyabin reported that AI-based models using full mfERG traces had a balanced accuracy of up to 0.795, precision of up to 0.844, recall of up to 0.866, and F1-score of up to 0.771.

H4. Compliance with the governance framework

In considering the alignment of the reviewed studies with the governance framework we note several points up front. The studies were retrospective and feasibility/pilot studies, without reported advancement to routine use in the clinic. Because the drugs under study are for autoimmune disorders, in addition the study populations often being mall, the subjects were predominantly female, consistent with the treatment indication. The use of eyes as the unit of observation raises the question of pseudo-replication and its potential impacts of performance estimates, though confidence intervals were not typically presented. At least one study noted under-representation of Asian in the study sample and the need for further assessment in larger and more diverse populations. Finally as is often the case in AI applications involving retinal pathology, other ocular pathologies were excluded, which limits generalizability to more diverse patient populations that have multiple retinal pathologies (e.g. diabetic retinopathy and drug-induced retinopathy).

Table 18: Use case H: Alignment with the governance framework (detail)
Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	Not aligned. Risk assessment and risk mitigation plans not provided in these pilot studies. But placement within a human-in-the-loop framework was explicitly considered in one or more studies	N/A	N/A	N/A	N/A	N/A
Human oversight	Partial alignment. One or more of the publications, which report feasibility/pilot studies in clinical settings, discuss the proper deployment with respect to human oversight, such as HITL. However, change management and staff training plans not discussed. Discussed is the fact that the available human oversight in some locations may be provided by generalists with less experience and expertise that retinal specialists, affording more opportunity for incremental benefits in underserved settings.	A	A	N/A	N/A	N/A
Validity & Robustness	Partial alignment. Reference standards defined. One or more studies note the limitation of the imbalanced data sets used that impair generalizability. Also, in one/more studies patients with other ocular pathology excluded so the two classes were HCQ retinopathy present versus normal retina, which limits generalizability to screening in patients with other coexistent ocular disorders that may affect the retina. Source population (deployment domain) not clearly defined in all studies No discussion of integrating data pre-processing (e.g. cropping retinal images) into routine use). In some studies unit of observation was “eyes” raising questions about pseudo-replication.	A	A	A	N/A	N/A
Transparency	One/more papers report adherence to tenets of the Declaration of Helsinki and obtained Institutional Review Board approval. One or more papers described explanations of results such as heatmaps of feature distributions.	A	A	A	N/A	N/A

Data Privacy	One or more of the referenced studies declared adherence to tenets of the Declaration of Helsinki and obtained Institutional Review Board Approval.	A	A	A	N/A	N/A
Fairness & Equity	One/more of the referenced studies report adherence to tenets of the Declaration of Helsinki and obtaining Institutional Review Board approval. One or more of papers acknowledge that data under-represents specific groups of persons such as Asians, who may display different findings and recommends further assessment with larger data sets with more diverse representations. Further discussion involved scenarios in which retinal specialists may not be available, such as under-resourced or under-represented locales, as also discussed in human oversight above.	A	A	A	N/A	N/A
Governance & Accountability	These studies which occurred in clinical settings were conducted according to the guidelines of the Declaration of Helsinki and approved by the respective Institutional Review Board o	A	N/A	N/A	NA	N/A

Abbreviations
SPEC: Collection of specifications, requirements
DEV: Development and change management
PreD: Pre-deployment & post-change sign-off
PstD: Post-deployment & post-change hyper-care
RU: Routine Use
A: Applicable
NA: Not Applicable

References

[282] *Kulyabin M, Kremers J, Holbach V, Maier A, Huchzermeyer C. Artificial intelligence for detection of retinal toxicity in chloroquine and hydroxychloroquine therapy using multifocal electroretinogram waveforms. Scientific Reports. 2024;Oct22;14(1);24853. (Journal full text) https://doi.org/10.1038/s41598-024-76943-4*

[283] *Kalra G, Talcott KE, Kaiser S, Ugwuegbu O, Hu M, Srivastava SK, Ehlers JP. Machine learning–based automated detection of hydroxychloroquine toxicity and prediction of future toxicity using higher-order OCT biomarkers. Ophthalmology Retina. 2022;Dec1;6(12):1241-1252. (Journal abstract)https://doi.org/10.1016/j.oret.2022.05.031*

APPENDIX 4: Content related to explainability and to Fairness & Equity

Content related to explainability

Illustrative examples

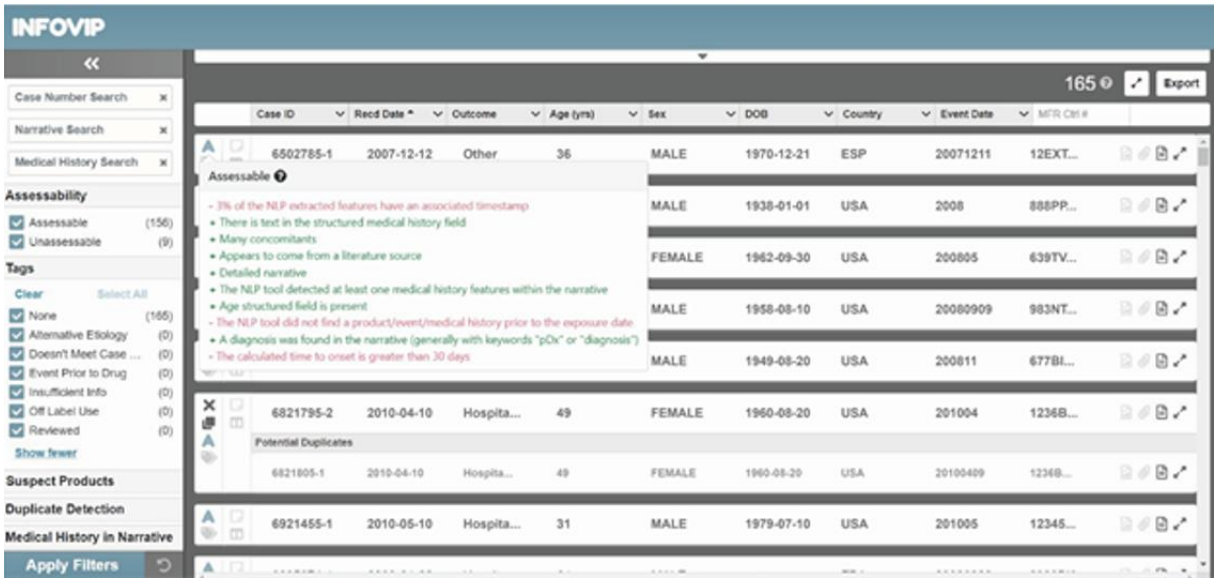
As stated above, diverse stakeholders in PV may require and benefit from explainability. However, the explainability that is required and how it is used, differs depending on the setting, i.e. who is asking and for what, in which task or process and in which system lifecycle stage.[284] In the following sections, examples are provided below to illustrate the different scenarios. Finally, the benefits of explainability described in the example are summarised at the end of this section.

Examples of explainability in artificial intelligence-supported pharmacovigilance tasks

Take for example a setting in which a PV officer is reviewing a case that has been selected by the AI as a potential signal but does not immediately see why the case was flagged. In such cases, the reviewer could benefit from being able to see which text in the case data led the AI to this conclusion. An actual example of this is described below.

The Information Visualization Platform (InfoViP) developed for the FDA's Center for Drug Evaluation and Research is an example of explainability supporting the human expert engaged in signal detection and assessment.[285] InfoViP uses Natural Language Processing (NLP) and several other components to process post-market data and visualize information, i.e. explanations, to support medical reviewers who detect and evaluate potential signals from the millions of adverse event reports submitted to the FDA's FAERS database. The NLP component, the Event-based Text-mining of Health Electronic Records (ETHER), coupled with modern frontend techniques, provide visual information by colour-coded highlighting of relevant text in the case narrative to help reviewers focus on signal-related information. An informed model further identifies cases containing enough information to assist reviewers assess the report quality and provides concrete explanations of these selections. All these functionalities combined with case deduplication and several filtering options facilitate speedy review by the medical reviewers, an otherwise humanly impossible task across millions of reports.[286]

Figure 10: FDA's Information Visualization Platform user interface illustrates the system capabilities, focusing on the features that positively contribute to classification for assessability.
Source:[287]



The example above illustrates a core benefit of explainability described by Albahri et al (2023).[288] Explainability can facilitate human experts in making 'sound and reliable' decisions. And ultimately, when the human decision and the accompanying explanation are retained, this information would nurture trust of the system owner and quality assurance staff who are tasked with ensuring compliance as well as the trust of regulators who may wish to inspect why certain cases are selected or rejected as signals.

Also, it is conceivable that explanations could lead a user to notice a bias or spurious correlation that is leading to incorrect predictions. Reporting this back to the development team can contribute towards future improvement. In this way, explainability is useful for ongoing vigilance against bias risk and performance issues that may appear post-deployment and for continually ensuring the trustworthiness of the decisions made. As a result, XAI explanations have resulted in increased trust and the perception of fairness in AI-supported decision making.[289]

Examples of pharmacovigilance stakeholders benefitting from explainability

While the likelihood of an individual from the general public requiring explainability in a PV setting is considered to be small, the possibility of this happening cannot be excluded entirely as the use of AI becomes more commonplace. Some conceivable scenarios are described below:

- When a reporter (healthcare professional or patient) directly reports a serious adverse event and the report is subsequently processed as a non-serious case by an AI triage system, upon being informed of this, the reporter undergoes a negative experience and may request an explanation from the MAH. When this scenario takes place in a non-AI setting or an AI-assisted triage setting, the reporter could receive an explanation from the PV officer who has made the final triage decision. However, when this takes place in a fully-automated AI triage process, unless some form of explainability is provided, the lack of explainability may impact trust and acceptance of the result by the reporter.
- When a consumer or a healthcare professional interacts with a medical information chatbot used by the MAH to provide drug product information and collect safety data and quality complaints, the individual may question and challenge any information that doesn't make sense.
- GenAI could be used in assisting a pharmacist in medication therapy management to prevent drug interactions. Both the pharmacist and the patient are directly exposed to the AI's

recommendations in this case.[290] Here, questions concerning the AI recommendations could be raised by both parties.

Examples of explainability in system development

A developer who is training the AI system benefits from explainability when it reveals which features are used by the AI to reach a specific prediction or when it reveals a bias in the training data. Especially in complex systems which lack inherent interpretability and explainability, an XAI tool could provide explanations that facilitate troubleshooting by revealing what to change or exclude in order to 'flip' the outcome.[291] Using the XAI explanation, one could discover that needle in the deep neural network haystack. However, in most cases, tweaking the system architecture of a deep neural network or specific features will be quite challenging even when XAI points the way. The XAI output is more likely to identify hidden biases in the training data which can be corrected as illustrated in the example below.

An example presented by Ribeiro et al (2016)[292] demonstrates how the XAI explanation provided by using Local Interpretable Model-Agnostic Explanations LIME could be understood by humans and reveal the likely cause of incorrect predictions. In this experiment, the model that was trained to distinguish images of dogs and wolves was first intentionally trained to associate wolves with snowscapes. In other words, the training data was deliberately biased by excluding images of wolves in other seasons. This resulted in predictions that included a wolf against a green background identified as a dog and a husky in a snowscape identified as a wolf. LIME was used to show subjects which areas of the image were used as features by the AI in its predictions to see if the subjects could identify the cause of the misidentification. The subjects successfully identified background snow as the potential feature that led the AI to make the incorrect predictions. Thus, XAI can be used to explain a prediction made by an inscrutable deep neural network and uncover the underlying issue in the training data and the resulting spurious correlation that led to the incorrect output.

In the context of PV, similar techniques could be used to highlight words in the text which are picked up by the AI as relevant features. In a real-life but unpublished example in which an AI triage system was misidentifying some serious cases, developers benefited from seeing which terms in the case were considered by the AI in its seriousness predictions. In this case, the XAI explanation revealed a focus on the drug name. Combined with the fact that the missed serious cases concerned Over The Counter (OTC) drugs, the developers discovered that the AI was basing decisions on the drug name and a learned spurious assumption that OTC drugs are not likely to cause serious events. Using the insight gained from explainability, the developers could reject the model in favour of another one, examine the training data for bias such as the lack of serious OTC cases or when there is no bias, solve the issue through feature engineering by instructing the AI not to consider the drug name in its decisions.

Explainability, therefore, can help developers make informed decisions when assessing AI models by uncovering hidden biases as well as features and spurious correlations that are resulting in incorrect predictions. Explainability may also reveal the underlying factors that result in performance differences between models that are trained on the same training data and aid the developer in model selection. In turn, transparent documentation of this process will go a long way towards nurturing trust in the system, not only for the developers but also for the system owners, users and the regulators.

Examples of explainability in artificial intelligence-systems interacting with health care professionals and patients

Another hypothetical example can be the case of a healthcare professional (HCP) who is requesting product-specific information via a chatbot provided by a marketing authorisation holder (MAH). Such a chatbot could have multiple objectives ranging from the provision of drug product information to the collection of adverse event and quality defect reports. When the HCP notices that the chatbot response is inadequate, i.e. not considering key medical terms or adverse events, or providing

questionable information, the HCP may contact the MAH for an explanation. Explainability, if it is available, may help the MAH troubleshoot the system and/or provide information back to the HCP.

While the scenario above is a fictive example, one example of a chatbot that is currently available is the Smart AI Resource Assistant for Health (SARAH) on the World Health Organization website. This is a prototype chatbot that is intended to provide tips on health topics and not medical advice as clearly stated on the landing page of SARAH.[293] On one hand, SARAH exemplifies how such an application could be of service to the public as it is available 24/7 and in eight languages. On the other hand, incidents of the chatbot providing inaccurate or incorrect information or being unable to answer some queries have unfortunately been reported in the media and taken up in the OECD AI incidents monitoring database.[294] This illustrates how, when a chatbot is deployed, the interacting patient or healthcare provider or the media may challenge the information that is provided. It is therefore conceivable that in PV, when a MAH deploys an AI system that interacts directly with the public, explainability for the public will also be required.

Finally, any system that interacts directly with the public in a medical setting warrants extra attention in that a HCP is likely to notice medically incorrect information, but most consumers and patients may not be able to do this. Individuals without a medical background will be at risk of accepting and acting on medically incorrect information. To illustrate this point, in a study of trust and medical advice provided by ChatGPT, persons without a medical background have been found to trust the chatbots for lower-risk health topics.[295] Without the medical background, a layperson is at increased risk of harm by not being able to recognize incorrect information. Thus, aside from being able to provide an explanation to an individual from the public who is challenging the AI output, system owners must thoroughly consider and mitigate the risks of an AI system that interfaces with the general public. This also touches on the subject of accountability since it is not the chatbot that is held accountable for any harm that befalls the individual.

Example where explainability is not available and not required

To illustrate a situation in which explainability is not necessary nor possible, consider first how the use of publicly available machine translation tools is now commonplace and how the public generally do not require explanations into how the AI translated the text. Consider also how translations in a GxP-regulated environment require a quality check regardless of whether the translation was carried out by a human or a machine. Furthermore, the quality check is normally carried out by an individual who is proficient in both the source and destination languages. Therefore, not only will the human be able to spot errors, but also in some cases, understand the underlying reasons for machine translation errors even without any explanation. An example of a translation issue with a self-evident root cause is the case of biased gender assignment that occurs when translating a genderless language such as Finnish to English.[296] The bilingual human reviewing the translations would easily notice the gender bias, understand why this has been introduced by the AI and can correct this accordingly. See also chapters x on human oversight, data privacy and fairness and equity.

Example where explainability could be available but is not required

In PV, another example of AI use which would not require explainability would be automatic de-identification of case narratives presented by Meldau et al 2024.[297] Automatic redaction of case narratives could ensure data privacy while allowing case narratives sharing which are crucial for providing more complete information to signal detection assessors. In this case, a system using a neural network combined with hand-engineered rules was trained to automatically detect names in case narratives for the purpose of redaction. Using a test set of over 5000 Yellow Card narratives from the MHRA and over 500 open source annotated and unannotated deidentified patient discharge summaries from i2b2.org, the system was able to identify 96% of names longer than three characters and 88% of all names for redaction. The false positive rate was 0.2%. It is not conceivable that stakeholders would require explanations during routine use of such a system. Even if stakeholders notice certain names are not being identified, it is conceivable that the reasons could be self-evident. Taking this system as an example, since it is trained to detect names that consist of

three letters or more, it may miss two-letter names which are common in Asia. When such a lapse is noticed, the developers and stakeholders would not require an explanation in order to take the necessary corrective measures such as re-training using a new dataset or adding new rules.

Example of recommended explainability combined with downstream human oversight

Even when there is downstream human quality control and the final decision rests with the human, explainability remains useful for supporting human decision and nurturing trust in the machine and in the decision that is made.

While the machine translation example above describes one end of the spectrum where explainability is not needed when risks are adequately covered human review, for AI in PV and other GxP environments, it is not conceivable to deploy a system that expects the human to make decisions based on AI predictions without any explanation. After all, the human needs to understand why the suggestion presented by the AI is trustworthy. Explainability is therefore essential for making a well-informed decision. The FDA's InfoVIP system (see above) demonstrates how a medical officer can make an informed decision supported by XAI visualisation of relevant features.

Another example from the WHO Collaborating Centre for International Drug Monitoring illustrates how explanations can be used in reviewing AI case deduplication. In this case, a pair of Norwegian cases were identified as probable duplicates, while on first glance for the reviewers the duplication was not obvious. Due to missing outcomes, onset dates and ages that were close but not matching, and no matches between the registered adverse drug reaction terms, these cases would not have been identified as duplicates if they were not flagged by the AI. The AI explanation revealed that the match score was based on six different drug substances which were identical between the cases in addition to the fact that these six drug substances are commonly not co-reported. The researchers verified that the cases were indeed duplicates by contacting the Norwegian national centre. The cases concerned the same incident but were reported by two different physicians from the same hospital, thus accounting for the differences.[298] This example illustrates how the explanation could be essential for the human in order to understand, further investigate the veracity of the prediction and finally reject or in this case accept the prediction.

Examples of xAI methods

Some xAI methods in use at the time of writing this report include:

- Local Interpretable Machine-Agnostic Explainability LIME

  See the example of Ribeiro et al (2016)[299] described earlier in this chapter.
- Shapley Additive exPlanations SHAP

An example of SHAP explainability in a supervised ML model used to support signal validation is presented by Imran et al (2024).[300]

- Trust scores that indicate the model's uncertainty for the output.[301]
- Confidence scores are a metric that is usually available and can be used to flag output that is uncertain for human review.[302]
- Visualisation through highlighting of text that was considered by the AI in its prediction and saliency maps using a heat map overlay to indicate areas of the input image that are relevant for the model's prediction.

Although assessing and processing images is not a mainstream activity in PV, saliency maps are mentioned as another example of XAI to complete the view of the current landscape. See examples in Plass et al (2023).[303]

In the case of PV where the data is predominantly text based, visual explanations are likely to take the form of highlighting relevant text within the case data. See also the FDA InfoViP described above as an example of explainability benefits.[304,305]

- Counterfactuals explanations or examples that show what characteristics in the input data when changed, would result in different output.[306]

References

[284] Royal Society. Explainable AI. Royalsociety.org. 2024. (Website accessed 11 August 2024)

[285] Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, Gish P, Lardieri A, Wu E, De S, Naidoo J. Information visualization platform for postmarket surveillance decision support. Drug Safety. 2020;Sep;43:905-915. (Journal abstract) https://doi.org/10.1007/s40264-020-00945-0

[286] Botsis T, Dang O, Kreimeyer K, Spiker J, De S, Ball R. A Decision-Support Platform Powered by AI and Humans-in-the-Loop Boosts Efficiency and Assures Quality in FDA's Pharmacovigilance. International Society of Pharmacovigilance, 23rd Annual Meeting 2024, Montreal, Canada. (Webpage accessed 21 March 2025)

[287] Botsis T, Dang O, Kreimeyer K, Spiker J, De S, Ball R. A Decision-Support Platform Powered by AI and Humans-in-the-Loop Boosts Efficiency and Assures Quality in FDA's Pharmacovigilance. International Society of Pharmacovigilance, 23rd Annual Meeting 2024, Montreal, Canada.

[288] Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion. 2023;Aug1;96:156-191. (Journal full text) https://doi.org/10.1016/j.inffus.2023.03.008.

[289] Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion. 2023;Aug1;96:156-191. (Journal full text) https://doi.org/10.1016/j.inffus.2023.03.008.

[290] Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the Role of Artificial Intelligence in medication therapy management. J Am Pharm 2023;Dec 2;64(2);422-428. (Journal full text) https://doi.org/10.1016/j.japh.2023.11.023

[291] Hauben M. Artificial intelligence in pharmacovigilance: Do we need explainability?. Pharmacoepidemiology and Drug Safety. 2022;Dec;31(12):1311-1316. (Journal full text) https://doi.org/10.1002/pds.5501

[292] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Internet]. arXiv:1602.04938 2016. (Journal full text) https://doi.org/10.48550/arXiv.1602.04938

[293] World Health Organization. Using AI to lead a healthier lifestyle. (Website accessed 21 March 2025).

[294] Organisation for Economic Co-operation and Development. OECD Legal Instruments. Recommendation of the Council on Artificial Intelligence. Oecd.org. 2019. (Webpage accessed 21 March 2025)

[295] Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. JMIR Medical Education. 2023;Jul10;9:e46939. (Journal full text) https://doi:10.2196/46939

[296] Savoldi B, Gaido M, Bentivogli L, Negri M, Turchi M. Gender bias in machine translation. Transactions of the Association for Computational Linguistics. 2021;Aug;18;9:845-874. (Journal full text) https://doi.org/10.1162/tacl_a_00401

[297] Meldau EL, Bista S, Melgarejo-González C, Niklas Norén G. WHO Uppsala Monitoring Centre. Automated De-identification of Case Narratives Using Deep Neural Networks for UK Yellow Card [Internet]. (Full text accessed 21 March 2025)

[298] Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Mining and Knowledge Discovery. 2007;Jun;14:305-328. (Journal full text) https://doi.org/10.1007/s10618-006-0052-8

[299] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Internet]. arXiv:1602.04938 2016. (Journal full text) https://doi.org/10.48550/arXiv.1602.04938

[300] Imran M, Bhatti A, King DM, Lerch M, Dietrich J, Doron G, Manlik K. Supervised machine learning-based decision support for signal validation classification. Drug Safety. 2022 May;45(5):583-596. (Journal full text) https://doi.org/10.1007/s40264-022-01159-2

[301] Plass M, Kargl M, Kiehl TR, Regitnig P, Geißler C, Evans T, Zerbe N, Carvalho R, Holzinger A, Müller H. Explainability and causability in digital pathology. The Journal of Pathology: Clinical Research. 2023;Jul;9(4):251-260. (Journal full text) https://doi.org/10.1002/cjp2.322

[302] Royal Society. Explainable AI. Royalsociety.org. 2024. (Website accessed 11 August 2024)

[303] Plass M, Kargl M, Kiehl TR, Regitnig P, Geißler C, Evans T, Zerbe N, Carvalho R, Holzinger A, Müller H. Explainability and causability in digital pathology. The Journal of Pathology: Clinical Research. 2023;Jul;9(4):251-260. (Journal full text) https://doi.org/10.1002/cjp2.322

[304] Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, Gish P, Lardieri A, Wu E, De S, Naidoo J. Information visualization platform for postmarket surveillance decision support. Drug Safety. 2020;Sep;43:905-915. (Journal abstract) https://doi.org/10.1007/s40264-020-00945-0

[305] *Botsis T, Dang O, Kreimeyer K, Spiker J, De S, Ball R. A Decision-Support Platform Powered by AI and Humans-in-the-Loop Boosts Efficiency and Assures Quality in FDA's Pharmacovigilance. International Society of Pharmacovigilance, 23rd Annual Meeting 2024, Montreal, Canada. (Webpage accessed 21 March 2025)*

[306] *Royal Society. Explainable AI. Royalsociety.org. 2024. (Website accessed 11 August 2024)*

Content related to Fairness and Equity

While not all of the examples provided below are specific to PV, they illustrate the potential

Impact of inadequate data, bias from underrepresented populations and explicit bias potentially leading to unfair treatment of specific populations, underserved populations, and potential treatment inequality.

Example of inadequate training of AI solutions and/or inadequate data sets that introduced unfair bias and resulted in inequity.

In the US, prescription opioids are tracked through electronic databases, Prescription Drug Monitoring Programs (PDMPs). While not a PV specific example, Bamboo Healths NarxCare® is an example of an AI-powered tool that leverages PDMPs to calculate an opioid risk metric to predict the likelihood of a potential overdose, and although it is intended to support medical decisions, there have been observations that patients who are high health care utilizers with complex medical conditions may be discriminated against and underserved for pain management because of a "high risk score".[307] The score is calculated based on limited data available in the PDMP and does not consider any other factors when calculating the risk score. One factor that influences the score is the number of prescribers. Patients treated at teaching hospitals with multiple healthcare prescribers may have "too many prescribing physicians" and they may be interpreted as seeking treatment from multiple physicians to obtain multiple prescriptions. An April 2021 study in Drug and Alcohol dependence found that "common data driven algorithms" misclassified 20% of patients with cancer who often see multiple specialists as patients seeking multiple physicians in an effort to obtain multiple opioid prescriptions. As noted by the authors, the PDMP data lacks diagnostic information and other critical patient context limiting ability to distinguish misuse from appropriate clinical use. An October 2021 study published in Drug and Alcohol Dependence conducted an independent validation study found that the NarxCare tool had a 17.2% false positive and 13.4% false negative.[308]

Bias introduced because of data limited to PDMP, interpretation of data, e.g. patients with complex medical conditions, multiple prescribers are perceived as a high risk for abuse, lack of context for patient population, e.g. cancer patients requiring prolonged opioid use, can influence high scores. The threat to fairness and equity for patients within subgroups who have a high score assigned because of bias, potentially may not receive adequate pain management when the high score is considered in isolation.

Within PV, the risk to fairness and equity are primarily from explicit biases that may result in negative impact or may result in discriminatory harm to subpopulations underserved by an AI solution. The NarxCare example, while not PV related, demonstrates both explicit bias from inadequate data, lack of context and implicit bias because negative stereotypes associated with "high health care utilizers" were applied.[309]

Example of bias applied because of under-represented populations

In Brazil, the assertiveness outcomes of the skin's lesions classification using artificial neural network in Caucasian patients and Brazilian patients were compared. The skin lesions were classified using basic architecture of convolutional neural networks (CNN). The ISIC database was used to train the neural network. This database was applied in about 25 thousand images of skin lesions. These images have included melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma lesions. The tests

performed with ISIC patients had accuracy rates close to 90%. However, the accuracy rate was less than 40% when the tests were carried out with Brazilian patients. Thus, there was a great discrepancy in the assertiveness of the Artificial Neural Network when applied to Caucasian patients and Brazilian patients (Ref. K. Mundim et al).

Example of Explicit negative bias

In Appendix 3, "Examples of explainability in system development" included an example describing an AI triage system that incorrectly identified serious cases. The AI solution incorrectly learned to predict any adverse events associated with an OTC drug of interest as being non-serious because serious events were under-represented in the training data. This can also be considered an example of explicit negative bias with the inadequate date set resulting in incorrect assessments not recognized because of an explicit bias that it was not likely the OTC products in question would have serious adverse events associated with the use of the products. Since populations that may not have the same means to seek treatment at a medical facility or access to a healthcare professional may be reliant on OTC products, and these groups have a high likelihood of being from minority groups, a systematic misclassification of serious reports for OTC products as being non-serious could be seen as a threat to fairness and equity.

References

[307] *Sammer MBK, Akbari YS, Barth RA, Blumer SL, Dillman JR, Farmakis SG. Use of Artificial Intelligence in Radiology: Impact on Pediatric Patients, a White Paper from the ACR Pediatric AI Workgroup. J Am Coll Radiol. 2023;Aug;20(8):731-742. (Journal full text) https://doi.org/10.1016/j.jacr.2023.06.003*

[308] *Cochran G, Brown J, Yu Z, Frede S, Bryan MA, Ferguson A, et al. Validation and threshold identification of a prescription drug monitoring program clinical opioid risk metric with the WHO alcohol, smoking, and substance involvement screening test. Drug Alcohol Depend.2021;228:109067. (Journal full text) https://doi:10.1016/j.drugalcdep.2021.109067*

[309] *Siegel Z. In a world of stigma and bias, can a computer algorithm really predict overdose risk? Ann Emerg Med. 2022;Jun;79(6):A16-A19. (Journal full text) https://doi.org/10.1016/j.annemergmed.2022.04.006*

APPENDIX 5: CIOMS Working Group membership and meetings

The CIOMS Working Group XIV on *Artificial intelligence in* pharmacovigilance included the following groups of stakeholders: academics, pharmaceutical companies, regulatory authorities, as well as national and international organisations.

Academia		
Name	**Company/Organisation**	**Country**
Altman, Russ	Stanford University	USA
Botsis, Taxiarchis	Johns Hopkins University School of Medicine	USA
Dogné, Jean-Michel	University of Namur	Belgium
Pharmaceutical companies		
Name	**Company/Organisation**	**Country**
Amelio, Justyna	AbbVie	UK
Barrios, Mariane	Merck Sharp & Dohme	Colombia
Bate, Andrew	GSK	UK
Bellur, Arvind	CSL Behring	USA
Berridge, Adrian	Takeda Development Center Americas, Inc	USA
Carroll, Hua	Biogen	USA
Cherkas, Yauheniya	Johnson & Johnson	USA
Comfort, Shaun	Genentech	USA
Cooper, Selin	AbbVie	UK
Diniz, Mariane	Bayer	Brazil
Domalik, Douglas	AstraZeneca	UK
Franco, Piero Francesco	Pfizer	Italy
Girod, Julie	Sanofi	USA
Grabowski, Neal	Sanofi	USA
Hauben, Manfred	Merck KGaA, Darmstadt, Germany	USA
Henn, Thomas	United Therapeutics	USA
Kara, Vijay	GSK	UK
Kempf, Dieter	Genentech	USA
Kidos, Kostadinos	Formerly Takeda Development Center Americas, Inc	USA
Lorenz, Denny	formerly Bayer AG	Germany

MacEntee Pileggi, Beth	Johnson & Johnson	USA
Patel, Ravi	United Therapeutics	USA
Reinhard Pietzsch, John	Bayer	Germany
Römming, Hans-Jörg	Merck KGaA, Darmstadt, Germany	Germany
Savage, Elizabeth	Johnson & Johnson	USA
Straus, Walter	Moderna	USA
Whitehead, James	AstraZeneca	UK
Regulatory authorities		
Name	**Company/Organisation**	**Country**
Ball, Robert	United States Food and Drug Administration (US FDA)	USA
Buch, Brian	Medicines and Healthcare products Regulatory Agency (MHRA)	UK
Durand, Julie	European Medicines Agency (EMA)	
Egebjerg Juul, Kirsten	Danish Medicines Agency (DKMA)	Denmark
Harrison, Kendal	Medicines and Healthcare products Regulatory Agency (MHRA)	UK
Hirokawa-Voorburg, Satoko	Health and Youth Care Inspectorate (HYCI)	The Netherlands
Horst, Alexander	Swissmedic	Switzerland
Jensen, Morten	Danish Medicines Agency (DKMA)	Denmark
Kikuchi, Yuki	Pharmaceuticals and Medical Devices Agency (PMDA)	Japan
Kjær, Jesper	Danish Medicines Agency (DKMA)	Denmark
Ling, Benny	Health Canada	Canada
Da Luz Carvalho Soares, Monica	Brazilian Health Regulatory Agency (ANVISA)	Brazil
Maniwa, Harumi	Pharmaceuticals and Medical Devices Agency (PMDA)	Japan
Matsunuga, Yusuke	Pharmaceuticals and Medical Devices Agency (PMDA)	Japan
McAteer, Richard	Health Canada	Canada
Mentzer, Dirk	Paul-Ehrlich-Institut (PEI)	Germany
Messelhäußer, Manuela	Formerly Paul-Ehrlich-Institut (PEI)	Germany
Moreira Cruz, Flávia	Brazilian Health Regulatory Agency (ANVISA)	Brazil
Perez, Nicolas	Swissmedic	Switzerland
Scholz, Irene	Swissmedic	Switzerland
Stammschulte, Thomas	Swissmedic	Switzerland

Tregunno, Phil	Medicines and Healthcare products Regulatory Agency (MHRA)	UK
National and international organisations		
Name	**Company/Organisation**	**Country**
Mathur, Roli	Indian Council of Medical Research	India
Meldau, Eva-Lisa	Uppsala Monitoring Centre/World Health Organization	Sweden
Norén, Niklas	Uppsala Monitoring Centre/World Health Organization	Sweden
Rosenfeld, Stephen	North Star Review Board	USA
Yau, Brian	World Health Organization	Switzerland
CIOMS		
Name	**Company/Organisation**	**Country**
Heaton, Stephen	Individual expert	Germany
Hill, Sanna	CIOMS	Switzerland
Le Louët, Hervé	CIOMS	Switzerland
Rägo, Lembit	CIOMS	Switzerland
Rannula, Kateriina	CIOMS	Estonia
Tsintis, Panos	CIOMS	UK

The Working Group XIV will have met ten times from 2022 to 2025 at the time of writing and most of the meetings at this stage will have been hybrid in nature.

1. Geneva, Switzerland 18-19 May 2022
2. Geneva, Switzerland 10-11 October 2022
3. Virtual meeting 19 January 2023
4. Virtual meeting 12 April 2023
5. Zurich, Switzerland 6-7 June 2023
6. Virtual meeting 8 November 2023
7. Virtual meeting 11 January 2023
8. Geneva, Switzerland 7-8 March 2024
9. Darmstadt, Germany 24-25 September 2024
10. *Geneva, Switzerland* *25-26 June 2025 (planned at time of writing)*

APPENDIX 6: Public consultation commentators

(to follow)