Council for International Organizations of Medical Sciences (CIOMS)

Artificial intelligence in pharmacovigilance

Report of the CIOMS Working Group XIV

Geneva 2025

Council for International Organizations
of Medical Sciences (CIOMS)

Artificial intelligence in pharmacovigilance

Report of the CIOMS Working Group XIV

Geneva 2025

DOI number https://doi.org/10.56759/cdob6397
Licence: CC BY-NC-SA 4.0.

ISBN 978-929036110-7

Suggested citation: Artificial intelligence in pharmacovigilance. CIOMS Working Group XIV report. Geneva, Switzerland: Council for International Organizations of Medical Sciences (CIOMS), 2025.

Further information can be obtained from CIOMS, P.O. Box 2100, CH-1211 Geneva 2, Switzerland, www.cioms.ch, e-mail: info@cioms.ch.

Design: Agence Gardeners, Annecy (France).

ACKNOWLEDGEMENTS

The Council for International Organizations of Medical Sciences (CIOMS) is thankful to the experts from academia, industry, international and national pharmacovigilance (PV) organisations and medicinal product regulators, including ICH founding members, who contributed their valuable time and vast experience to create this consensus report.

Special thanks go to those members who led meetings and undertook active roles; while valuing all sizes of input in the discussions, drafting, re-drafting and reviewing of the report. We gratefully acknowledge the chapter and section leads and co-leads in particular: Hua Carroll, Julie Durand, Manfred Hauben, Thomas Henn, Satoko Hirokawa-Voorburg, Vijay Kara, Benny Ling, Denny Lorenz, Elizabeth MacEntee Pileggi, Niklas Norén, Ravi Patel, and Walter Straus.

The Working Group (WG) XIV Editorial Team merits mention with additional appreciation to Elizabeth MacEntee Pileggi for co-ordinating the team's work. CIOMS would like to extend kind thanks to others in this team too: Justyna Amelio, Andrew Bate, Taxiarchis Botsis, Hua Carroll, Julie Durand, Manfred Hauben, Thomas Henn, Vijay Kara, Benny Ling, Denny Lorenz, Lembit Rägo, Monica Da Luz Carvalho Soares, Niklas Norén, Ravi Patel, and Walter Straus. CIOMS is also much obliged to Cheryl Renz for proofreading the report.

The WG had an invaluable Glossary Team, headed by Douglas Domalik, that helped to define many of the terms in this field, thereby making the report accessible to a wide readership. The Glossary Team members included Neal Grabowski, Thomas Henn, Vijay Kara, Benny Ling, and Manuela Messelhäuser.

The WG invested considerable time into developing a series of use cases (see Appendix 3) to demonstrate the applicability of artificial intelligence in PV, and CIOMS wishes to thank the members of this team: Vijay Kara, Hans-Jörg Römming, Taxiarchis Botsis, Adrian Berridge, and Manfred Hauben.

CIOMS appreciates the contributions of the many reviewers (see Appendix 6), who participated in the Public Consultation of the draft report when it was hosted online on the CIOMS website during nearly six weeks in 2025, and which was promoted by the WG's extensive networks. All comments and suggestions were gratefully received and considered.

The WG met for a number of in-person meetings and CIOMS warmly welcomed the generosity of Takeda Pharmaceutical Company Ltd. and Merck KGaA, who kindly hosted two of these meetings.

Finally, we would like to courteously thank a member who thoughtfully gave his time and contributed his expertise in a key role but who unfortunately cannot be named.

At CIOMS, Hervé Le Louët and Lembit Rägo shared the chairing of the WG meetings, and the project was managed by Lembit Rägo, Sanna Hill, Kateriina Rannula, and Sue le Roux.

Geneva, Switzerland, 2025

Lembit Rägo, MD, PhD
Secretary-General, CIOMS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ADR	Adverse Drug Reaction
AE	Adverse Event
AERS	Adverse event reporting systems
AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
AIDA	Artificial Intelligence and Data Act (Canada)
AI-MD	Artificial Intelligence Medical Device
AKI	Acute Kidney Injury
ATC	Anatomical Therapeutic Chemical
BCR	Binding Corporate Rules
BLEU	Bilingual Evaluation Understudy
CDC	Centers for Disease Control and Prevention
CGRP	Calcitonin gene-related peptide
CIOMS	Council for International Organizations of Medical Sciences
CNN	Convolutional Neural Networks
COU	Context of use
CSV	Computerized System Validation
EDSTP	Emerging Drug Safety Technology Program
EEA	European Economic Area
EHR	Electronic Health Records
EMA	European Medicines Agency
ETHER	Event-based Text-mining of Health Electronic Records
EU	European Union
EU AI Act	European Artificial Intelligence Act
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
GVP	Good Pharmacovigilance Practices
GxP	Good [x] Practices

HCP	Healthcare Professional
HFE	Human Factors and Ergonomics
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
IT	Information Technology
KPI	Key Performance Indicator
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Models
LRS	Likelihood of Retinopathy Score
MAH	Marketing Authorisation Holder
MedDRA	Medical Dictionary for Regulatory Activities
mfERG	Multifocal electroretinography
MHRA	Medicines and Healthcare products Regulatory Agency (UK)
ML	Machine Learning
ML-DSF	Machine Learning-Enabled Device Software Functions
NLP	Natural Language Processing
NLQ	Natural Language Query
OCT	Optical Coherence Tomography
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
PoC	Proof-of-Concept

PPV	Positive Predictive Value
PSMF	Pharmacovigilance System Master File
PV	Pharmacovigilance
QA	Quality Assurance
QC	Quality Control
QMS	Quality Management System
RAG	Retrieval Augmented Generation
ROI	Return on Investment
RWD	Real-World Data
RWE	Real-World Evidence
SARAH	Smart Artificial Intelligence Resource Assistant for Health
SD-OCT	Spectral Domain – Optical Coherence Tomography
SHAP	Shapley Additive exPlanations
SME	Subject Matter Expert
SOP	Standard Operating Procedures
SQL	Structured Query Language
SVM	Support Vector Machines
US FDA	U.S. Food and Drug Administration
USPI	US Prescribing Information
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[i] 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 standardisation in drug safety monitoring.

The thalidomide tragedy of the late 1950s and early 1960s exposed severe deficiencies in global drug safety practices, highlighting the need for comprehensive data collection and international harmonisation. In response, the World Health Organization (WHO) established the Programme for International Drug Monitoring in 1968, initiating efforts to share individual case reports between countries and harmonise data practices. Building on these foundational efforts, the late 1980s and the 1990s saw key 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 standardised adverse event (AE) collection and reporting in addition to signal detection processes.

Advancements in technology are transforming PV, with examples ranging from smart-phone and web apps for direct AE reporting,[1] to the integration of data across large health networks to enable nearly real-time protocol-based assessment of AEs.[2] The application of artificial intelligence (AI) to PV offers the hope of enhancing both the efficiency and quality of PV, but it calls for explainable, reliable, and responsible AI use, recognising that its usefulness requires human decision making and its acceptability needs to conform to regulatory expectations. Equally important is the ethical and responsible use of data, which underpins the integrity of AI outcomes and fosters public confidence in these technologies.

As AI continues to evolve and impact biomedical research, its increased integration in and impact on PV practice is inevitable. Given AI's significant potential to enable transformative advancements, it is imperative that we engage in rigorous and forward-thinking discourse: how do we envision its development, validation, and deployment 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

[i] Medicine
In this report, we use medicines for products that are used to treat, prevent, or diagnose medical conditions as well as some that restore, correct or modify how the body works. In this report, these are products that fall within the scope of national and regional medicines regulatory authorities' activities. Vaccines and medicine-device combinations fall within our description of medicines too. Other terms used interchangeably with medicines include drugs, medications, and medicinal products.

Adopted from: CIOMS Working Group XI

applications, we recognise 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. Consequently, this report intends 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 or particularly applicable to PV rather than considerations for the more general use of AI.

This report aims to provide guidance to individuals and organisations interested in developing solutions for the use of AI in PV, including regulators, industry, academic researchers, clinicians, patients, ethicists, technology vendors, and global organisations.

Preface – References

1 Liyanage PH, Madhushika MT, Liyanage PLGC. Effectiveness of mobile applications in enhancing adverse drug reaction reporting: A systematic review. BMC Digit Health, 2025;3:15. https://doi.org/10.1186/s44247-025-00153-9 (Journal full text)

2 Desai RJ, Matheny ME, Johnson K, Marsolo K, et al. (2021). Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework. NPJ Digit Med. 2021;4:170. https://doi.org/10.1038/s41746-021-00542-0 (Journal full text)

EXECUTIVE SUMMARY

The Council for International Organizations of Medical Sciences (CIOMS) report on artificial intelligence (AI) in pharmacovigilance (PV) addresses a rapidly emerging cross-disciplinary field that is at the intersection of PV, computer science, mathematics, 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. The latter must be clearly defined and understood by many people from different backgrounds to propel research and practical implementation in an effective, safe and responsible manner. The diverse pool includes professionals, researchers, and decision makers working in PV in biopharmaceutical industry, regulatory authorities, and academia. It also includes software vendors that develop AI solutions PV, 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 one's organisation.

Perhaps more than other CIOMS report topics, the potential hazards 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 organisations, through which AI in PV must grow. This report provides a set of guiding principles and corresponding organisations 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, fairness and equity, and governance and accountability. Key points to consider for these guiding principles are elaborated throughout the report and summarised 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) recognised that the field of AI is progressing so rapidly that a prescriptive document would likely be quickly outdated. Instead, the WG decided to focus on 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 report 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 systems, and corresponding impacts on the safety and well-being of individuals and society. The level of risk, and corresponding intensity of oversight, depends on two considerations: 1. whether the decision is a high stakes decision, i.e. are the outputs used to make a critical decision(s) for which an error has substantial adverse consequences to humans; and 2. whether the AI solution is intended to be used in unchecked, stand-alone mode versus with a human-computer 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. For a given PV application, 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 and documented at regular intervals and adapted if needed.

Human oversight. Human oversight supports performance optimisation 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" where the decision is the end result 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 benchmarks, selecting appropriate data for model development and testing/validation in a realistic setting, an ongoing quality assessment process and retraining or dynamic/online learning 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 continually and 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, a cross-disciplinary exercise, and span a diverse range of relevant examples. Evaluations should use a sufficient representation of relevant data types (e.g. data sources such as spontaneous reports, clinical trials, and literature), reporter/patient characteristics, and a variety of medicines, vaccines, and AEs, to mitigate the chance of, and detect, biases, promote adequate and generalisable performance across the intended deployment domain, assess usability, and identify circumstances associated with underperformance. As many applications in PV focus on recognition of rare events or patterns (e.g. safety signals and duplicates), enrichment strategies to obtain representative test sets with high enough prevalence of the outcome may be required. Special care should be taken to attempt to ensure that performance evaluation results generalise 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, with a corresponding value proposition, should be communicated, including model development and architectures, expected inputs and outputs, and the nature of human-computer interaction. To fully characterise 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 return plausible hypotheses 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, regulatory compliance and scientific hypothesis generation.

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. However, explainable AI methods have limitations, and they only provide plausible hypotheses, 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 Large Language Models (LLMs) for patient re-identification, which may be addressed by pre-deployment data-protection-impact-assessments. 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 and other relevant stakeholders, such as software vendors, should recognise 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.

Fairness & Equity. Key regulatory and ethical imperatives for the fair and equitable use of AI in PV include: supporting fairness and equity, avoiding propagating or amplifying harmful explicit biases, discrimination and inaccurate results during model development and deployment, and underserving certain subpopulations, which may even permeate the initial decision of whether or not to implement an AI solution. 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. Training and performance evaluation of reference data sets should be scrutinised for adequate representation and performance evaluated 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 solutions are used safely, responsibly and ethically, and in compliance with 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 solutions.

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 and refinement of a proposed governance framework grid (see Chapter 9 on Governance & Accountability) of the aforementioned guiding principles for practical use.

Future considerations for development and deployment of artificial intelligence in pharmacovigilance. Increasing deployment of AI in PV is expected to prioritise 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 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 could 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. routinised 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 could eventually develop refined medical and scientific judgement.

It is critical that the guiding principles outlined in this report remain as core considerations and responsibly applied in specific context of use. They will need to evolve and adapt with advancements and application of AI in PV and medicine in general, which requires flexibility and full understanding of the process, data, and capabilities and limitations of AI. 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 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. The definition encompasses systems ranging from those explicitly programmed to perform tasks based on human expertise, to machine learning (ML) based methods, including more complex approaches such as deep neural networks. We acknowledge however that organisations that use an AI system may apply narrower definitions in their own processes.

In the context of PV, the use of AI systems and activities is aimed at enhancing drug safety monitoring, patient safety and regulatory compliance with the overall objective to inform decisions to optimise treatment benefit-risk. 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 should be designed to address specific objectives within PV. The overall AI solution could be developed with one or many AI systems. An AI system encompasses the model(s) and components necessary for operation, including user interfaces and data processing pipelines. At the core of these systems are AI models. These models utilise parameters to learn patterns or relationships within data, enabling the systems to adapt and improve model performance over time or support knowledge retrieval systems.

Simpler AI systems, such as statistical methods for signal detection, have been widely utilised 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 (NLP). These advancements have resulted in a significant increase in their use. In addition, 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, medical imaging and diagnostics, drug discovery and development, 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 plausible inferences by utilising accumulated data, rather than relying

solely on explicit rules or human intervention. This approach facilitates the development of AI systems that provide new, improved, or complementary solutions. A critical enabler for AI success within PV will be the ability to link and analyse 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. Linkage to external datasets may entail additional privacy implications and risks. To mitigate these risks, privacy-preserving record linkage approaches can be employed, enabling secure and ethical data integration while maintaining patient confidentiality.[9,10] 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 utilising AI including ML.

The start of systematic safety monitoring 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 summarising data.

Individual case safety reports (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, with further dissemination to other safety databases (e.g. regulatory authorities). As the number of product approvals and the patient exposure grow, so does the number of reported AEs. The increased volume of ICSRs, coupled with stringent regulations impacting PV, creates significant challenges in ICSR processing and compliance.

Once a signal is detected as a result of individual or aggregate analysis of AE 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.[11] 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 AE reports include:[12,11]

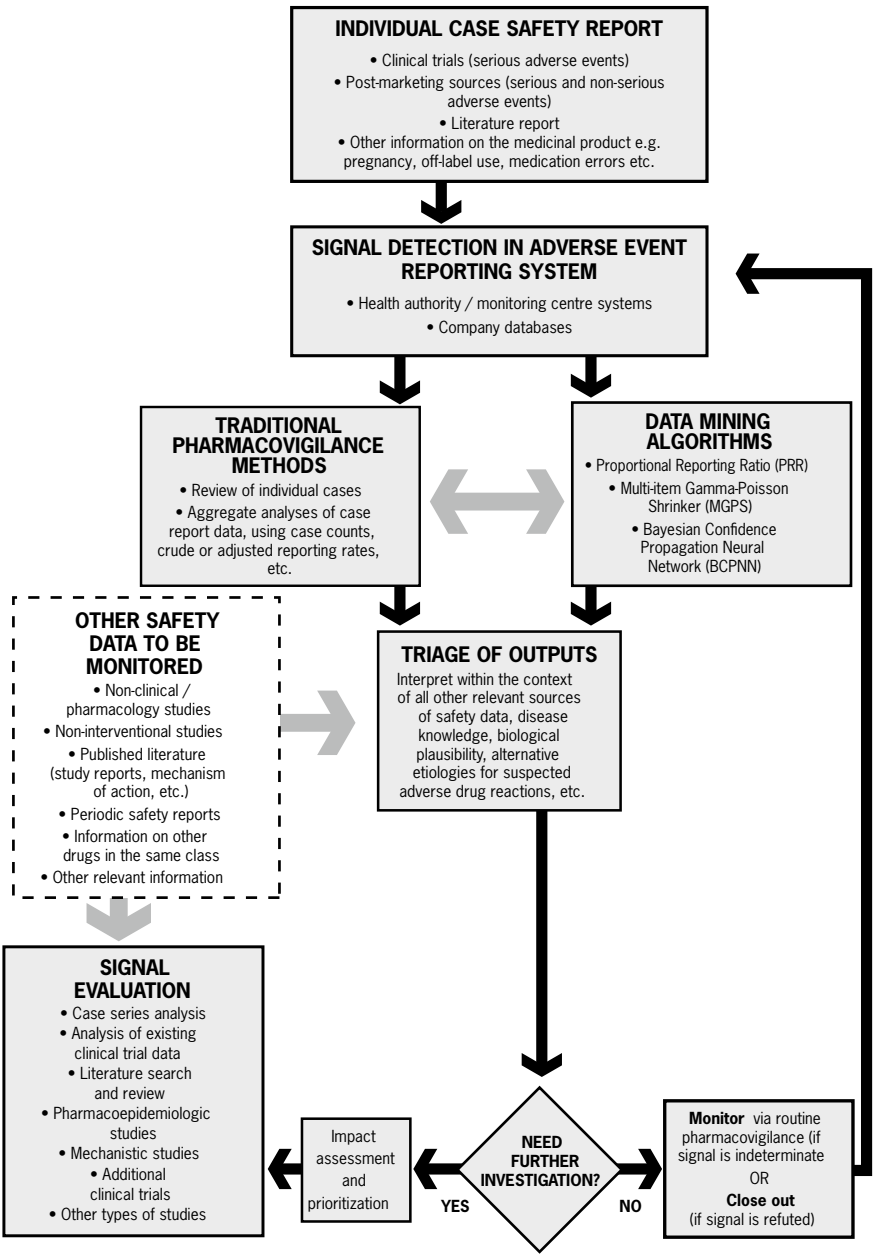
- Review of ICSRs or case series in a PV database or in published medical or scientific literature; and
- Aggregate analyses of case reports using absolute case counts, simple reporting rates, proportions or estimated exposure-adjusted reporting rates.

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, real-world data (RWD), literature, and information from clinical trials etc.

Once safety concerns (including important identified risks or important potential risks) and missing information are identified, risk management activities are put in place 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 Direct Healthcare Professional communications (DHPCs).

Figure 1. Representative signal management process

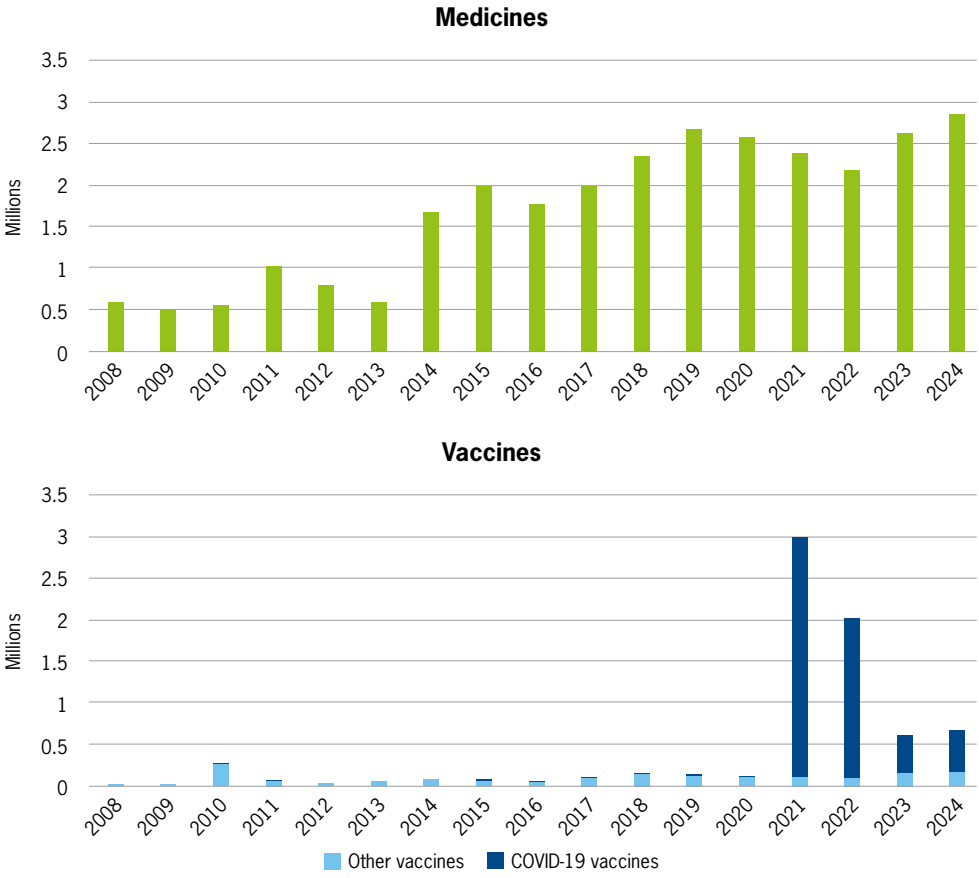
Source: Modified from CIOMS Working Group report VIII[11]



The COVID-19 pandemic has further emphasised the need for advanced methods in PV, as it has led to a significant rise in safety reports (see Figure 2 and Figure 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 medicines.

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

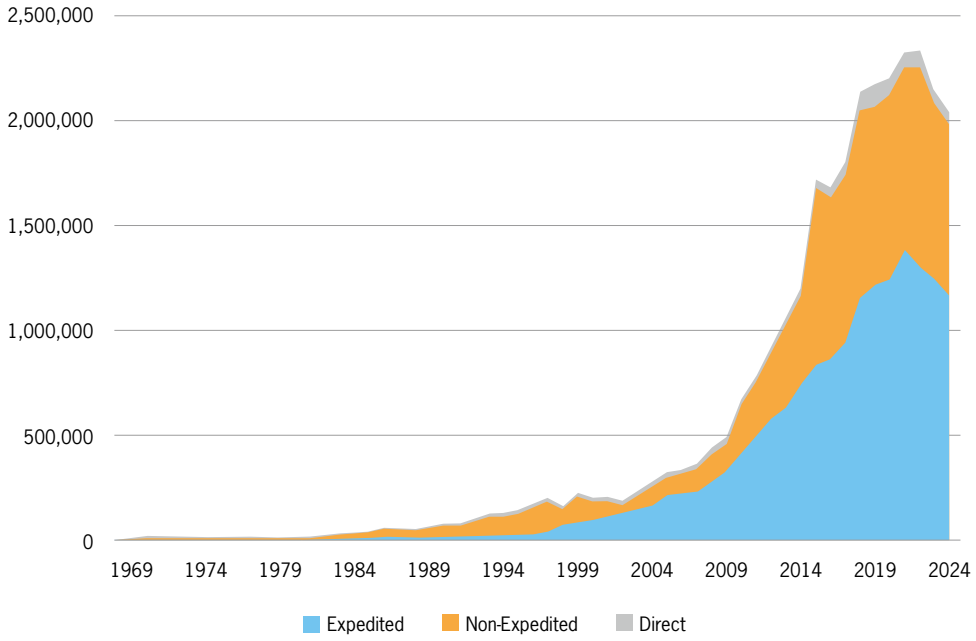
Source: VigiBase accessed April 2025. Figure reproduced with permission.



Number of reports received annually in VigiBase (accessed April 2025) stratified by medicines and vaccines.

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

Source: Constructed using FDA FAERS database.[15]



Number of reports received annually in the FDA FAERS database (accessed October 2025) stratified expedited, non-expedited and reports submitted directly to FDA. This may include multiple versions of the same cases e.g. in the form of follow-up reports.

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 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 need.

Rapid evolution of artificial intelligence

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

[i] *BERT: Bidirectional Encoder Representations from Transformers*

However, the landscape has been evolving beyond this framework thanks to emerging technologies like GenAI, knowledge graphs and ontologies. 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 ML learning techniques.

1.1. Scope

This document aims to guide those working in PV in addition to organisations developing AI solutions for the PV domain, such as regulators, medicinal products industry professionals, software vendors, international and national PV organisations, 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. Recognising the rapid evolution and application of AI technology, the CIOMS Working Group XIV developed this document to guide the development and integration of AI systems into PV activities.

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

The scope deliberately excludes broader healthcare data applications outside the direct purview of PV, 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.

Chapter 1 – References

1 Organisation for Economic Co-operation and Development (OECD). *Explanatory memorandum on the updated OECD definition of an AI system*. OECD Artificial Intelligence Papers, No. 8. Paris: OECD Publishing; 2024. 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. https://doi.org/10.1002/pds.5501 (Journal full text)

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 Saf. 2023;Apr46(4):433. https://doi.org/10.1007/s40264-023-01273-9 (Journal full text)

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

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

6 Wright T, Yan P, Easterbrook M. Machine learning to identify multifocal ERG deficits in patients taking hydroxychloroquine. Invest Ophthalmol Vis Sci. 2019;Jul22;60(9):5959. (Journal abstract accessed 15 October 2025)

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). https://doi.org/10.3390/ijerph18073839 (Journal full text)

8 Li MH, Mestre TA, Fox SH, Taati B. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *J Neuroeng Rehabil.* 2018;Dec15;15(1):3. https://doi.org/10.1186/s12984-018-0446-z (Journal abstract)

9 Pathak A, Serrer L, Zapata D, King R, Mirel LB, Sukalac T, et al. Privacy preserving record linkage for public health action: opportunities and challenges. *J Am Med Inform Assoc.* 2024;Nov1;31(11):2605-2612. https://doi.org/10.1093/jamia/ocae196 (Journal abstract)

10 Eisinger-Mathason TSK, Leshin J, Lahoti V, Benner B, Snyder M, Kohane I, et al. Data linkage multiplies research insights across diverse healthcare sectors. *Commun Med.* 2025;5:58. https://doi.org/10.1038/s43856-025-00769-y (Journal full text)

11 Council for International Organizations of Medical Sciences (CIOMS). *Practical aspects of signal detection in pharmacovigilance: CIOMS Working Group VIII report.* Geneva: Council for International Organizations of Medical Sciences; 2010. (Full text accessed 15 October 2025)

12 U.S. Food and Drug Administration (FDA). *Good pharmacovigilance practices and pharmacoepidemiologic assessment: guidance for industry.* Silver Spring (MD): U.S. Food and Drug Administration; 2005;Mar. (Full text accessed 21 March 2025).

13 Alvager T, Smith TJ, Vijai F. Neural-network applications for analysis of adverse drug reactions. *Biomed Instrum Technol.* 1993;Sep-Oct;27(5):408-411. (Journal abstract accessed 15 October 2025)

14 Këpuska V, Bohouta G. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC).* Las Vegas (NV): IEEE; 2018. p. 99-103. https://doi: 10.1109/CCWC.2018.8301638. (Journal abstract)

15 U.S. Food and Drug Administration (FDA). *FDA Adverse Event Reporting System (FAERS) Public Dashboard.* [Internet]. Silver Spring (MD): U.S. Food and Drug Administration; 2025. (Webpage accessed 29 October 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. https://doi.org/10.1093/jamiaopen/ooaf003 (Journal full text)

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: *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI).* Las Vegas (NV): IEEE; 2023. p. 1005-1010. 2023;Dec13:49-56. https://doi:10.1109/CSCI62032.2023.00015 (Journal abstract)

CHAPTER 2.
LANDSCAPE ANALYSIS

2.1. Use of artificial intelligence in pharmacovigilance to date

AI may directly or indirectly impact all aspects of PV (see Figure 1: Representative signal management). In this chapter, we discuss systems 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 attributes or features especially prominent in PV applications and processes. For example, AI systems for general translation tasks are out of scope, but PV specific translations, e.g. of AE reports or reporter source documents (e.g. original case reports, transcriptions of a call), are in scope. Additionally, 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.[1] The reader is also referred to the many perspectives and commentaries that discuss the use of AI in PV[2,3,4,5,] and the cautionary notes that have been provided.[6]

2.1.1. Adverse event capture

AI systems have been proposed and evaluated for a variety of tasks related to NLP of social media content to identify references to (personal experiences of) medicine use and AEs that may provide the basis for AE reports. These tasks include identifying relevant posts,[7,8] identifying relevant parts of such posts,[9] normalising descriptions of AEs or medicinal products within such posts to standardised terminologies like the Medical Dictionary for Regulatory Activities (MedDRA) or the Anatomical Therapeutic Chemical (ATC) classification system code,[10] and classifying the relationship between AEs and drugs mentioned in the same posts.[11]

Similarly, AI methods have been developed to support screening the scientific literature for AEs that may be captured on AE reports.[12,13]

2.1.2. Individual Case Safety Report Processing

An area of ICSR processing where AI systems have been in routine use by some organisations since at least the 2010's is duplicate detection, which relates to the identification of multiple unlinked records describing the same AE in a particular patient.[14] Duplicate detection methods based on ML and probabilistic record linkage have been implemented for VigiBase,[15] US Food and Drug Administration Adverse Event Reporting System (FAERS),[16] and EudraVigilance.[17] The use of NLP to improve duplicate detection by extracting and incorporating information from free text has also been explored.[16] Rule-based methods are more widely used and easier to implement but do not perform as well.[14,18]

Another area where AI has been used to support ICSR processing is in the encoding of information on AEs[19,20] or medicinal products[21] on AE reports in standard terminologies based on verbatim fields and/or free-flowing case narratives. NLP has also been applied to extract relevant information from case narratives and map it to structured fields[22,23,24,25,26,27] and for ICSR translation.[28]

Several organisations who process large numbers of case reports have also automated repetitive, labour-intensive tasks using rule-based so-called Robotic Process Automation (RPA) technologies.[29] These operate on the user interface of other computer systems mimicking actions that humans otherwise would take.[30] They may for example automate duplicate checking and importing of cases, as described by TransCelerate.[31]

Other applications of AI systems during ICSR processing include methods that have been developed to create narratives from structured data,[29] help support triage of incoming reports for human review,[32,33] help support individual case causality assessment,[34] and automate redaction of person names in case narratives.[35]

2.1.3. Signal detection and analysis

The earliest examples of real-world use of (simple, narrow, and rule-based) AI in PV are from the late 1990s. At this point, disproportionality analysis, first conceptualised in the 1970s,[36] began to be implemented *as part of rule-based triage algorithms* to help direct the attention of PV specialists in their analysis of large national and international collections of individual case reports.[37,38,39,40] Since then, various incremental improvements have been introduced and evaluated including automated adjustment for confounding through e.g. regression,[41,42] or propensity scores,[43] extensions to drug-drug interactions,[44,45,46,47,48,49] and other possible risk factors for adverse reactions.[50] Methods to detect AEs associated with the production process or with substandard or counterfeit medicines have also been explored.[51,52,53] 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,[54] the time-to-onset of the reported reactions,[55] or a combination of e.g. Naranjo scores and the proportions of reports on a drug-AE combination coming from healthcare professionals (HCPs) and marketing authorisation holders (MAHs), respectively.[56]

NLP has been applied to mine regulatory information,[57] scientific literature, and clinical notes[58,59,60,61] for information on already known/unknown and potentially serious adverse effects. This may support and streamline decision making, especially during early signal assessment and prioritisation.

Some published AI-based signal detection exercises provide tantalising glimpses of how elegant AI solutions may uncover truly novel AEs.[62] At the same time, 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.[63]

Several organisations have developed predictive models for ICSR prioritisation to assess causal associations between drugs and AEs and/or inform a regulatory action.[64,65,33,66,67] These can be used to prioritise 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[68,69] and there have been efforts to provide ML-based decision support for signal validation[70] and to automatically visualise relevant information on case

reports to facilitate human review during signal assessment.[71] ML 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.[72]

Applications of unsupervised learning have been developed to support signal detection and analysis, especially seeking to bring together reports describing similar or related AEs. These include network analyses of AEs (and to a lesser extent drugs),[73,74,75,76,77] cluster analysis of AE reports,[78,79]and data-driven derivations of semantic representations of AEs and drugs.[80,81]

Datasets with information about drug side effects and indications such as DrugBank[82] and SIDER,[83] as well as those with information on pharmacology and chemical structures such as Bio2RDF,[84] have been leveraged to enhance PV signal detection and analysis,[85,86,87] or derive knowledge graphs that can serve as downstream inputs for AI-based predictive signal detection.[88,90] There have also been AI applications that help retrieve scientific papers relevant to the analysis of possible adverse effects.

2.1.4. Early applications of generative AI in pharmacovigilance

Early applications of generative Large Language Models (LLMs) in PV have started to be explored. They include also applications where generative LLMs are prompted or post-processed to more restricted outputs, as a basis for e.g. classification or named entity recognition. Examples to date include use of generative LLMs to simplify the patient communication from a regulatory authority,[89] summarisation for drug labelling documents[90] and of case narratives,[91] screening scientific literature and social media,[13,92] search of drug safety documentation,[5] Q&A for drug labelling,[93] PV context-aware generation of Structured Query Language (SQL) code,[94] and drafting follow-up letters to reporters.[95]

2.1.5. 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, Table 1 presents examples of AI solutions that have been adopted for routine use in PV by various PV organisations and are described in the public domain. The deployment of AI solutions by pharmaceutical companies may be largely based on software vendor implementations, which are not described in the public domain.[96] Similarly, several AI solutions deployed by the European Medicines Agency (EMA) are described in public domain,[97] but not yet in separate scientific publications. See also the use cases presented in Appendix 3.

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[21]
Duplicate detection	FAERS,[98] VigiBase[15]
Automated triages of individual case reports	Swedish Medical Products Agency[34], pharmaceutical companies[99]
Automated triages for quantitative signal detection	Databases of various regulatory authorities, international organisations, and pharmaceutical companies
Predictive models for quantitative signal detection	VigiBase,[56,100] Netherlands pharmacovigilance centre Lareb[101]
Adverse event cluster analysis for signal detection and assessment	VigiBase[80,102]
Literature surveillance for safety data	EudraVigilance[103] Netherlands pharmacovigilance centre Lareb[104]

2.2. Regulatory considerations

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.[105] Countries have developed different regulatory frameworks and guiding principles to ensure the ethical use and trustworthiness of AI systems, and legislations of AI are being implemented (i.e. European Artificial Intelligence Act [EU AI Act][106], Artificial Intelligence and Data Act [AIDA]).[107,111] In addition, there have been published reflection and discussion papers on the use of AI in medicinal products by the EMA and the United States Food and Drug Administration (FDA), as well as a draft guidance on AI use to support regulatory decision making for drug products by the US FDA.

2.2.1. 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 The Organisation for Economic Co-operation and Development (OECD). Select publications defining guiding principles and recommended best practices for safe and responsible AI use in regulated fields were reviewed by the CIOMS Working Group XIV. Although these publications were not developed specifically for PV, the described guiding principles for AI were determined to be applicable to the field of PV. It should be acknowledged that some discretion was used by the CIOMS Working Group XIV to establish the guiding principles for PV from these various publications, as some of the principles were described in conjunction with other principles. Of note, the US reference used by the CIOMS Working Group XIV has since been archived and no longer represent the current US policy; however, this reference has been retained as

it was used to inform the development of guiding principles for this report. For more current AI policies from other countries, please refer to the Artificial Intelligence Policy Tracker.[108] Table 2 provides an overall comparison of the guiding AI principles from select governmental institutions and international organisations. A non-exhaustive description of the AI 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 organisations

Source: CIOMS Working Group XIV

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	EU[109,110]	Australia[111]	Canada[113]	Singapore[112]	UK[113]	US[114]	PAHO[115]	WHO[116]	OECD[117]
Human Oversight	✓	✓	✓		✓	✓	✓	✓	✓
Validity & Robustness	✓	✓	✓		✓		✓	✓	✓
Data Privacy	✓	✓			✓	✓	✓		
Transparency	✓	✓	✓	✓	✓		✓	✓	✓
Accountability	✓	✓	✓	✓	✓	✓	✓	✓	✓
Societal well-being	✓	✓		✓			✓	✓	✓
Environmental Well-being	✓	✓						✓	✓
Fairness & Equity	✓	✓	✓	✓	✓	✓	✓	✓	✓
Explainability	✓	✓		✓	✓	✓		✓	✓
Safety	✓	✓	✓		✓	✓		✓	✓
Governance	✓				✓			✓	

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

On September 9, 2024, the EMA finalised its Reflection paper on the use of AI in the medicinal product lifecycle.[118] The reflection paper addresses the use of AI/ML in the safe and effective development, manufacturing and use of medicines.

The 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/ 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 notes that AI/ML tools can effectively support activities such as AE report management and signal detection, in line with applicable Good Pharmacovigilance Practices (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 AE 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 Good [x] Practices (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.[118]

The EMA is planning to develop further guidance on the use of AI in the medicines lifecycle, including in PV.[118]

2.2.3. US FDA Discussion Paper on Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products

In May 2023, the U.S. Food and Drug Administration (FDA) published a discussion paper on "Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products".[119] The US FDA acknowledges the increased use of AI/ML in the lifecycle of drug development with novel approaches in data mining, analysing large multi-omics, PK/PD modelling, 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 post-marketing safety surveillance, the US FDA sees the potential to: i) automate the processing and prioritisation of ICSR using AI/ML, due to the increasing volume of reports and complexity of data sources; ii) classify ICSRs on the likelihood of causal relationship between the drug and AE; iii) determine the seriousness of the outcome of ICSRs; and iv) automate aggregate reports for multiple AEs for a particular product.

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

The US 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[120]. 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. It considers AI broadly, i.e. not limited to specific subsets of AI such as ML. 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)" 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 the 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.[121]

2.2.5. US FDA Emerging Drug Safety Technology Program

The US FDA 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 US 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.[122]

2.2.6. Guidance on use of Large Language Models

Since the release of ChatGPT (Generative Pre-trained Transformer) on November 30, 2022, there has been significant work in exploring how GenAI 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 organisations 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 (LLMs) in regulatory science and for medicines regulatory activities (EMA[123]);
- Guide on the use of generative AI (Canada[124]);
- Initial policy considerations for generative artificial intelligence (OECD[125]);
- WHO Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. (WHO[126]).

2.2.7. Guidelines for safe Artificial Intelligence

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

- Regulatory considerations on artificial intelligence for health, WHO 2023;[127]
- Ethics Guidelines for Trustworthy AI, European Commission 2019;[128]
- Recommendation of the Council on Artificial Intelligence, OECD 2019, amended 2023;[129]
- Good Machine Learning Practice for Medical Device Development: Guiding Principles, US FDA, Health Canada, MHRA 2021;[130]
- Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles;[131]
- ISO/IEC 23894:2023 Information technology - Artificial Intelligence – Guidance on risk management.[132]

Chapter 2 – References

1 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.* 2022;Oct;36(5):295-306. https://doi.org/10.1007/s40290-022-00441-z (Journal abstract)

2 Basile AO, Yahi A, Tatonetti NP. Artificial intelligence for drug toxicity and safety. Trends Pharmacol Sci. 2019;Sep1;40(9):624-635. https://DOI:10.1016/j.tips.2019.07.005 (Journal full text)

3 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 Saf. 2022;May;45(5):439-448. https://doi:10.1007/s40264-022-01164-5 (Journal full text)

4 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.) https://doi.org/10.1016/j.hlpt.2023.100743 (Journal full text)

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

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

7 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 Saf. 2014;May;37:343-350. https://doi.org/10.1007/s40264-014-0155-x (Journal full text)

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

9 Nikfarjam A, Sarker A, O'connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc. 2015;May1;22(3):671-681 https://doi.org/10.1093/jamia/ocu041 (Journal full text)

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

11 White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the internet. Clin Pharmacol Ther. 2014;Aug;96(2):239-246. https://doi:10.1038/clpt.2014.77 (Journal full text)

12 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 Saf. 2024;Jan;47(1):71-80. https://doi.org/10.1007/s40264-023-01367-4 (Journal abstract)

13 European Medicines Agency (EMA). *AI Observatory 2024*. Amsterdam: European Medicines Agency; 2025;May 6. EMA/154528/2025. (Webpage accessed 23 September 2025)

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

15 Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;Jun;14:305-328. https://doi.org/10.1007/s10618-006-0052-8 (Journal abstract)

16 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 Saf. 2017;Jul;40:571-582. DOI:10.1007/s40264-017-0523-4 (Journal full text)

17 Personal communication from Tom Paternoster-Howe (March 2025).

18 Kiguba R, Isabirye G, Mayengo J, et al. Navigating duplication in pharmacovigilance databases: a scoping review. BMJ Open 2024;14:e081990. https://doi:10.1136/bmjopen-2023-081990 (Journal full text)

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

20 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. https://doi.org/10.1186/2041-1480-3-15 (Journal full text)

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

22 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. Pharm Med. 2018;32 (6):391-401. https://doi.org/10.1007/s40290-018-0251-9 (Journal full text)

23 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. J Biomed Semantics. 2016;Aug1;62:78-89. https://doi.org/10.1016/j.jbi.2016.06.006 (Journal full text)

24 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. J of Biomed Informatics. 2017;Sep1;73:14-29. https://doi.org/10.1016/j.jbi.2017.07.012 (Journal full text)

25 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. Pharm Med. 2021;Sep;35(5):307-316. https://doi.org/10.1007/s40290-021-00398-5 (Journal abstract)

26 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. Front Drug Saf Regul. 2022;Nov14;2:1020943. https://doi.org/10.3389/fdsfr.2022.1020943 (Journal full text)

27 Abatemarco D, Perera S, Bao SH, Vermeer P, Botsis T. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. *Pharm Med.* 2018;32(6):391-401. https://doi.org/10.1007/s40290-018-0251-9 (Journal full text)

28 Römming H-J, Pushparajan R. *AI translation assistant for pharmacovigilance*. Poster presented at: DIA Europe 2021; 2021; Virtual conference. (Webpage accessed 21 March 2025)

29 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 (Journal full text)

30 van der Aalst WMP, Bichler M, Heinzl A. Robotic process automation. *Bus Inf Syst Eng*. 2018;60(4):269-272. https://doi.org/10.1007/s12599-018-0542-4 (Journal full text)

31 TransCelerate BioPharma Inc. *Successful implementation of robotic process automation (RPA) in the individual case safety report (ICSR) management process leads to enhanced ability to protect patients: featured solution – interactive (ICSR) & automation technologies tool (IATT)*. [Internet]. Philadelphia (PA): TransCelerate BioPharma Inc.; 2023. (Webpage accessed 23 September 2025)

32 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. Pharmacoepidemiol Drug Saf. 2022;Dec;31(12):1300-1307. https://doi.org/10.1002/pds.5553 (Journal full text)

33 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. https://doi.org/10.1371/journal.pdig.0000409 (Journal full text)

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

35 Meldau EL, Bista S, Melgarejo-González C, Norén GN. Automated redaction of names in adverse event reports using transformer-based neural networks. BMC Med Inform Decis Mak. 2024;Dec23;24(1):401. https://doi.org/10.1186/s12911-024-02785-9 (Journal full text)

36 Finney DJ. Systematic Signalling of Adverse Reactions to Drugs. Methods Inf Med. 1974;13(01):1-10. https://doi.org/10.1055/s-0038-1636131 (Journal full text)

37 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. Eur J Clin Pharmacol. 1998;Jul;54:315-321. https://doi.org/10.1007/s002280050466 (Journal abstract)

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

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

40 Council for International Organizations of Medical Sciences (CIOMS). *Practical aspects of signal detection in pharmacovigilance: CIOMS Working Group VIII report*. Geneva: Council for International Organizations of Medical Sciences; 2010. (Full text accessed 15 October 2025)

41 Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. Stat Anal Data Min ASA Data Sci J. 2010;Aug;3(4):197-208. https://doi.org/10.1002/sam.10078 (Journal abstract)

42 Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C. Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clin Pharmacol Ther. 2011;Feb;89(2):243-250. https://doi.org/10.1038/clpt.2010.285 (Journal full text)

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

44 van Puijenbroek E, Egberts A, Heerdink E, Leufkens H. 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*. 2000;56(9-10):733-738. https://doi.org/10.1007/s002280000215 (Journal full text)

45 Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug–drug interactions in a spontaneous reports database. Br J Clin Pharmacol. 2007;Oct;64(4):489-495. https://doi.org/10.1111/j.1365-2125.2007.02900.x (Journal full text)

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

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

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

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

50 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 Saf. 2020;Oct;43(10):999-1009. https://doi.org/10.1007/s40264-020-00957-w (Journal full text)

51 Juhlin K, Karimi G, Andér M, Lucas C, Star K, Norén GN. Using VigiBase to identify substandard medicines: detection capacity and key prerequisites. *Drug Saf*. 2015;38(4):373-382. https://doi.org/10.1007/s40264-015-0271-2 (Journal full text)

52 Trippe ZA, Brendani B, Meier C, de Jong H, Norén GN, Juhlin K. Identification of substandard medicines via disproportionality analysis of individual case safety reports. *Drug Saf*. 2017;40(4):293-303.https://doi.org/10.1007/s40264-016-0499-5 (Journal abstract)

53 Mahaux O, Bauchau V, Zeinoun Z, Van Holle L. Tree-based scan statistic – application in manufacturing-related safety signal detection. *Vaccine*. 2019;37(1):49-55. https://doi.org/10.1016/j.vaccine.2018.11.044. (Journal full text)

54 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 Saf. 2014;Aug37:617-628. https://doi.org/10.1007/s40264-014-0204-5 (Journal full text)

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

56 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. Pharmacoepidemiol Drug Saf. 2018;Feb;27(2):199-205. https://doi.org/10.1002/pds.4364 (Journal full text)

57 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. J Biomed Inform. 2018;Jul1;83:73-86. https://doi.org/10.1016/j.jbi.2018.05.019 (Journal full text)

58 Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc. 2011;Sep1;18(5):668-674. https://doi.org/10.1136/amiajnl-2011-000096 (Journal abstract)

59 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. J Am Med Inform Assoc. 2013;May1;20(3):446-452. https://doi.org/10.1136/amiajnl-2012-001083 (Journal abstract)

60 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. Clin Pharmacol Ther. 2024;Jun;115(6):1391-1399. https://doi.org/10.1002/cpt.3226 (Journal full text)

61 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. https://doi.org/10.1101/2024.07.09.24310100 (Journal full text)

62 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). https://doi.org/10.1038/clpt.2011.83 (Journal full text)

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

64 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 Saf. 2020;Apr;43:329-338. https://doi.org/10.1007/s40264-019-00897-0 (Journal abstract)

65 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. Comput Biol Med. 2021;Aug1;135:104517. https://doi.org/10.1016/j.compbiomed.2021.104517 (Journal abstract)

66 Lieber T, Gosselt HR, Kools PC, Kruijssen OC, Van Lierop SNC, Härmark L, van Hunsel F. Natural language processing for automated triage and prioritization of individual case safety reports for case-by-case assessment. Front Drug Saf Regul. 2023;3: 1120135. https://doi.org/10.3389/fdsfr.2023.1120135 (Journal full text)

67 European Medicines Agency. *2024 AI Observatory*. 6 May 2025. EMA/154528/2025. (Webpage accessed 23 September 2025)

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

69 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 Saf. 2025;Mar15:1-3. https://doi.org/10.1007/s40264-025-01529-6 (Journal full text)

70 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 Saf. 2022;May;45(5):583-596. https://doi.org/10.1007/s40264-022-01159-2 (Journal full text)

71 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 Saf. 2020;Sep;43:905-915. https://doi.org/10.1007/s40264-020-00945-0 (Journal abstract)

72 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. https://doi.org/10.1002/cpt.1789 (Journal full text)

73 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. Int J Neural Syst. 2005;Jun;15(03):207-222. https://doi.org/10.1142/S0129065705000219 (Journal abstract)

74 Ball R, Botsis T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS?. Clin Pharmacol Ther. 2011;Aug;90(2):271-278. https://doi.org/10.1038/clpt.2011.119 (Journal abstract)

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

76 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

77 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 Saf. 2024;Aug15:1-8. https://doi.org/10.1007/s40264-024-01471-z (Journal full text)

78 Norén GN, Meldau EL, Chandler RE. Consensus clustering for case series identification and adverse event profiles in pharmacovigilance. Artif Intell Med. 2021;Dec1;122:102199.

79 Okada T, Hashiguchi M, Hori S. Classification of patient characteristics associated with reported adverse drug events to neuraminidase inhibitors: an applicability study of latent class analysis in pharmacovigilance. Int J Clin Pharm 2022;44; 1332-1341. https://doi.org/10.1007/s11096-022-01477-6 (Journal full text)

80 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. Sci Rep. 2017;Nov27;7(1):16416. https://doi.org/10.1038/s41598-017-16674-x (Journal full text)

81 Simms AM, Kanakia A, Sipra M, Dutta B, Southall N. A patient safety knowledge graph supporting vaccine product development. BMC Med Inform Decis Mak. 2024;Jan4;24(1):10. https://doi.org/10.1186/s12911-023-02409-8 (Journal full text)

82 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 Res. 2024;Jan5;52(D1):D1265-1275. https://doi.org/10.1093/nar/gkad976 (Journal full text)

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

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

85 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. https://doi.org/10.1186/s12859-015-0774-y (Journal full text)

86 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. J Am Med Inform Assoc. 2016;Sep1;23(5):968-78. https://doi.org/10.1093/jamia/ocv127 (Journal full text)

87 Bai C, Wu L, Li R, Cao Y, He S, Bo X. Machine learning-enabled drug-induced toxicity prediction. *Adv Sci (Weinh).* 2025;12(12):2413405. https://doi.org/10.1002/advs.202413405 (Journal full text)

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

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

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

91 Dowdy K, Hoffman A, Giles T, Kugele D, et al. Summarizing FAERS narratives with generative AI: methods, resource requirements, and quality assessment. In: *FDA Symposium on Scientific Computing and Digital Transformation*; 2024. Silver Spring (MD): Center for Drug Evaluation and Research, U.S. Food and Drug Administration; Booz Allen Hamilton. (Full text accessed 23 September 2025)

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

93 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. Regul Toxicol Pharmacol. 2024;May1;149:105613 https://doi.org/10.1016/j.yrtph.2024.105613 (Journal full text)

94 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. https://doi.org/10.1093/jamiaopen/ooaf003 (Journal full text)

95 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. https://doi:10.2196/65651 (Journal full text)

96 Barbieri MA, Battini V, Carnovale C, Cocco M, Papoutsi DG, Heckmann NS, Sessa M. Artificial intelligence in pharmacovigilance signal management: a review of tools, implementations, research, and regulatory landscape. Expert Opin Drug Saf. 2025;1-16. https://doi.org/10.1080/14740338.2025.2545926 (Journal full text)

97 European Medicines Agency (EMA). 2024 AI Observatory. 2025 May 6. EMA/154528/2025. (Webpage accessed 23 September 2025)

98 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. Front Drug Saf Regul. 2022;Jun15;2:918897. https://doi.org/10.3389/fdsfr.2022.918897 (Journal full text)

99 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 Health. 2023;26(6):S281. https://doi.org/10.1016/j.jval.2023.03.1556 (Journal full text)

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

101 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. Pharmacoepidemiol Drug Saf. 2018;Feb;27(2):199-205. https://doi.org/10.1002/pds.4364 (Journal full text)

102 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. Ther Adv Drug Saf. 2022;Aug;13:20420986221118972. https://doi.org/10.1177/20420986221118972 (Journal full text)

103 EudraVigilance. European Medicines Agency (EMA). (Webpage accessed 15 October 2025)

104 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. Reprod Toxicol. 2022;Oct1;113:150-154. https://doi.org/10.1016/j.reprotox.2022.09.001 (Journal full text)

105 Organisation for Economic Co-operation and Development (OECD). The state of implementation of the OECD AI principles four years on. OECD Artif Intell Pap. 2023;(3). Paris: OECD Publishing. https://doi.org/10.1787/835641c9-en. (Full text accessed 15 October 2025)

106 *ArtificialIntelligenceAct.eu.* [Internet]. Brussels: European Union; 2024. (Webpage accessed 15 October 2025)

107 Government of Canada. *Artificial Intelligence and Data Act (AIDA).* [Internet]. Ottawa: Government of Canada; 2023. (Webpage accessed 15 October 2025)

108 *AI Policy Tracker.* [Internet]. (Webpage accessed 25 Aug 2025)

109 European Medicines Agency (EMA). *Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle.* Amsterdam: European Medicines Agency; 2024;Sep 9. (Full text accessed 21 March 2025).

110 European Commission, Directorate-General for Communications Networks, Content and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.* Luxembourg: Publications Office of the European Union; 2020. https://data.europa.eu/doi/10.2759/002360 (Full text)

111 Department of Industry, Science, Energy and Resources (Australia). *Australia's artificial intelligence ethics principles.* Canberra: Australian Government; 2019. (Webpage accessed 21 March 2025).

112 Ministry of Health Singapore. *Artificial intelligence in healthcare guidelines (AIHGLE) 1.0*. Singapore: Ministry of Health; 2021. (Full text accessed 21 March 2025).

113 Government of the United Kingdom, Department for Science, Innovation & Technology. *Implementing the UK AI regulatory principles: guidance for regulators*. London: Department for Science, Innovation & Technology; 2024. (Full text accessed 21 March 2025).

114 United States. Office of Science and Technology Policy. *The Blueprint for an AI Bill of Rights: making automated systems work for the American people*. [Internet]. Washington (DC): The White House; 2022. (Webpage accessed 15 October 2025)

115 Pan American Health Organization (PAHO). *Artificial intelligence in public health: digital transformation toolkit, knowledge tools*. Washington (DC): Pan American Health Organization; 2021. PAHO/EIH/IS/21-011. (Full text accessed 21 March 2025).

116 World Health Organization (WHO). *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO. (Full text accessed 21 March 2025).

117 Organisation for Economic Co-operation and Development (OECD). *OECD AI principles overview*. Paris: Organisation for Economic Co-operation and Development; 2019, updated 2024. (Full text accessed 21 March 2025).

118 Heads of Medicines Agencies (HMA), European Medicines Agency (EMA) Joint Big Data Steering Group. *Multi-annual artificial intelligence workplan 2023-2028*. Amsterdam: European Medicines Agency; 2023. (Full text accessed 21 March 2025).

119 U.S. Food and Drug Administration (FDA). *Using artificial intelligence and machine learning in the development of drug and biological products*. Silver Spring (MD): U.S. Food and Drug Administration; 2023, revised 2025. (Full text accessed 21 March 2025)

120 U.S. Food and Drug Administration (FDA). *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*. Silver Spring (MD): U.S. Food and Drug Administration; 2025;Jan. (Full text accessed 15 October 2025)

121 U.S. Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER). *Artificial intelligence in drug manufacturing*. Silver Spring (MD): U.S. Food and Drug Administration; 2023. (Full text accessed 21 March 2025)

122 U.S. Food and Drug Administration (FDA). *CDER Emerging Drug Safety Technology Program (EDSTP)*. Silver Spring (MD): U.S. Food and Drug Administration; 2024. (Full text accessed 21 March 2025)

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

124 Government of Canada. *Guide to the use of generative AI*. Ottawa: Treasury Board of Canada Secretariat; 2023. (Full text accessed 21 March 2025)

125 Organisation for Economic Co-operation and Development (OECD). *Initial policy considerations for generative artificial intelligence*. Paris: Organisation for Economic Co-operation and Development; 2023. (Full text accessed 21 March 2025)

126 World Health Organization (WHO). *Ethics and governance of artificial intelligence for health: guidance on large multi-modal models*. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO. (Full text accessed 21 March 2025)

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

128 European Commission. *Ethics guidelines for trustworthy AI*. [Internet]. Brussels: European Commission; 2019 (Full text accessed 21 March 2025).

129 Organisation for Economic Co-operation and Development (OECD). *Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (2019, amended 2024)*. Paris: Organisation for Economic Co-operation and Development; 2024. (Full text accessed 21 March 2025)

130 U.S. Food and Drug Administration (FDA), Health Canada, Medicines and Healthcare products Regulatory Agency (MHRA). *Good machine learning practice for medical device development: guiding principles*. Silver Spring (MD): U.S. Food and Drug Administration; 2021. (Full text accessed 21 March 2025)

131 U.S. Food and Drug Administration (FDA). *Transparency of machine-learning enabled medical devices: guiding principles*. [Internet]. Silver Spring (MD): U.S. Food and Drug Administration; 2022. (Full text accessed 15 October 2025)

132 ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management. (Webpage accessed 15 October 2025)

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 in the execution of core PV tasks. 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 process controls. A risk is characterised by both the anticipated impact and the likelihood of negative outcomes.[1]

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 solutions, 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.
- A risk-based approach should be reviewed and adapted as needed at regular intervals and whenever changes in the system's performance dictate so.

3.1. Introduction

3.1.1. Regulatory considerations

Regardless of the integration of AI elements, PV systems are expected to comply with existing regulations and GVP.[2,3] In accordance with GVP, a wide range of PV processes are considered critical to achieving the goals and objectives of PV, including collection and handling of ICSRs, signal management, and periodic safety reports.[6]

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.[4,5,6] The EU AI Act[7] 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.[8] The 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).[9]

During development and other stages of an AI solution's lifecycle, applicants and developers should consider engaging actively with regulatory authorities and seek suitable scientific advice, as relevant and depending on the level of risk to individual patients, public health or the regulatory decision making. 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.[12,10] Due to its fast-moving nature, the use of AI technology in PV will pose challenges to both regulators, required to adapt and keep abreast of this evolving field,[10] and industry PV stakeholders, required to maintain regulatory compliance (see Chapter 10 on Future considerations for development and deployment of artificial intelligence in pharmacovigilance).

3.1.2. Motivation and interplay with other guiding principles

While the integration of AI systems into PV processes may help address human errors, inconsistencies and limitations, it is associated with some risks and challenges. 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.

A risk-based approach is applicable to, and influenced by, 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. Conversely, an AI solution may be risk-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 validation 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 may affect specific

groups and cause them to be under-served or biased against, and those risks should be appropriately mitigated against (see Chapter on Fairness & Equity).

3.1.3. Types of risks

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

Risks to patient safety and public health

Inadequate use of AI systems in PV, or their poor performance, may impede the fulfilment of PV objectives: detection, collection, assessment, understanding and prevention of adverse effects of medicinal products, which may come at the cost of patient safety, public health and compliance to regulatory requirements. Unreliable or inaccurate 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 AEs 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 (e.g. 'Black swan' events).[11] An initially robust AI system 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 system is intended to support.

Risks to user trust and engagement

The lack of transparency and interpretability of certain AI algorithms, or their use in tasks that are perceived as cognitively challenging for humans (e.g. causality assessment), may hinder trust and acceptance by users, including PV professionals[12] (see Chapter on Transparency). Lack of trust from users may also result from previous poor experience with AI systems of insufficient validity and robustness, leading to mistrust of AI systems 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.[13] 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 solution. Conversely, some users may put excessive trust in AI solutions, 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. It is also important to recognise that some PV problems may not require an AI solution.

Other types of risks

Other risks include misalignment or misuse,[14] and risks related to data privacy, cybersecurity, intellectual property, liability, or economic and reputational aspects.

The rest of this chapter mainly focusses on the impact that the use of AI systems in PV processes could have on patient safety. Other risks and challenges are further discussed in the chapters on Data Privacy, Fairness & Equity, Transparency, Human oversight and Governance & Accountability.

3.2. Risk assessment

3.2.1. General considerations

Organisations planning to deploy AI 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 solution's lifecycle from development to routine use.

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

Artificial intelligence technology

The level of risk may depend on the type of system used (e.g. static vs dynamic model), the underlying data (type and quality), the novelty of the technology (i.e. risks may be better characterised with older approaches) 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 (meaning the generation of outputs that may lead to seemingly coherent and convincing outputs that may be deceiving for humans-in/on-the-loop) may make the detection and mitigation of issues more challenging and require consideration of further guardrails (see Section on Risk mitigation). Multi-agent systems may carry specific risks linked to the autonomy individual agents are granted and to the inter-agent dependencies with potential cascading failures.[15]

As the AI landscape continues to evolve, so will AI-related risk areas. New risks may emerge while current challenges, including those associated with GenAI/LLMs may be addressed.

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 system intervenes (e.g. automated triage of relevant cases as a preliminary step to signal review) and whether the solution is assistive or directly supports a PV process;
- The relative importance of the model outputs in the decision-making vs other information sources or activities;
- 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 solution 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).

3.2.2. 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.

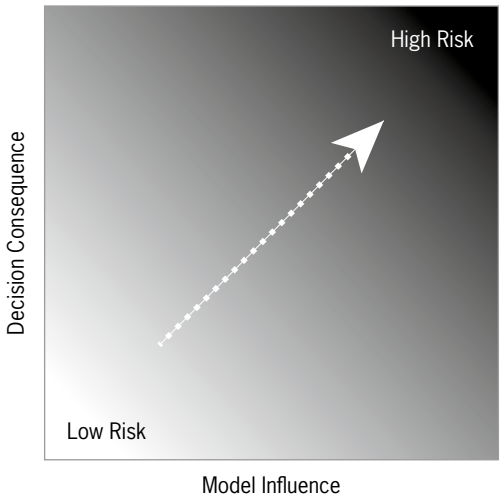
Credibility assessment framework

The US 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 medicinal products (see also Chapter Landscape analysis).[161] Similar frameworks have been proposed for the use of computational models in medical device submissions[16] or drug development.[17] 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. what will be modelled, how model outputs will be used and whether other information will be used in conjunction with the model outputs.
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 assessment involves subject matter expertise. As illustrated in Figure 4, the model risk moves from low to high as decision consequence or model influence increases.

Figure 4: Model risk matrix

Source: U.S. Food and Drug Administration.161



Algorithmic impact assessment

In Canada, the algorithmic impact assessment (AIA) tool[18] is designed to help departments and agencies better understand and manage the risks associated with automated decision systems. It is composed of a multitude of questions that consider factors within risk areas (i.e. project, system design, algorithm, decision, impact and data) and mitigation areas (i.e. consultations and de-risking and mitigation measures), which contribute to an assessment score. The value of each question is weighted based on the level of risk it introduces or mitigates in the automation project. For the risk areas, there are 65 questions with a maximum score of 169, and for the mitigation, there are 41 questions with a maximum score of 75. The score percentage range determines the impact level of the automated decision system into four levels:

- Level I – little to no impact (0% to 25%);
- Level II – moderate impact (26% to 50%);
- Level III – high impact (51% to 75%); and
- Level IV – very high impact (76% to 100%).

The algorithmic impact assessment is required prior to the production of any automated decision system under the Directive on Automated Decision-Making.[19]

3.3. Issue detection and risk mitigation

Issue detection through continuous monitoring

Defining when to mitigate risk 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 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.[20] Testing should be based on pre-defined key performance indicators (KPIs) and acceptance criteria, considering the human performance, and account for the identified risk areas, e.g. low quality data.

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 (HITL). 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. AI-assisted human oversight, i.e. the use of AI models to help monitor the main AI solutions, may also be considered (see also Chapter on Human oversight).

Reactive mitigation approaches

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

- *HITL:* Increased or full human review/quality control (QC), indefinitely or until performance levels are back within acceptance criteria, e.g. if a seriousness detection algorithm fails to detect seriousness criteria in some cases, i.e. false negatives, mitigation could involve reviewing all cases classified as non-serious until the issue is understood and addressed;
- Model re-training: targeted re-training of the underlying models using recent or challenging examples;
- *Decommissioning* of the system when mitigation options appear inefficient or costly, in which case alternative approaches should be considered.

Other (proactive) mitigation approaches

- LLM-specific strategies, including grounding techniques such as retrieval augmented generation (RAG) or other guardrails against hallucinations[21], or contingency protocols to address multi-agent system failures;
- Articulation of the level of uncertainty or confidence scoring of AI outputs;[16]
- Approaches to *combat automation bias or complacency*,[22] e.g. mock data simulations or injection of simulated false positive outputs for verification / assessment in a training environment;
- Red teaming approaches in very high-risk situations. Red teaming is when a group of people is authorised and organised to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture. The Red Team's objective is to improve enterprise cybersecurity by demonstrating the impacts of successful attacks and by demonstrating what works for the defenders (i.e. the Blue Team) in an operational environment. This is also known as Cyber Red Team.[23]

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

3.4. Review and documentation of risk-based approaches

The risk-based approach should be reviewed and the oversight measures adapted as needed at regular pre-determined intervals or whenever the AI solution shows performance issues. The evolving nature of AI technology and the emergence of new technical options for risk mitigation also call for dynamic, adaptable approaches to risk assessment frameworks.

Finally, AI components, especially those deployed in critical PV processes, should be included in the organisation's business continuity plan. The aim is to ensure that the PV system's objectives and regulatory compliance are maintained in case of failure or performance degradation of the AI solution.

The key components of the AI-related risk management strategy should be documented (see also Chapter on Transparency), including:

- AI system risk assessment;
- Testing plan with KPIs, acceptance criteria and results of testing and validation activities including any comparative assessments;
- Planned mitigation measures including human oversight strategy and criteria for more stringent or reduced QC, and continual monitoring after deployment;
- Plans for periodic re-assessment and update of the risk management strategy;
- Business continuity plan.

Chapter 3 – References

1 Council for International Organizations of Medical Sciences (CIOMS). *Practical approaches to risk minimisation for medicinal products*. Geneva: Council for International Organizations of Medical Sciences; 2014. (Full text accessed 21 March 2025)

2 European Medicines Agency (EMA). *Guideline on good pharmacovigilance practices (GVP): Module I – pharmacovigilance systems and their quality systems*. London: European Medicines Agency; 2012. (Full text accessed 21 March 2025)

3 U.S. Food and Drug Administration (FDA). *Good pharmacovigilance practices and pharmacoepidemiologic assessment: guidance for industry*. Silver Spring (MD): U.S. Food and Drug Administration; 2005;Mar. (Full text accessed 21 March 2025).

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

5 European Medicines Agency (EMA). *Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle*. Amsterdam: European Medicines Agency; 2024;Sep9. (Webpage accessed 21 March 2025)

6 U.S. Food and Drug Administration (FDA). *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*. Silver Spring (MD): U.S. Food and Drug Administration; 2025;Jan. (Full text accessed 21 March 2025)

7 European Parliament, 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. Off J Eur Union*. 2024;L 1689:1-159. (Webpage accessed 21 March 2025)

8 European Medicines Agency (EMA). *Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle*. Amsterdam: European Medicines Agency; 2024;Sep9. (Webpage accessed 21 March 2025)

9 Government of Canada. *The Artificial Intelligence and Data Act (AIDA) – companion document*. [Internet]. Ottawa: Government of Canada; 2022. (Webpage accessed 21 March 2025)

10 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. https://doi:10.1038/d41573-022-00190-3. (Journal full text)

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

12 Ball R, Talal AH, Dang O, Muñoz M, Markatou M. Trust but verify: lessons learned for the application of AI to case-based clinical decision-making from postmarketing drug safety assessment at the US Food and Drug Administration. *J Med Internet Res*. 2024;Jun 6;26:e50274. https://doi:10.2196/50274 (Journal full text)

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

14 Department for Science, Innovation and Technology (DSIT), AI Safety Institute. *International AI Safety Report 2025*. London: UK Government; 2025;Jan29. (Webpage accessed 18 September 2025)

15 Siva Kumar RS. *Taxonomy of failure modes in agentic AI systems*. [Internet]. Redmond (WA): Microsoft Security Blog; 2025;Apr24. (Webpage accessed 18 September 2025)

16 U.S. Food and Drug Administration (FDA). *Assessing the credibility of computational modeling and simulation in medical device submissions*. Silver Spring (MD): U.S. Food and Drug Administration; 2023. (Full text accessed 21 March 2025)

17 International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *General principles for model-informed drug development (M15): draft version endorsed on 6 November 2024*. Geneva: ICH; 2024. (Full text accessed 21 March 2025)

18 Government of Canada. *Algorithmic Impact Assessment tool*. [Internet]. Ottawa: Government of Canada; 2020 (Webpage accessed 21 March 2025)

19 Government of Canada. *Directive on automated decision-making*. [Internet]. Ottawa: Government of Canada; 2019 (Webpage accessed 21 March 2025)

20 International Society for Pharmaceutical Engineering (ISPE). *GAMP 5 guide: a risk-based approach to compliant GxP computerized systems*. 2nd ed. Tampa (FL): International Society for Pharmaceutical Engineering; 2022. (Abstract accessed 21 March 2025).

21 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 pharmacovigilance and other medical safety critical settings. Sci Rep. 2025;Jul 31;15(1):27886. https://doi.org/10.1038/s41598-025-09138-0 (Journal full text)

22 Adler-Milstein J, Redelmeier DA, Wachter RM. The limits of clinician vigilance as an AI safety bulwark. *JAMA*. 2024;331(14):1173-1174. https://doi:10.1001/jama.2024.3620 (Journal full text)

23 National Institute of Standards and Technology (NIST). Glossary. [Internet]. Gaithersburg (MD): NIST Computer Security Resource Center; 2025. (Webpage accessed 23 October 2025)

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 solution.

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 system should follow a risk-based approach.
- Quality assurance (QA) principles should apply to the conduct of the human oversight of AI solutions 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.

4.1. Introduction

4.1.1. Motivation

Human oversight is required to minimise the risk that an AI solution undermines human autonomy or causes other negative or unintended effects.[1] The principle of protection of human autonomy requires that humans remain in control of the AI systems.[2] Human agency and oversight are key requirements of trustworthy AI according to several regulatory frameworks, including the Assessment List for Trustworthy Artificial Intelligence (ALTAI), the EU AI Act, and the Canadian AIDA, for high-risk systems[4,3,4] (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 human 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 systems in PV to monitor and mitigate risks.

A key challenge and important starting point for defining an AI QA approach is to strike a balance between the efficiency boost that an AI solution 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.[5]

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 solutions by humans should be proportional to the potential impact of an undetected mistake or spurious output by the AI system.

4.2. Considerations on human involvement and oversight

4.2.1. Multidisciplinary expertise

The successful integration of AI systems into PV systems requires that multidisciplinary human expertise is mobilised as appropriate throughout the lifecycle of the solution, from development to routine use. This multidisciplinary expertise is usually obtained through a close collaboration between domain experts, which may include, as applicable, 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 Chapters on Validity & Robustness and 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 the effective integration of AI capabilities into PV processes. As such, PV professionals should be engaged in the design, development, pre-deployment and testing/piloting/revisions of AI systems to ensure that the systems are fit for purpose and widely accepted by the end-users that the PV professionals themselves will ultimately be.

4.2.2. Mechanisms of human oversight

Human oversight may serve different objectives and be achieved through governance mechanisms at different stages.[6] 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: 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. The concept of 'human on many loops', a special case of HOTL, addresses the scalability of monitoring multiple AI models.[7] 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.[6] The delineations of these three terms may vary according to sources[9] and their practical implementation may differ according to individual organisations and use cases. As an example, after the decision has been made to build an AI system to support a PV process (HIC), human oversight may be exercised as early as during the development phase to help define the system's context of use or support the identification or development of reference datasets (HOTL) and, when deployed, to perform QCs 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.

4.2.3. 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 its performance (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 system 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 QCs. Intermediate approaches may also be envisaged where, for instance, an AI system flags cases it struggles with to a human specialist.

The metrics and KPIs used to monitor the performance of deployed AI solutions should be pre-defined as part of the testing and validation 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), in complex or ambiguous cases, 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 deployment 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 solution can be monitored once deployed. In a static AI system, one could perform a 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 systems with high performance may also warrant specific monitoring strategies as humans are more prone to miss very rare errors than frequent ones ('low prevalence effect').[8]

4.3. 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. On the other hand, the fast-evolving pace of AI capabilities and the dynamic nature of AI performance may make it challenging for organisations to accurately forecast staffing needs and maintain optimal and sustainable human resource models. The evolving landscape 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 systems 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 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 QC 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.[9] 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.[6]

Beyond PV professionals, staff working in QA also need to develop an understanding of the organisation's human oversight strategy and of approaches to validating AI systems, 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 & Accountability). Structured competency development programs with defined learning and career progression frameworks should be considered.

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). Cross-functional training between e.g. PV professionals, data scientists, and QA teams may be a worthwhile approach. 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. For example, staff should be trained on identified areas where the system's outputs may require human review or decision, due to known limitations. Training modalities (e.g. classroom-based vs online, live vs asynchronous) should be adapted to the system's complexity, limitations, the supported use case or task including decision and action points and the specific needs of the organisation or the individual.[10,11,12,13]

Chapter 4 – References

1 European Commission. *Ethics guidelines for trustworthy AI.* [Internet]. Brussels: European Commission; 2019 [Internet]. Brussels: European Commission; 2019. (Webpage accessed 21 March 2025)

2 World Health Organization (WHO). *Ethics and governance of artificial intelligence for health: WHO guidance.* Geneva: World Health Organization; 2021. (PDF accessed 19 September 2025)

3 European Parliament, 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. Off J Eur Union.* 2024;L 1689:1-159. (Webpage accessed 21 March 2025)

4 Government of Canada. *The Artificial Intelligence and Data Act (AIDA) – companion document.* [Internet]. Ottawa: Government of Canada; 2022 (Webpage accessed 21 March 2025)

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

6 Enqvist L. 'Human oversight' in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology. 2023;Jul3;15(2):508-535. https://doi.org/10.1080/17579961.2023.2245683 (Journal full text)

7 Hillis JM, Payne K. Health AI needs meaningful human involvement: lessons from war. *Nat Med.* 2024;30:3397-3398 https://doi.org/10.1038/s41591-024-03311-0 (Journal abstract)

8 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. J Vis. 2008;Nov1;8(15):15. https://doi.org/10.1167/8.15.15 (Journal full text)

9 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. https://doi:10.1007/s40264-018-0746-z. (Journal full text)

10 Keen PG. Decision support systems: a research perspective. In: *Decision Support Systems: Issues and Challenges*. Proceedings of the International Institute for Applied Systems Analysis (IIASA) Conference Series. 1980;Jun23;11:23-27. (E-book abstract accessed 15 October 2025)

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

12 Power DJ. A brief history of decision support systems. DSSResources.com. 2007;Mar10;3. (Webpage accessed 21 March 2025)

13 Turban E. *Decision support and business intelligence systems*. New Delhi: Pearson Education India; 2011.

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 QA 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 AI systems 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 AI 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 generalisable performance across the intended deployment domain, assess usability, and identify circumstances where the model may underperform.
- Many PV applications focus on recognition of rare events or patterns (e.g. safety signals and duplicates) and may require enrichment of test sets with the event of interest. If so, special care should be taken to attempt to ensure that performance evaluation results generalise to real-world settings.

5.1. Introduction

Ensuring the validity and robustness of AI solutions is central to ensuring patient safety, 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 systems regardless of whether they develop them in-house or acquire them from other organisations. 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 models 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 organisation. In general, this will be considered separately from ensuring the validity and robustness of the core AI model (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 the ability and approach to leveraging AI. 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 AE reports is for the most part voluntary, and reporting practices vary over time, between organisations and types of reporters. This may impact the types of AEs 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 generalisability of AI models 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 organisations (see also Chapter on Fairness & Equity).

Generally, the variable quality and consistency of AE 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 regulation, and regulatory expectations regarding validity and robustness may differ from those of the business itself.

Performance evaluation and testing are crucial considerations in ensuring the validity and robustness of AI models and the focus of this chapter. It will usually be more effective to account for the same considerations also during development and training of AI models. For example, bias mitigation for ML classifiers[1] may improve performance and so may cost-sensitive learning when different types of errors are associated with different costs.[2,3] At the same time, it may not be necessary or feasible to do so to achieve good performance. For example, LLMs can be capable of zero-shot learning, with solid performance on language tasks for which they have not been specifically trained. Also, PV organisations may be offered already developed AI systems where they cannot influence AI model development. In all these scenarios, it remains important to ensure and demonstrate adequate performance on the relevant tasks in independent testing with conditions reflecting the intended use.

5.2. Specification and design

5.2.1. 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[4]. Similarly, if an AI model for recognising AEs in free text is intended for broad use, its evaluation should include reports related to various medicinal products and AEs, from both patients and health professionals, in relevant languages, etc. Ideally, there should be sufficient representation of relevant types of

data in the training and test sets to promote adequate and generalisable performance across the intended deployment domain, assess usability, detect biases, and identify circumstances where the model may underperform. For more capable AI systems, it may also be relevant to specify refusal policies which determine which tasks the system will permit and refuse.

Design of AI systems may also account for an AI model's susceptibility to overfitting, computational complexity and robustness to outliers, especially if test sets will not be large and diverse enough to reliably capture their impact during performance evaluation. Complex methods highly dependent on skilful design and deployment by human experts may not readily transfer to similar 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.

5.2.2. Multidisciplinary collaboration

Ensuring the validity and robustness of AI models 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.

AI systems addressing the complex relationships between drugs and AEs often require a HITL, 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 that are central to causality assessment[5,6] but 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 AE encoding may lend themselves to automation with minimal human intervention.

5.2.3. Definition of reference standards

Test sets must be aligned with the intended deployment domain(s) 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, in which case a set of real examples may be classified (annotated) by a human specialist. Approaches to mitigate inconsistencies in such annotations 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 execution of the PV task of

interest, sometimes bringing value by harmonising and making explicit decision processes that may otherwise remain implicit and variable within an organisation. Based on the size and scale of the project, 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 model; for example, annotations may be performed preferably by individuals blinded to the specific AI model 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 organisations may have different internal conventions regarding how strong the conviction should be that two AE reports refer to the same event for them to be classified as suspected duplicates. This may vary even within an organisation 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 NLP 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. The ambiguity is even more pronounced for signal detection and causality assessment tasks, where human experts may often disagree on whether there is sufficient evidence of a causal association between a drug and an AE (at a given point in time).

A general challenge in PV has been ensuring sustainable and reusable access to reference sets. 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.

5.3. Performance evaluation

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

Many of the metrics relevant to performance evaluation for AI models 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 ranked orderings, unsupervised learning, or content generation.

AI systems, like humans, will typically not achieve perfect performance on more complex classification tasks. In fact, there can be an inherent ambiguity as to what is the correct classification of some instances in real-world applications also for domain experts (e.g. for lack of information). Therefore, performance is typically assessed statistically for a sample of cases referred to as the test set. *Recall* measures how many of the target events

are correctly identified (*recalled*) by the AI solution. *Sensitivity* is a synonym. *Precision* measures the proportion of target events among all events highlighted by the AI solution. *Positive predictive value* (PPV) is a synonym.

The balance between precision and recall (and correspondingly between sensitivity and specificity) can typically be tweaked and 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.[i]

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). Any user-driven design decisions should be fixed and finalised before AI model developers first access test sets. This is especially important for more complex methods with numerous analytical choices regarding model architecture, hyper-parameters, and model initialisation.[7] Various potential sources of dependence between development and evaluation should 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 AEs, and the deployment domain is broader.[8]

5.3.1. Benchmarking

Ideally, performance should be compared against relevant benchmark methods, if available. For example, AI-based signal detection methods may currently be compared against standard disproportionality measures, if this is an organisation's baseline method. 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,[9,10] sets of established adverse drug reactions,[11,12,13,14,15] and clinically relevant drug-drug interactions.[16] 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 model for different parts of the deployment domains (See also Chapter on Fairness & Equity). Along the same lines, sensitivity

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

analyses can help assess the robustness of the AI model and its evaluation to variations in specification and design.

5.3.2. Special considerations for low-prevalence settings

Many PV applications focus on recognising 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 analysing 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.[17] 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.[ii]

This low prevalence of positive controls (i.e. class imbalance) limits our ability to achieve accurate performance evaluation and requires special care and consideration. 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. 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. For a deeper elaboration regarding this, see for example Norén et al 2025.[18]

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 model, may less likely be included in the test set. This does not mean that rebalancing approaches should necessarily be avoided but if they are used, this should be acknowledged and critically assessed.

Similarly, *precision* 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 as the baseline prevalence of the target event is inflated. For reliable precision estimates, the prevalence of positive controls in the test sets should match as far as possible that of the intended deployment. For a specific AI model, precision is straightforward to estimate by applying the AI model to a random sample and annotating all highlighted instances. However, such test sets for precision are tied to the AI model in question and will need to be developed again, or at least extended, if the model is modified. They are not useful to estimate recall.

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.

[ii] $0.10 \times 1/10^6$

5.3.3. 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 model. Access to and ability to inspect representative, concrete examples of an AI model's classification of individual instances in a test set is also important. 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' which has not been redacted by the method, this may undermine end-users' trust in the solution, even if overall recall is excellent, because the error seems trivial. On the other hand, if the false negative is 'AF' and it is hard to know, even for a domain expert, from the surrounding text if these are initials or an abbreviation for *atrial fibrillation*, then one should perhaps consider the overall precision metric to be conservative. Review of correctly classified instances may in turn give insights regarding an AI system'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 method, 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 review 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.

5.3.4. Unsupervised learning

For AI systems performing unsupervised learning like cluster analysis, patterns are identified in a data-driven manner without access to human-annotated reference sets. They require other approaches to performance evaluation. In some cases, one may rely on human subjective review and assessment of the AI output, but then potential cognitive biases must be considered and mitigated. A possible solution may be to present the results of several different AI models to a blinded, domain expert and ask which one they prefer. There are also performance evaluation approaches specifically designed for unsupervised learning like intruder detection analysis where domain experts are asked to spot an unrelated "intruder" among items an AI model has grouped as related, and coherence is measured by the intruder detection rate.[19]

5.3.5. Generative output

GenAI models can create open-ended, often longer, pieces of text (or other content) that may be used without restriction or further post-processing into a pre-defined set of options (e.g. yes/no or MedDRA Preferred Terms). Examples of such applications include text summarisation, translation, report generation, and lay-language rewrites. There typically does not exist a single correct output and aspects of the text such as fluency and coherence may need to be evaluated along with task-specific metrics. This is a rapidly evolving field and at the time of writing; multiple evaluation metrics are often used in parallel. For example, generic metrics for readability and toxicity may be obtained, and when a ground truth reference text exists (e.g. for translation or summarisation tasks), measures of syntactic (e.g. Bilingual Evaluation Understudy or BLEU) or semantic (e.g. BERT score) text overlap can be computed. Similarly, retainment of key entities can be measured against human-annotated reference sets, if available. Human evaluation may also be obtained prospectively but should then be designed and executed with care, as in the context of unsupervised learning discussed

above; pairwise preference testing or evaluation of success rates for human task execution supported by an AI system output are among the options. While human evaluation remains the gold standard, generative LLMs can also be prompted to rate AI system outputs on various dimensions as part of a holistic performance evaluation. Such LLMs-as-a-judge approaches scale well but require a strong evaluator model that should ideally be distinct from the model under evaluation, and human calibration is typically required. Recent examples of performance evaluation for GenAI applications in PV include that of summarising AE reports[20] and that of LLM-generated clinical reasoning in the context of individual case causality assessment for COVID-19 vaccine reports.[21]

5.3.6. Non-deterministic systems

A deterministic AI system will always generate the same output for a given input. Predictive models like support vector machines and decision trees are of this nature. So are certain LLMs (e.g. masked encoders like BERT) and other deep neural networks used for classification tasks, once their weights have been fixed at the end of training / fine-tuning. This is so, even though their model fitting may include stochastic components, and re-training a model on the same data may result in different parameters.

In contrast, methods for unsupervised learning like cluster analysis, network analysis, or data-driven derivation of semantic vector representations of AEs may be stochastic and generate different results when executed repeatedly on the same data. The same is true for generative LLMs, which will typically produce different outputs for the same prompt, without changes to the underlying models. For such AI models, stability of output can be a key additional performance metric, reflecting how similar the results of repeated analyses are. Sometime, the level of stochasticity can be directly controlled by hyper-parameters. There may also be specific measures taken to reduce the variability in output depending on the method. While repeatability of results can sometimes be artificially ensured through seeding the pseudo random number generator, this may not be possible for proprietary models and does not improve the inherent (in)stability of the AI solution, which should be evaluated. Whether a non-deterministic AI system is appropriate for a given application will depend on its context of use and the possible negative effects of variability in the output, according to the principles of a risk-based approach.

5.4. Assessing artificial intelligence systems with human-in-the-loop

Many AI systems 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 requiring a different nature of testing than described above. To date, there is 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 system is integrated with the PV system and how humans will interact with the system.[22] For example, performance evaluation for an NLP-based system to identify and extract AEs 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[23]).

5.5. Continuous integration and deployment

Deployed models should be monitored in real-world use with a focus on maintained or improved performance. In some circumstances, there may be reason 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 QCs 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 model. 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 system incorporates ML components or not. Its frequency should follow the risk-based approach and may include data-driven safeguards to identify, for example, substantial data drift or performance degradation. Such observations may trigger remedial actions that could include additional evaluation, and possible retraining, stopping use of the algorithm and/or introducing QC measures to maintain confidence in its results. In the case of dynamic AI models continually fine-tuned or otherwise updated in (near) real-time after deployment, automated detection of model drift may also trigger re-validation activities. Documentation of activities and acceptance criteria for re-introducing AI solutions under such circumstances may also be required. As an example, they may include 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 RWD. To support this approach, the US FDA, Health Canada, and Medicines and Healthcare products Regulatory Agency (MHRA) described a "Predetermined Change Control Plan" for ML-enabled device software functions (ML-DSF). Their general principles might conceivably be applied to AI systems 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 (QMS) governing the PV process being modified.

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

Chapter 5 – References

1 Hort M, Chen Z, Zhang JM, Harman M, Sarro F. Bias mitigation for machine learning classifiers: a comprehensive survey. *ACM J Responsib Comput.* 2024;1(2):Article 11. https://doi.org/10.1145/3631326 (Journal full text)

2 Elkan C. The foundations of cost-sensitive learning. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01), Vol. 2*. San Francisco (CA): Morgan Kaufmann Publishers; 2001;973-978. (Journal full text)

3 Araf I, Idri A, Chairi I. Cost-sensitive learning for imbalanced medical data: a review. *Artif Intell Rev.* 2024;57:80. https://doi.org/10.1007/s10462-023-10652-8 (Journal full text)

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

5 Bradford Hill A. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295-300. https://doi:10.1177/003591576505800503 (Journal full text)

6 Meyboom RHB, Hekster YA, Egberts ACG, et al. Causal or casual? *Drug Saf.* 1997;17:374-389 https://doi.org/10.2165/00002018-199717060-00004 (Journal full text)

7 Duin RP. A note on comparing classifiers. *Pattern Recognit Lett.* 1996;May 1;17(5):529-536. https://doi.org/10.1016/0167-8655(95)00113-1 (Journal abstract)

8 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. https://doi.org/10.1162/99608f92.5a8a3a3d (Journal full text)

9 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. Scient Data. 2014;Nov11;1(1):1-0. https://doi.org/10.1038/sdata.2014.43. (Journal full text)

10 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 Saf. 2023;Feb;46(2):109-120. https://doi.org/10.1007/s40264-022-01258-0 (Journal full text)

11 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. https://doi.org/10.1007/s40264-012-0002-x (Journal abstract)

12 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. https://doi.org/10.1007/s40264-013-0097-8 (Journal abstract)

13 Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Research. 2016;Jan4;44(D1):D1075-1079. https://doi.org/10.1093/nar/gkv1075 (Journal full text)

14 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. https://doi.org/10.1038/sdata.2018.1 (Journal full text)

15 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 Saf. 2021;Jan;44:83-94. https://doi.org/10.1007/s40264-020-00996-3 (Journal full text)

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

17 Painter J, Powell G, Bate A. *PVLens: enhancing pharmacovigilance through automated label extraction.* [preprint]. *arXiv.* 2025;Mar26. https://doi.org/10.48550/arXiv.2503.20639 (Journal full text)

18 Noren N, Meldau E-L, Ellenius J. Critical appraisal of artificial intelligence for rare-event recognition: principles and pharmacovigilance case studies. arXiv:2510.04341 [cs.LG] https://doi.org/10.48550/arXiv.2510.04341 (Journal full text)

19 Chang J, Boyd-Graber J, Gerrish S, Wang C, et al. *Reading tea leaves: how humans interpret topic models*. [unpublished manuscript]. 2009. (Full text accessed 27 April 2025)

20 Dowdy K, Hoffman A, Giles T, Kugele D, et al. Summarizing FAERS Narratives with Generative AI: Methods, Resource Requirements, and Quality Assessment. FDA, SYMPOSIUM, SCIENTIFIC COMPUTING + DIGITAL TRANSFORMATION, 2024. Center for Drug Evaluation and Research, U.S. Administration, Booz Allen Hamilton. (PDF accessed 23 September 2025)

21 Abate A, Poncato E, Barbieri MA, et al. Off-the-shelf large language models for causality assessment of individual case safety reports: a proof-of-concept with COVID-19 vaccines. *Drug Saf*. 2025;48:805-820. https://doi.org/10.1007/s40264-025-01531-y (Journal full text)

22 The US Food and Drug Administration (FDA). 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)

23 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 Saf. 2024;Jan;47(1):71-80 https://doi.org/10.1007/s40264-023-01367-4 (Journal abstract)

CHAPTER 6.
TRANSPARENCY

Principle

Transparency regarding AI involves disclosing information between organisations 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, perceived benefits, and impact on their rights.

Key messages

- Declaring when and how AI systems 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 sufficiently described, including their model architectures, expected inputs and outputs, and the level and type of human-computer interaction.
- To give a clear picture of an AI model's effectiveness and limitations in a PV application, performance evaluation results for the specific task should be presented and describe the scope and nature of the test set(s), 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. A lack of explainability should be acknowledged and discussed.

6.1. Introduction

Transparency provides stakeholders with relevant information regarding the nature and use of an AI system. 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 organisations not involved in their development to inspect and scrutinise the design and performance of AI systems, 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 system lifecycle. For example, during the design phase the business owner should be transparent toward developers regarding the specification and requirements for an AI system, whereas in the pre-deployment phase developers should be transparent toward the business owner regarding the nature and performance of an AI system. During routine use, the most important

form of transparency may be from the organisation toward end users (and in some cases regulatory authorities).

6.2. 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, awareness, and responsibility among stakeholders, including developers, PV professionals and decision makers, regulatory authorities, HCPs, and patients. Confidence scores and other metrics communicating the AI model's certainty in a prediction or output can be a valuable component of such disclosure. However, the validity and robustness of such scores and metrics must also be ensured and their meaning clearly communicated to end users.

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 organisations, who in turn need to be transparent toward regulatory authorities. At the same time, those individuals who utilise AI solutions to process or analyse 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 AE. 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.

6.3. Transparency regarding the artificial intelligence model

Ensuring sufficient transparency of the AI models used in PV 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 also an ethical imperative, ensuring that all parties understand the systems 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 or on-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 or multi-agent system, 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. For AI solutions based on GenAI models, 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 bespoke ML components have been developed should be disclosed. This would include their size, scope, annotation guidelines, quality assurance, and creation date, along with reflections on how well they align with the intended deployment domain and a justification for their use.
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 with 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, as well as metrics conveying the AI model's certainty regarding specific outputs.
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, or known biases or under-served populations (See Chapter on Fairness & Equity).

To allow other developers and researchers to fully replicate an AI model and possibly even modify it for further use, an organisation 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.

6.4. 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 characterises 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".[1]

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".[2]

The OECD Transparency and Explainability Principle 1.3 states:[3]

"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 Appendix 3.

6.4.1. Benefits of explainability

Explainability can be beneficial because it may:

- Nurture trust in an AI system, by enabling stakeholders to make sense of and contextualise an AI solution's output;
- Allow individuals affected by an AI system's output to challenge and influence the outcome;
- Support and speed up human decision-making which builds on or integrates an AI system output;[4,5]
- 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;[6]
- Enable more complete documentation, audit, and human oversight of AI systems;
- 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;[7,8]
- 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.[9] They may range from the PV and QA 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.

6.4.2. 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.[9] This is also referred to as *ante-hoc* explainability. Examples may include lower-dimensional decision trees, rule-based classifiers, and regression models.

In contrast, a growing field of research seeks to obtain *post-hoc* explainability 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 on top of the AI solution[10] to trace and explain the basis for a specific, already generated output. Some post-hoc explainability 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 methods that provide post-hoc explainability of a specific output by fitting simpler, inherently interpretable models to the local context of a specific output. For examples of specific methods in use at the time of writing this report, please see Appendix 4.

When a post-hoc method is used to gain explainability, it must itself comply to the applicable regulatory requirements for computerised systems. In other words, the post-hoc explainability method should be verified regarding its fitness for purpose and the process integrating such methods must be validated. Post-hoc explanations offer an approximate understanding of the relationship between the data and the predictions.[11] The explanations can be imperfect or incomplete and/or provide only a partial explanation.

Generative LLMs can be prompted to produce apparent rationales for their outputs. However, such rationales are language artifacts that do not provide direct access to the model's internal computation and may merely be post-hoc rationalisations. For PV use, organisations may prefer evidence-anchored justifications (explicit links to input spans or source documents) over free-form 'chain-of-thought,' and require faithfulness tests before accepting LLM-produced rationales as part of the audit trail.

6.4.3. 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 AI system they are considering can provide it, bearing in mind that explainability may not be an appropriate goal for all AI solutions.[12]

The level of explainability of an AI solution's output should not be the sole determining factor for model selection. For example, applications like machine translation depend on the higher capabilities and improved performance offered by deep neural networks and typically do not require inherent explainability. A negative impact of limited explainability may be mitigated by ensuring high transparency regarding other aspects of the AI system and its performance, coupled with extra care to achieve validity and robustness and human oversight.[13] 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.[14]

Similarly, while explainability of an AI model's output can sometimes help identify issues with validity and robustness or fairness and equity, explainability alone does not prove that the system is fit for purpose, nor does it vouch for the trustworthiness of the system.[15] It attempts to clarify what factors led to a specific output but is not indicative of an AI model's general performance or of its fairness and equity. For example, even if an inherently explainable AI model 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.[15] Also, explainability is no guarantee of transparency – an organisation 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[9] and explanations must be accessible to people with a wide range of literacy and educational attainment.[10] Since humans will need to process and contextualise 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.

6.5. Transparency regarding performance

Transparency regarding an AI model's assessed performance of a specific PV task communicates how well an AI model operates in practice and complements the insights into the design, implementation, and decision-making processes provided through model transparency. It provides a bridge between theoretical capability and practical utility. Without a clear view of how an AI model 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 model'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, including quantitative metrics, qualitative examples, and comparisons to benchmarks. Thereby organisations provide the necessary context to build trust and appropriate reliance on AI systems. This transparency allows for informed decision making by PV professionals and decision-makers, 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 clear documentation of the data used for performance evaluation, including the rationale for its selection, how it was acquired, cleaned and transformed, and any processes for managing missing or erroneous data.

Table 4 outlines relevant aspects to disclose to ensure transparency regarding the estimated performance of an AI model. 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 model

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). 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 customised 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 and 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).

Chapter 6 – References

1 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. Geneva: International Organization for Standardization; 2020. (Abstract accessed 21 March 2025)

2 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)

3 Organisation for Economic Co-operation and Development (OECD). *Transparency and explainability (Principle 1.3).* [Internet]. Paris: OECD AI Policy Observatory; 2024. (Webpage accessed 21 March 2025)

4 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 Saf. 2020;Sep;43:905-915. https://doi.org/10.1007/s40264-020-00945-0 (Journal abstract)

5 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)

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

7 Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York (NY): Association for Computing Machinery; 2016;Aug13:1135-1144. https://doi.org/10.1145/2939672.2939778 (Journal full text)

8 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. Inf Fusion. 2023;Aug1;96:156-191. (Journal full text) https://doi.org/10.1016/j.inffus.2023.03.008

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

10 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. https://doi.org/10.1038/s41746-020-0254-2 (Journal full text)

11 Pinheiro LC, Kurz X. Artificial intelligence in pharmacovigilance: a regulatory perspective on explainability. Pharmacoepidemiol Drug Saf. 2022;Dec1;31(12):1308-1310. https://doi:10.1002/pds.5524 (Journal full text)

12 Royal Society. *Explainable AI*. [Internet]. London: Royal Society; 2019;Nov. (Website accessed 11 August 2024)

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

14 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;Nov22;1(2):1-9. https://doi.org/10.1162/99608f92.5a8a3a3d (Journal full text)

15 Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;Jan1;19(1):121-127. https://doi.org/10.1136/amiajnl-2011-000089 (Journal full text)

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. Legislation, regulations and guidance documents provide certain 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

- Application of AI in PV that may involve protected data should consider the standard principles for research activities involving human subjects.
- The use of AI applications in PV requires additional attention to ensure that appropriate safeguards are in place to address data privacy requirements.
- The applications of ethical principles most relevant for the use of AI in routine PV are data privacy, fairness, and equity.
- PV professionals should recognise 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.

7.1. Introduction

Although data privacy has been recognised as an implicit legal right for well over a century,[1] 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, organisation, 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 1970.[2] In the US, public concerns about the potential misuse of collected data led to the US Privacy Act (1974),[3] which established a code of fair information practices that governs the collection, maintenance, use, and dissemination of information about individuals that is maintained in systems of records by federal agencies. 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.[4] 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. Over the following decades, the guidelines have influenced many subsequent data protection regulations/laws, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. 104-191, 110 Stat. 1936 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.

7.2. Ethical considerations

While data privacy concerns are widely recognised in the use of AI, at the time of publication, there has been limited attention paid to specific ethical considerations applied to the use of AI in PV.[5] Many publications that refer to ethics and AI, such as the WHO Guidance, Ethics and Governance of Artificial Intelligence for Health,[6] emphasise several basic principles that were first elaborated in the Belmont Report (1979).[7] The Report was designed to provide an ethical framework for clinical and behavioural research; however, it has subsequently been applied to certain PV activities (acknowledging that most PV is not typically 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 that individuals are free to decide whether to participate in research. In clinical research, informed consent is recognised as one application of this principle. In the setting of PV, the principle may be applied to data privacy and the right to control one's personal information (acknowledging that for the purpose of PV, statutes may infringe). Beneficence emphasises that research should be designed and conducted to maximise benefits and minimise harm to participants. As applied to PV, this principle is captured in ongoing benefit-risk assessment. Justice focuses upon fairness and non-discrimination. In PV, justice can be applied as fairness and equity, i.e. that the benefits of PV knowledge be distributed equitably across populations.

The Report's original intent was to establish universal principles applicable to clinical and behavioural research, not to address public health activities. In the succeeding decades, as certain areas in public health have expanded (e.g. through access to data and associated research methods that were unavailable in the 1970's), its principles have been applied to areas such as disease surveillance and PV. As an example, post-authorisation safety studies are often required as a condition of product licensure and may use RWD sources to generate real-world evidence (RWE). The report of the CIOMS Working Group XIII is on *Real-World Data and Real-World Evidence in Regulatory Decision Making*.[8]

As applied to AI applications in PV (e.g. training and validity testing, and generalisability), data privacy (drawn from Respect for Persons[6]) and fairness and equity (drawn from Justice[6]) are particularly relevant for ethical considerations. Fairness in PV requires non-discriminatory practices, ensuring that findings are representative of the population exposed to a product, and equity is essential to ensure that PV benefits are shared broadly, a topic explored further in the following chapter.

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 1996 HIPAA Pub. L. 104-191, 110 Stat. 1936, used in the United States, and the GDPR, 2016, which is employed in the European Union (EU). 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[9] and was developed to ensure data privacy during the transition of medical information from analogy to digital. HIPAA also introduced administrative standards for health care data to improve efficiency in the health care industry. In short order, the rapid adoption of digital technologies in health care (e.g. EHRs) and the interest in using electronic data for research and other purposes led to additional regulations to support the use of EHRs according to standards that would ensure administrative efficiency while protecting patient privacy and security (HIPAA Privacy Rule, 2000; Security Rule, 2003; Health Technology for Economic and Clinical Health [HiTech] Act Breach Notification Rule, 2009). The HIPAA Privacy rule addresses the use and disclosure of individuals' health information, called protected health information (PHI), by organisations subject to the Privacy Rule, called "covered entities", as well as standards for individuals' privacy rights to understand and control how their health information is used. A major goal of the Privacy Rule is to assure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and wellbeing.

HIPAA emphasises the confidentiality, integrity and availability of health data, and requires regulated entities to make reasonable efforts to limit the use, disclosure of, and requests for PHI to the minimum necessary amount to accomplish a particular purpose. It specifies patients' rights to access and amend PHI. To protect patient confidentiality, HIPAA recognises 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 or disclosed 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 approved for use in the US, routine PV activities conducted by application holders are typically exempt from certain HIPAA requirements for patient authorisation to disclose and use PHI. Medicinal products are governed in the US FDA regulations that require, among other things, monitoring the quality and safety of US FDA-regulated products, which is conducted in part through AE reporting, product tracking, recalls, and post-marketing surveillance. While certain PV activities are exempt from certain HIPAA requirements, 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 anonymisation of data (employed where possible); use of technical, administrative, and physical safeguards to prevent unauthorised access, use, and disclosure of PHI; and requirements for Business Associate Agreements (where vendors or partners are engaged by a covered entity). Within the US, the US FDA and Centers for Disease Control and Prevention (CDC, Atlanta) are responsible for public health

matters as part of their official mandate. Among its responsibilities, the US FDA is responsible for "protecting the public health by assuring the safety, efficacy, and security of human and veterinary drugs, biological products, medical devices ...". The CDC responsibilities include protecting "America from health, safety and security threats, both foreign and in the U.S.", including whether diseases are chronic or acute, curable or preventable. This is accomplished in part by conducting "critical science and providing health information that protects (the US) against expensive and dangerous health threats". The Privacy Rule permits covered entities to disclose PHI, without authorisation, to public health authorities that are legally authorised to receive such reports for the purpose of preventing or controlling disease, injury, or disability. Covered entities are generally required to reasonably limit disclosures of PHI made without individual authorisation for public health activities to the minimum amount necessary to accomplish the public health purpose.[10]

In contrast to public health authorities and the private sector, academic involvement in PV is generally conducted as a research activity, e.g. Post-Approval Safety Studies (PASS), and are subject to different oversight, including the use of institutional review boards (IRBs, aka ethical review boards) to assure that studies meet appropriate ethical standards, and are to mitigate data privacy concerns, including data use agreements, where applicable when collaborating with other organisations. This includes the use of de-identified and limited data sets, and compliance with both HIPAA and the Common Rule (45 CFR 46, Subpart A – HHS Policy for Protection of Human Subjects, which governs the ethical conduct of research involving human subjects), along with other applicable requirements.

Example: General Data Protection Regulation

The GDPR (Regulation [EU] 2016/679) is a comprehensive regulation overseeing personal data protection in the EU and succeeds the earlier Data Protection Directive (Directive 95/46/EC), which was issued contemporaneously with HIPAA, 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".[11]

The GDPR incorporates principles such as lawfulness, fairness, transparency, accuracy and integrity, purpose limitation, data minimisation, confidentiality, and storage limitation. Compliance is a major feature of the GDPR with organisations such as pharmaceutical companies required to have a Data Protection Officer responsible for overseeing compliance. Penalties for non-compliance may be significantly greater than those under HIPAA, with fines up to 4% of global turnover.

Several safeguard measures may be used to help ensure data privacy, such as data encryption (preventing access without a decryption key), anonymisation (where possible) or 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 AEs without consent in order to

ensure oversight of the safety and effectiveness of medicinal products – provided that certain safeguards are in place (and subject to the individual laws of the respective member states). 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 minimisation as well as administrative, technical and organisational measures to protect personal data.

In fulfilling its responsibilities to assure the safety, effectiveness, and quality of medicinal products authorised for use in the EU, the EMA 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 emphasises the principles of data minimisation, purpose limitation, lawfulness, fairness and transparency in its data use. 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 (EEA) 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 or Independent Ethics Committee (IEC) oversight and an emphasis upon adherence to principles of data minimisation 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.

7.2.1. Other data privacy laws regulations

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

Table 5: Data privacy regulations for secondary use of data in Brazil

Source: CIOMS Working Group XIV

Aspect	Brazil
Governing Law	General Data Protection Law (Lei Geral de Proteção de Dados, LGPD), Law No. 13.709/2018
Health Data Classification	Sensitive personal data includes health data; additional safeguards for children and adolescents
Consent Requirements (Research & PV)	Consent required in principle; exceptions allowed (e.g. legal/ regulatory obligations, implementation of public policies, protection of health, and research by authorised institutions)
Secondary Use of Data (e.g. RWE)	Permitted when justified by legal bases (e.g. regulatory obligations, public policies, health protection, or research); ANVISA's regulatory activities exempt from consent
De-identification Standards	Anonymisation and pseudonymisation encouraged to reduce reliance on consent; focus on transparency, purpose limitation, and data minimisation
Cross-border Data Transfer	Applies to processing of personal data of individuals in Brazil regardless of processor location; international transfers permitted if LGPD requirements and safeguards are met

Aspect	Brazil
Pharmacovigilance Exemptions	Adverse event reports and technical complaints handled by ANVISA without consent under regulatory/legal obligations and health protection grounds
Regulator Guidance on Biomedical Use	LGPD implemented nationally; ANVISA Ordinance No. 1,184/2023 establishes Personal Data Protection Policy (including inventories, security measures, impact reports, contracts compliance, and data protection culture)
Oversight Body	National Data Protection Authority (ANPD); implementation in health sector by ANVISA
Key References	LGPD, Law N°13.709 (2018);[12] ANPD regulations;[13] ANVISA Ordinance No. 1,184/2023[14]

Table 6: Data privacy regulations for using secondary data in China

Source: CIOMS Working Group XIV

Aspect	China
Governing Law	Personal Information Protection Law of the People's Republic of China (PIPL); Data Security Law of the People's Republic of China, Cyber Security Law of the People's Republic of China.
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 anonymisation
Cross-border Data Transfer	Strict rules: security assessments, contracts, individual consent; limited adequacy
Pharmacovigilance Exemptions	AE reporting permitted but must minimise identifiable data
Regulator Guidance on Biomedical Use	PIPL + draft health data governance rules; evolving
Oversight Body	Cyberspace Administration of China (CAC) (oversees cybersecurity and data protection) and National Health Commission (regulatory authority establishes and implements standards for medical and health data)

Aspect	China
Key References	PIPL (2021);[15] CAC draft regulations on health data;[16] State Council health data measures (2022)[17]; Data Security Law of the People's Republic of China;[18] Cyber Security Law of the People's Republic of China;[19] Regulation on Network Data Security Management;[20] Measures for the Security Assessment of Outbound Data Transfer;[23] Provisions on Promoting and Regulating Cross-border Data Flows;[21] Law of the People's Republic of China on Basic Medical and Health Care and the Promotion of Health;[22] Regulation of the People's Republic of China on the Administration of Human Genetic Resources:[23] Measures for the Standard Contract for the Outbound Transfer of Personal Information.[24]

Table 7: Data privacy regulations for secondary use of 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	Pseudonymisation encouraged; full anonymisation for broader reuse
Cross-border Data Transfer	Allowed to countries with adequacy or with SCCs/ Binding Corporate Rules (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 and European Data Protection Board (EDPB)
Key References	GDPR (Regulation EU 2016/679); EDPB Guidelines 03/2020; EMA Module VI (GVP); BDSG (Germany)

Table 8: Data privacy regulations for secondary use of 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 pseudonymised data may be used for public interest or research
Secondary Use of Data (e.g. RWE)	Allowed with pseudonymisation/anonymisation and research purpose declaration
De-identification Standards	Recognises both anonymised and pseudonymised 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);[25] Act on Anonymized Medical Data That Are Meant to Contribute to Research and Development in the Medical Field;[26] PPC Guidelines;[27] MHLW guidance on GPSP and human research ethics[28]

7.3. Practical considerations to support data privacy

As these examples indicate, there are regulations that support 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 anonymise data.

As a regulated industry, pharmaceutical companies must comply with the data privacy and reporting requirements of all countries in which their products are approved. As an example, the EMA requires adherence to GVP and to data protection principles from the GDPR. Ensuring compliance requires attention to evolving country-specific regulations, 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 customisation

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 by a covered entity require Business Associate Agreements. Under GDPR, 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 is a range of potential consequences for data breaches, from requirements for notification to data protection authorities up to and including significant fines and penalties.

7.3.1. 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. ICSRs. Additionally, LLMs, whether open or closed, permit nearly instantaneous planned (or unplanned) linking of data sources that would otherwise not have occurred, or would 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 recognised. These attributes raise the question of whether current data privacy tools are sufficient to prevent re-identification of de-identified 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, sex, and date of birth, illustrating that few data points were needed to uniquely identify individuals.[29] 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 minimised, 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 generalisation 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.[30]

The EMA and Health Canada now require public sharing of clinical trial reports as part of the drug approval process. Standards for data anonymisation 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.[31] 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). Re-identification was best informed by social media and death records, although it was uncertain if the re-identification had been successful. Based on the ≤0.09 probability risk threshold of re-identification established by EMA[32] and accepted by Health Canada, the authors concluded that existing anonymisation guidance was sufficient to provide an adequate level of data protection (defined as non-zero, but a small number of low confidence re-identifications); however, they also observed that the findings may not apply to studies of rare diseases, nor to studies that employ qualitative rather than quantitative methods for anonymisation. With rapid advances in AI, the time required to replicate the re-identification 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. Today's AI is advancing rapidly and risks for re-identification will only increase as AI methods become more sophisticated. A few examples illustrate the breadth of some recent data breaches (when AI was not as advanced as currently), 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.[33]
- 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 of the US population, purportedly with the intent of using AI to compromise US government officials.[34]
- A purposeful domestic data breach intended to embarrass a political opponent, involving a cyber-attack on an Asian health care plan that resulted in 1.5 million patient records.[35]
- An unintentional release of Indian government biometric, and other personal data, in a database containing records of 1.2 billion individuals.[36] 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 GenAI, 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 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.[37] 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. As GenAI is rapidly evolving, regulations to safeguard data privacy in GenAI will need to evolve as well.

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 ML 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 organisations that maintain closed LLMs may exercise control of prompts as well as the data contained in the models. Open LLM models lack this safeguard thereby increasing the risk for re-identification, as these models have access to diverse sources of data that are not necessarily 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 that bypass data privacy considerations or through models that are trained on personal data. Additionally, as noted above, re-identification can occur even with data that have been presumed de-identified (e.g. where postal code, birth date, gender, ethnicity have been retained). 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) or ensuring suitable controls (e.g. access controls, monitoring of inputs and outputs to mitigate data leaks).[38,39]

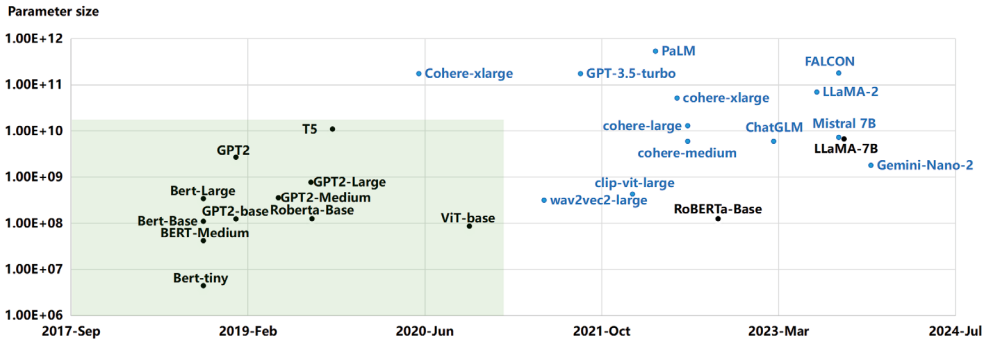
— GenAI programs can integrate otherwise discrete data sources such as census and vital statistics and public health data, which may be linked to a de-identified health record data set (e.g. in the setting of an active PV activity e.g. a post-authorisation safety study) leading to the possibility of re-identification.

— Other privacy risks may include: (i) data persistence (i.e. data being retained within the AI system and never deleted once retention time has lapsed); (ii) complexity of upholding data subject rights, such as access right if an AI system was trained on a dataset that the system no longer holds as such; and (iii) transparency: when a model is trained through data scraping (e.g. on the web) without a data subjects' knowledge.

Individual use cases of GenAI/AI must comply with local and relevant globally applicable legislation. 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 5). 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 5: State of research on privacy protection for Large Language Models (as of June 2025)

Source: Modified from Yan B et al 2025[40]



The timeline axis represents the release time of LLMs, while the vertical axis indicates the size of parameters. Blue data points represent LLMs that have received limited attention in the literature on privacy protection, while black data points signify models that have been studied alongside privacy concerns. The green background draws attention to the central cluster of LLMs that could help boost privacy protection. At the time of publication of the article, recently introduced larger LLMs had not been as fully evaluated in the scientific literature with respect to data privacy as earlier, smaller models.265

7.4. 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 minimisation, anonymisation, 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 globalised PV landscape involving more numerous and complex interdependencies (e.g. business partnerships, international vendors conducting PV activities). The ongoing challenge for PV professionals, regulatory agencies, industry, as well as PV organisations 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 re-identification, 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 having existing data privacy policies in place and adhering to data privacy regulations, efforts to mitigate risks to data privacy when applying AI to PV may include the below.

- Recognition that the technology is advancing rapidly, requiring ongoing monitoring, e.g. to assure that data de-identification measures are adequate.
- Attentiveness to policies that may be introduced by GenAI firms to mitigate the risk of re-identification.[41]
- Legislation may also impose criteria for GenAI models intended to mitigate systemic risks.[42,43]
- On-premise or private cloud deployment, advanced anonymisation techniques, federated learning architectures, differential privacy methods, and comprehensive contractual safeguards governing data handling, retention, and cross-border transfer with AI service providers.
- 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 organisations, the organisation 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.

Chapter 7 – References

1 Warren SD, Brandeis LD. The right to privacy. *Harv Law Rev.* 1890;4(5):193-220. (Webpage accessed 27 April 2025)

2 *Gesetz zum Schutz personenbezogener Daten (Hessisches Datenschutzgesetz – HDSG). Gesetz- und Verordnungsblatt für das Land Hessen I.* 1970;S.61. (Webpage accessed 15 October 2025)

3 United States. *Privacy Act of 1974*, Pub. L. No. 93-579, § 552a, 88 Stat. 1896 (1974) (codified as amended at 5 U.S.C. § 552a). (PDF accessed 15 October 2025)

4 Organisation for Economic Co-operation and Development (OECD). *Guidelines on the protection of privacy and transborder flows of personal data*. OECD Doc. C(80)58/FINAL. Paris: Organisation for Economic Co-operation and Development; 1980;Sep 23. (Webpage accessed 15 October 2025)

5 Jain A, Salas M, Aimer O, Adenwala Z. Safeguarding patients in the AI era: ethics at the forefront of pharmacovigilance. *Drug Saf*. 2024;48:119-127. https://doi.org/10.1007/s40264-024-01483-9 (Journal abstract)

6 World Health Organization (WHO). *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization; 2021. (Webpage accessed 25 March 2025)

7 U.S. Department of Health & Human Services (HHS). *The Belmont Report: ethical principles and guidelines for the protection of human subjects of research*. Washington (DC): U.S. Department of Health & Human Services; 1979. (Webpage accessed 21 March 2025)

8 Council for International Organizations of Medical Sciences (CIOMS). *Real-world data and real-world evidence in regulatory decision making: CIOMS Working Group report*. Geneva: Council for International Organizations of Medical Sciences; 2024. https://doi.org/10.56759/kfxh6213 (Full text)

9 United States, Department of Health & Human Services (HHS). *Health information privacy: HIPAA for professionals*. [Internet]. Washington (DC): U.S. Department of Health & Human Services; 2024. (Webpage accessed 21 March 2025)

10 United States, Department of Justice, National Security Division. *Data Security Program* [Final rule implementing Executive Order 14117]. 28 CFR Part 202. Washington (DC): U.S. Department of Justice; 2025;Apr 8. (Webpage accessed 22 September 2025)

11 GDPR.eu. *What is GDPR, the EU's new data protection law?* [Internet]. Brussels: GDPR.eu; 2024. (Webpage accessed 21 March 2025)

12 Brazil. Law No. 13,709 of 14 August 2018. Provides for the protection of personal data and amends Law No. 12,965/2014 (the Brazilian Internet Civil Framework) [Internet]. Brasília: Government of Brazil; 2018. (Webpage accessed 15 October 2025)

13 Brazil. Law No. 13,853 of 8 July 2019. Amends Law No. 13,709 of 14 August 2018 to provide for the creation of the National Data Protection Authority (ANPD) and makes other provisions [Internet]. Official Gazette of the Federal Government. Brasília: Government of Brazil; 2019;Jul 9. (Webpage accessed 15 October 2025)

14 Brazil. ANVISA Ordinance No. 1,184 of 17 October 2023. Establishes the Personal Data Protection Policy of the Brazilian Health Regulatory Agency (ANVISA) [Internet]. Brasília: Agência Nacional de Vigilância Sanitária; 2023;Oct17. (Webpage accessed 15 October 2025)

15 National People's Congress (China). *Personal Information Protection Law of the People's Republic of China*. Beijing: National People's Congress; 2021. (Webpage accessed 15 October 2025):

16 Cyberspace Administration of China (CAC). *Official website of the Cyberspace Administration of China*. [Internet]. Beijing: Cyberspace Administration of China; 2025. (Webpage accessed 15 October 2025)

17 National Health Commission of the People's Republic of China (NHC). *Official website of the National Health Commission*. [Internet]. Beijing: National Health Commission; 2025. (Webpage accessed 15 October 2025)

18 National People's Congress (China). Data Security Law of the People's Republic of China. Beijing: National People's Congress; 2021. (Webpage accessed 15 October 2025)

19 National People's Congress (China). *Cybersecurity Law of the People's Republic of China*. [Internet]. Beijing: National People's Congress; 2016;Nov7. (Webpage accessed 15 October 2025)

20 State Council of the People's Republic of China. *Regulations on network data security management*. [Internet]. Beijing: State Council of the People's Republic of China; 2024;Sep30. (Webpage accessed 15 October 2025)

21 State Council of the People's Republic of China. *Regulation on the protection of non-public data*. [Internet]. Beijing: State Council of the People's Republic of China; 2024;May30. (Webpage accessed 15 October 2025)

22 National People's Congress (China). Law of the People's Republic of China on Basic Medical and Health Care and the Promotion of Health. [Internet]. Beijing: National People's Congress; 2019;Dec28. (Webpage accessed 15 October 2025)

23 State Council of the People's Republic of China. *Regulations on the Administration of Human Genetic Resources (2024 Revision).* [Internet]. Beijing: State Council of the People's Republic of China; 2024. (Webpage accessed 15 october 2025)

24 National People's Congress (China). Standard Contract Measures for the Export of Personal Information. [Internet]. Beijing: National People's Congress; 2023;Feb22. (Webpage accessed 15 october 2025)

25 Japan. Act on the Protection of Personal Information [Internet]. Amended June 12 2020; Personal Information Protection Commission; 2020. (Webpage accessed 15 October 2025)

26 Japan. Act on Anonymized Medical Data That Are Meant to Contribute to Research and Development in the Medical Field [Internet]. Law No. 85 of 2017 (enacted May 9 2017; effective May 30 2018). Ministry of Health, Labour and Welfare & other competent ministries; 2017. (Webpage accessed 15 October 2025)

27 Japan. Ministerial Ordinance on Good Post-marketing Study Practice for Drugs (Ordinance No. 171 of 2004) [Internet]. Tokyo: Ministry of Health, Labour and Welfare; 2004. (Full text accessed 15 October 2025)

28 Ethical Guidelines for Medical and Biological Research Involving Human Subjects (Japan) Japan. Ethical Guidelines for Medical and Biological Research Involving Human Subjects [Internet]. Public Notice of MEXT / MHLW / METI No.1 of 2021. Tokyo: Ministry of Health, Labour and Welfare; 2021. (Journal full text accessed 15 October 2025)

29 Sweeney L. *Simple demographics often identify people uniquely.* Carnegie Mellon University, Data Privacy Working Paper No. 671. Pittsburgh (PA): Carnegie Mellon University; 2000;Jan;1-34. (Full text accessed 21 March 2025)

30 El Emam K, Dankar FK, Vaillancourt R, Roffey T, Lysyk M. Evaluating the risk of re-identification of patients from hospital prescription records. *Can J Hosp Pharm.* 2009;Jul;62(4):307-313. https://doi:10.4212/cjhp.v62i4.812 (Journal full text)

31 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. https://doi.org/10.1186/s13063-020-4120-y (Journal full text)

32 European Medicines Agency (EMA), 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.* Amsterdam: European Medicines Agency; 2018. (Webpage accessed 21 March 2025)

33 Starks T. *U.S. charges Chinese military hackers with massive Equifax breach.* [Internet]. Atlanta (GA): CNN; 2020. (Article accessed 21 March 2025)

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

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

36 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)

37 Janmey V, Elkin PL. Re-identification risk in HIPAA de-identified datasets: the MVA attack. *AMIA Annu Symp Proc.* 2018;Dec 5;2018:1329-1337. (Journal full text accessed 15 October 2025)

38 Ayyamperumal SG, Ge L. *Current state of LLM risks and AI guardrails.* [preprint]. *arXiv.* 2024;Jun16. https://doi.org/10.48550/arXiv.2406.12934 (Journal full text)

39 Asthana S, Zhang B, Mahindru R, DeLuca C, Gentile AL, Gopisetty S. *Deploying privacy guardrails for LLMs: a comparative analysis of real-world applications.* [preprint]. *arXiv.* 2025;Jan 21. https://doi.org/10.48550/arXiv.2501.12456 (Journal full text accessed 22 September 2025)

40 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-Confid Comput.* 2025;Feb 28;100300 https://doi.org/10.1016/j.hcc.2025.100300. (Journal full text)

41 OpenAI. *Usage policies.* [Internet]. San Francisco (CA): OpenAI; 2025;Jan 29. (Webpage accessed 22 September 2025)

42 European Parliament, Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (Article 51 § 1, a). Off J Eur Union.* 2024;Jun 13;L 1689:1-145. (Webpage accessed 22 September 2025)

43 European Parliament, Council of the European Union. *Annex XIII: criteria for the designation of general-purpose AI models with systemic risk (in Regulation (EU) 2024/1689, Artificial Intelligence Act). Off J Eur Union.* 2024;Jun 13;L 1689:136-138. (Webpage accessed 22 September 2025)

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, 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 when possible for areas where 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) that are relevant to populations anticipated to have exposure to the specific medicinal product being evaluated.
- Screening and identifying explicit or potential bias using appropriate statistical methodologies when possible is key to implementing mitigation measures to reduce risk, determining AI applicability and limitations, and establishing expected performance acceptance criteria.
- Scrutinise 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.
- Fairness and equity in ICSRs remain limited due to the known limitations of spontaneous reporting systems, with some countries reporting significantly more than others and providing more contextual data for analysis, such as RWD. Consequently, our understanding of routine usage is often limited among underserved populations.

8.1. Introduction

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 to various degrees depending on the type of technology, data source and application of AI.

Not all fairness and equity concepts, considerations or negative consequences associated with the use of AI will be uniquely specific to PV. Fairness and equity considerations are challenging and can be influenced by cultural differences, historical inequalities, perceptions, socio-economic differences etc., and may appear subjective. That does not negate the need to address these considerations. To limit bias, intentional actions are required throughout the AI lifecycle, from concept and design through implementation, and while in production, to reduce discriminatory risk.[1]

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 systems 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 AI, including 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.[2]

8.2. 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.[3] When utilising AI solutions in PV, it is essential to implement proactive strategies to mitigate potential harm caused by AI systems that are being developed for high-risk PV activities.

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 means 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 organisation applying AI to PV activities is responsible for ensuring that the solution meets the defined business requirements, supports patient safety activities, and does not introduce bias that may inadvertently place patients at risk, in 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.

8.3. 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 recognised as biased. Conversely, the output could be accurate, fair, and equitable; however, results may be rejected by the human with a bias towards the AI system as being of poor performance.

8.3.1. 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 recognised. 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 organised data, e.g. regions with less systemic PV reporting systems. Otherwise, scenarios may be biased towards groups represented by the training data, and since the training data does not represent all groups, e.g. all ethnicities, AI systems with inadequate training data could result in poor system performance and discrimination against the under-represented groups.

Inadequate data - whether as a result of data not being available or not organised 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 recognised, and this may negatively impact patients' health outcomes. For example, an algorithm developed to detect Acute Kidney Injury (AKI) using clinical data predominately representing older non-black men may not be reliable when used to detect AKI in younger female patients and in ethnicities not represented in the data.[4] 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.

It is acknowledged that it may not always be possible to find datasets for development and testing that are fully representative of the intended population. In some cases, the gap between apparently similar datasets may be too wide to bridge. In others, appropriate care can be taken to re-purpose a dataset. For this, the developer should provide appropriate documentation and demonstrate the appropriateness of models trained on imperfect data.

Historically, there have been examples of bias influencing PV activities because of data limitations such as known under-reporting or stimulated reporting of AEs, with inadequate data or imbalance of data that could result in misleading or inaccurate results. Rofecoxib (Vioxx), a Cox-2 inhibitor prescribed for osteoarthritis pain, provides an important example of stimulated reporting where a significant number of AE reports were received once withdrawn.[5,6,7] Impact of safety alerts on measures of disproportionality in spontaneous reporting databases exemplifies the notoriety bias. Drug safety. Litigation, such as class action lawsuits that are pursuing product liability claims, can result in stimulating high volume of reported AEs during the process of legal firms identifying potential plaintiffs.[10] Solicited reports 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.[8] Reporting practices, data availability, and variability should be considered to understand limitations and limit potential bias. These data biases, if introduced into an AI solution, will potentially magnify the negative impact and remain undetected with difficult identification of underlying bias.

8.3.2. Underserved groups

Under-representation can directly result in underserved population segment(s) and potentially not recognise 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, resulting in a negative impact or outcome. This can occur if data are exclusive to a specific group, if data are exclusionary, or if nuances of a subgroup are not understood. For example, a case prioritisation algorithm may underperform in reports from certain countries in Asia if reports from Asia were under-represented in a training data set and differed in important ways from other countries represented. Inclusion of appropriate subject matter experts (SMEs), who can support identification and assessment of limitations of representation to ensure that these groups can be accounted for when developing and applying AI solutions, is fundamentally important.

During clinical trials, such potential harm may be missed by the Investigator 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 AEs 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 minimisation activities based on data with limited representation of served

population) could result in minimisation measures not properly addressing safety of patients in the population. If unable to mitigate insufficient data in AI solutions, it may require non-AI PV safety measures (e.g. robust monitoring measures for special populations).

The detailed identification of groups that could be disfavoured, identification of low-volume events that are disproportionate to the data set, along with deploying comprehensive strategies to address insufficient data, when possible, can reduce potential bias and discrimination against underserved populations.

8.3.3. 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 which 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 utilisation, not need.[9] 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.

Human-defined parameters and how a model processes data could introduce bias or produce inaccurate results. If individuals select or design 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 younger adult and may select data that does not account for paediatric or geriatric populations).

The development strategy should have a conscious systematic approach to limit bias and achieve complete and accurate data representation accounting for representative groups. Strategies could include review and adjustments of AI solutions as necessary, including the avoidance of historical biases. Documenting how distinct groups are represented in the training and test data may provide insight to limitations, bias, and potentially 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.[10]

8.4. 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 are 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.

To assess the impact of potential bias, methods of analysing the fairness can be utilised. Current techniques for assessing fairness in AI systems are focused on normative (value-based), procedural (process focused) and algorithmic (technical) approaches, as described by Li & Chignell.[11] Normative approaches focus on societal norms, shared values and principles to achieve an ideal standard for development of AI systems. Procedural approaches allow for self-assessment using a defined framework such as checklists or decision trees. Algorithmic/statistical approaches rely on a technical solution to support fair algorithmic decision making. Formal impact assessments of AI systems should include screening for potential bias that could negatively impact fairness and equity. This screening allows for identification of potential risk areas and responsive mitigation measures to minimise negative outcomes. When possible, applying appropriate statistical methods to test for bias as part of a detailed assessment supports development of mitigation strategies related to AI. An example of applying statistical tests of fairness is the use of t-tests assessing bias against minority/majority, single case bias, and unsuccessful favouritism toward minority/majority.[11]

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, determining 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 mitigation strategies 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.

8.5. Key mitigation strategies

- Evaluate each AI solution for fairness and equity, outlining the assessment method, results, and any measures taken to mitigate.
- Consider common biases across each phase of AI model development and key mitigation strategies to address biases at the different phases of the AI model lifecycle.[4] Ensure that training and test data sets are complete and representative of all relevant groups.
- Perform sensitivity analysis when possible, evaluating AI model results for equity by changing subgroups/populations to confirm expected results and highlight underserved populations. This is especially important when an AI solution has a lower level of explainability. Examples:
  - Modify sex and gender input and evaluate impact to the output;
  - Refer to an example noted in the Section Underserved groups regarding a case prioritisation algorithm underperforming in certain countries such as Asia.
- 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.

Chapter 8 – References

1 Hasanzadeh F, Josephson CB, Waters G, et al. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *npj Digit Med*. 2025;8:154. https://doi.org/10.1038/s41746-025-01503-7 (Journal full text)

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

3 Hamid AAA, Rahim R, Teo SP. Pharmacovigilance and its importance for primary health care professionals. *Korean J Fam Med*. 2022;Sep;43(5):290-295 https://doi.org/10.4082/kjfm.21.0193 (Journal full text)

4 Nazer LH, Zatarah R, Waldrip S, Ke JXC, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health.* 2023;2(6):e0000278. https://doi.org/10.1371/journal.pdig.0000278 (Journal full text)

5 Pariente A, Gregoire F, Fourrier-Reglat A, Haramburu F, Moore N. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. Drug Saf. 2007;30(10):891-898. https://doi:10.2165/00002018-200730100-00007. (Journal full text)

6 Catalogue of Bias. Richards GC, Onakpoya IJ. Reporting biases. In: Catalog Of Bias 2019: www.catalogueofbiases.org/reportingbiases (Webpage accessed 22 September 2025)

7 Haguinet F, Bate A, Stegmann J-U. The futility of adverse drug event reporting systems for monitoring known safety issues: a case study of myocardial infarction with rofecoxib and other drugs. *Pharmacoepidemiol Drug Saf*. 2024;33(1):e5719. https://doi.org/10.1002/pds.5719. (Journal full text)

8 Jokinen J, Bertin D, Donzanti B, et al. Industry assessment of the contribution of patient support programs, market research programs, and social media to patient safety. *Ther Innov Regul Sci.* 2019;53:736-745. https://doi.org/10.1177/2168479019877384 (Journal full text)

9 Obermeyer Z, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366:447-453. https//doi:10.1126/science.aax2342 (Journal full text)

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

11 Li J, Chignell M. FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI Ethics.* 2022;2:837-850 https://doi.org/10.1007/s43681-022-00145-9 (Journal abstract)

CHAPTER 9.
GOVERNANCE & ACCOUNTABILITY

Governance - Principle

Governance refers to the human management system 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 to fulfil PV obligations. 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, maintaining and managing AI systems. This ensures operations remain within expected parameters throughout the AI lifecycle while addressing any unforeseen consequences.

Key messages

- Governance requires the established PV QMS system to include a comprehensive approach across all lifecycle stages of an AI system as well as the processes it impacts and should therefore be established as early as possible.
- Accountability rests with the organisation that owns and operates the AI solution for PV and requires clearly defined roles and responsibilities for stakeholders involved in it; 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 solution in PV can serve as a structured guide to help relevant parties to document key elements throughout the lifecycle of the AI system.
- Governance and accountability should be independent of the business' utilisation and value proposition of the AI system to facilitate unbiased decision making.

9.1. 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 systems 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 solutions.

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. These include the refinement of the proposed governance framework grid for practical use.

9.2. Governance framework

A governance framework grid (referred to as grid) for AI solutions in PV (see Table 9) 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 system developed for PV use.

In addition to serving as a structured guide for planning and overseeing AI systems, 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 systems enter routine use phase. If a risk emerges that warrants significant modification to the AI system, 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 organisation 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 system to a PV organisation, 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 solution. These principles should guide the development or selection of a vendor system, deployment, ongoing management, and decommission. 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 system. It should also be noted that during the lifecycle of an AI solution, it may be necessary to go back to a previous phase to address certain needs and discoveries. 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.

The below lifecycle phase descriptions accompany Table 9 on the Governance framework grid on pages 92 and 93.

Collection of specifications, requirements: This is the initial phase where the stakeholders are identified and engaged, and the project's objectives, value proposition, scope and features are defined. The multidisciplinary team of PV professionals, data scientists, AI/ML engineers, software engineers, IT specialists, and other domain experts (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 system 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 system, 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 including extensive AI specific tailored review of outputs 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 system and for adapting the processes in which the system 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, PV processes impacted by deployment, etc.

Post-deployment & post change "hyper-care": Following deployment, this phase is critical for the immediate monitoring of the AI system'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 differs from traditional software hypercare; for AI systems, 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 systems 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 system being suspended and sent back to pre-deployment for enhancements.

Table 9: Governance framework grid

Source: CIOMS Working Group XIV

	Collection of specifications, requirements	Development & change management	Pre-deployment sign-off	Post-deployment hyper-care	Routine use
Risk-based approach	Risk assessment (theoretical) ▶ AI system ▶ 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
Human oversight	Multidisciplinary expertise Consider human oversight strategy	Define oversight strategy Change management, including staff training plan	Fine-tune human oversight strategy Staff training roll-out	Implement oversight strategy Intensive or targeted intervention	Routine oversight activities Adjust and fine-tune oversight strategy as needed
Validity & Robustness	Specification of use case & deployment domain Specification of reference standard(s) Specification of benchmarks Requirements on reproducibility	Model training & validation Development or acquisition of reference standards	Performance evaluation Benchmark comparisons Formal execution of system testing	Performance monitoring	Continuous integration and deployment Periodic performance monitoring
Transparency	[From organisation to developer] Model requirements ▶ Intended use ▶ Human-computer interaction ▶ Explainability ▶ Expected outputs ▶ Performance evaluation requirements ▶ Scope ▶ Reference standard	[From developer to organisation] Model ▶ Architecture ▶ Parameters ▶ Acceptable inputs ▶ Expected outputs ▶ Standard AI components ▶ Model training & validation ▶ Known limitations Explainability in support of model development, debugging, and documentation	[From developer to organisation] 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 organisation and from organisation to end user] Performance evaluation ▶ Deviations Explainability in support of assessing Validity & Robustness and Fairness & Equity	[From organisation to end user (and regulatory authorities)] Disclosing use of AI Explainability in support of building trust with end users

	Collection of specifications, requirements	Development & change management	Pre-deployment sign-off	Post-deployment hyper-care	Routine use
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 minimisation) ▸ Data protection impact assessment	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 hyper-care	Ongoing processes to identify and rectify data privacy issues in routine use
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	Greater attention to model shift, inadequate training data	Routine Monitoring -Poor performance related to model shift, inadequate training data (underrepresented populations, special populations)
Governance & Accountability	Consideration on how AI system 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

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. Efforts should be made to ensure AI system and training data description is transparent, and limitations are highlighted to reduce inappropriate application or incorrect conclusions. 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 system'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 system. When relevant, it may involve the development of clear annotation guidelines for human domain experts to ensure solid method development and performance evaluation. When evaluating AI system performance, special considerations should be given for low-prevalence settings (see also the discussion of performance evaluation in the Chapter on Validity & Robustness). The next step is to propose appropriate mitigation strategies such as defining "human-in-the-loop" within an AI system 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 at regular intervals in the Routine phase based on real-life observations. This sequential approach highlights the need for evolving risk management as the AI system 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 hyper-care, 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 system version control from the rest of the software versioning.

9.2.1. 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 and segregation of duties ensure a broad and balanced review of the AI system. The governance body oversees the development, deployment, performance, 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 systems that changing inputs, rules or other unforeseen issues may disrupt the system at varying degrees, some significantly. The appropriate frequency and scope of reassessment of a deployed AI system should be assessed. There should be measures ready to intervene or even disable the AI system if necessary. Once the governing body defines that the AI system has reached 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 or suspension to the AI system, 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.

9.3. Traceability and version control

Traceability and version control are crucial aspects of managing AI systems, 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 system and are often required for audit purposes (see also the discussion of AI systems 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 systems, 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 system should comprise its entire lifecycle, and may cover the justification, initial scoping and conception, development, deployment, validation, post-deployment, and decommissioning. 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 system 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 systems undergo continual experimentation and iterative improvement. Transparency between the development team and the PV organisation is crucial to ensure efficiency and that the system 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 system 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 AI components would be in a document such as the PV System Master File (PSMF), in the EU.

The continual improvement and adaptation of AI systems post-deployment should also be documented. It may be triggered by human domain experts or built into the deployment of the AI system 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.[1]

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 system lifecycle.

9.3.1. Roles and responsibilities in artificial intelligence-enhanced pharmacovigilance systems

Organisations 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 organisations 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, 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 systems and PV experts is crucial. This collaboration can help to ensure that AI systems meet PV requirements and governance principles. Regular audits and qualification of vendors and AI systems, ongoing performance monitoring, business continuity planning 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 systems into the PV system must include appropriate regulatory documentation (see Chapter on Transparency), such as in the PSMF.

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 system KPIs.

Chapter 9 – Reference

1 U.S. Food and Drug Administration (FDA), Health Canada, Medicines and Healthcare products Regulatory Agency (MHRA). *Good machine learning practice for medical device development: guiding principles*. Silver Spring (MD): U.S. Food and Drug Administration; 2021. (Full text accessed 21 March 2025)

CHAPTER 10.
FUTURE CONSIDERATIONS FOR DEVELOPMENT AND DEPLOYMENT OF ARTIFICIAL INTELLIGENCE IN PHARMACOVIGILANCE

10.1. 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 (see Appendix 3) detail various AI systems under evaluation, and at stages of deployment, and provide an assessment of their effectiveness within the discipline. To try to predict into the future, it is essential to recognise that the trajectory of AI is dynamic and highly unpredictable. Indeed, the only truly predictable elements are that AI is expected to be ubiquitously deployed in medical sciences with the potential to revolutionise 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 HCPs 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 revolutionise 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/or 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.[1,2,3,4,5]

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

Advanced AI systems are poised to take PV beyond current boundaries, for example into tasks supported by automated or augmented decision making, AI agents and quantum computing to improve AI training, modelling and simulation, and optimising personalised medicine.[6,7,8,9] 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.[10,11] While current limitations of AI models are acknowledged earlier in this report, a more explicit recognition of the ongoing evolution of AI infrastructure is recommended. It is envisioned that evolution of AI systems will improve and address some, if not all, of the limitations of the current models.

In the future, an expert AI system, designed specifically for PV, may emulate the judgement and decision-making processes of seasoned professionals or organisations 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.

In the future, 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.).[12,13,14,15]

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.

10.3. 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 systems 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 systems 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 utilisation of AI is likely to significantly hamper patient safety. The principles of Human Factors and Ergonomics (HFE) can assist in simplifying AI design and consequently optimise human performance ensuring better understanding of AI outcome. HFE is a scientific discipline that focuses on understanding interactions between humans and other elements of a system to optimise human well-being and overall system performance and uses principles, data, and methods to design and improve systems, products, and environments.[16]

The oversight and risk mitigation of such advanced AI systems demands 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.

As covered in Chapter 3 on Risk-based approach, risk mitigation needs to consider identifying rare, unexpected anomalies (Black Swan incidents). 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.

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 HITL 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.[17,18,19]

This evolving landscape will require PV professionals to develop new skillsets, AI system benchmarking tools and metrics, and undergo specialised training to effectively oversee AI-driven systems. This includes, but is not necessarily limited to, AI-aided testing and benchmarking, KPIs, workflow improvements, and drift and bias monitoring. The focus must extend beyond traditional oversight methods to include competencies in understanding, interpreting, and guiding AI behaviours. In addition, there must be implementation of dynamic and continuous training of PV professionals and technicians overseeing AI systems in PV to ensure appropriate monitoring. However, HITL oversight may be insufficient for certain highly complex and/or very high-risk applications. In these instances, new tools may need to be developed in conjunction with training to maintain adequate human oversight. By cultivating these skills and adjunct oversight tools, 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 robustness and considers multidisciplinary 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 processing vast amounts of data in real time or near real time, there is a need for scalable 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 and creation of cross functional and multidisciplined PV teams to meet the demands of validating AI-enabled systems. As discussed in Chapter 5, best practices for critical appraisal of AI in generative applications are still evolving and will likely become better understood and more consistently utilised. AI use with some advanced technologies would need the creation of new standards and validation methods and consideration of how it is deployed for optimising

the outputs and real-time / near-real-time safety data generated, e.g. neurotechnology such as implantable chips, 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 scrutinised, 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 is vital for understanding failures and enhancing system improvements rather than serving as a constant requirement.

This shift may significantly influence PV decision making, emphasising timely interventions and near-real-time root cause analysis. Looking ahead, organisations 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. The trajectory of AI in PV is poised for further rapid growth. The incorporation of big data analytics, federated learning, and blockchain will enhance data security, interoperability, and global collaboration. For example AI-powered chatbots and virtual assistants will facilitate real-time ADR reporting by engaging with patients and HCPs seamlessly.[20,21] As noted in Chapter 7, leaks of personal data have been increasing in frequency, with some at enormous scale.[22] 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 without the individual's approval carries risks for emotional well-being, stigmatisation, 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 commoditisation of data to maximise 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 anonymisation might incorporate blockchain or similar technology to make complete anonymisation 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. A further challenge for data privacy and deployment is heterogeneity of data privacy regulations and data sharing across regions.

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 realise AI's potential while safeguarding public confidence.

Fairness & Equity

Chapter 8 on fairness and equity 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 example, of higher healthcare costs associated with more advanced monitoring, genetic profiling and/or personalised risk/remediation, or 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.

Governance & Accountability

Chapter 9 of 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 emphasises 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.

10.4. 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 cause a paradigm shift in this discipline, which is likely to be focused on rapid or real-time data collection, assessment and reporting. For example, providing us 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.

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.

Chapter 10 – References

1 Ventola CL. The nanomedicine revolution: part 1: emerging concepts. P T. 2012;Sep;37(9):512-525. (Journal full text accessed 28 April 2025)

2 Ventola CL. The nanomedicine revolution: part 2: current and future clinical applications. PT. 2012;Oct;37(10):582-591. (Journal full text accessed 28 April 2025)

3 Ventola CL. The nanomedicine revolution: part 3: regulatory and safety challenges. PT. 2012;Nov;37(11):631-639. https://doi:10.1007/s13300-012-0003-x (Journal full text)

4 Ventola CL. Big data and pharmacovigilance: data mining for adverse drug events and interactions. *PT.* 2018; Jun;43(6):340-351. (Journal full text)

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

6 Shneiderman B. *Human-centered artificial intelligence: reliable, safe & trustworthy.* [preprint]. *arXiv.* 2020;Feb23. https://doi.org/10.48550/arXiv.2002.04087 (Journal full text)

7 National Academies of Sciences, Engineering, and Medicine. *Opportunities and challenges for digital twins in biomedical research: proceedings of a workshop–in brief*. Washington (DC): The National Academies Press; 2023. https://doi.org/10.17226/26922. (Journal full text)

8 Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. *npj Digit Med.* 2024;7:77. https://doi.org/10.1038/s41746-024-01073-0 (Journal full text accessed 19 September 2025)

9 Zhang K, Zhou H-Y, Baptista-Hon DT, Gao Y, et al. Concepts and applications of digital twins in healthcare and medicine. *Patterns*. 2024;5(8):101028., https://doi.org/10.1016/j.patter.2024.101028. (Journal full text accessed 19 September 2025)

10 Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. *Commun ACM.* 2020;63(5):82–89. https://doi.org/10.1145/3376898 (Journal full text accessed 28 April 2025)

11 Hauben M. Artificial intelligence and data mining for the pharmacovigilance of drug–drug interactions. *Clin Ther.* 2023;45(2):117-133. doi:10.1016/j.clinthera.2023.01.002. (Journal full text accessed 28 April 2025)

12 Sardari S, Hheidari A, Ghodousi M, Rahi A, et al. Nanotechnology in tissue engineering: expanding possibilities with nanoparticles. Nanotechnology. 2024; 35 392002. DOI 10.1088/1361-6528/ad5cfb. (Journal full text accessed 24 October 2025)

13 Zhu X, Wang Z, Teng F. A review of regulated self-organizing approaches for tissue regeneration. *Prog Biophys Mol Biol*. 2021;167:63–78. https://doi.org/10.1016/j.pbiomolbio.2021.07.006. (Journal full text accessed 28 April 2025)

14 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;Aug2;9(1):17. https://doi.org/10.1186/s42234-023-00118-1. (Journal full text accessed 28 April 2025)

15 Abyzova E, Dogadina E, Rodriguez RD, Petrov I, et al. Beyond Tissue replacement: The Emerging role of smart implants in healthcare. Mater Today Bio. 2023 Aug 29;22:100784. doi: 10.1016/j.mtbio.2023.100784. (PubMed accessed 24 October 2025)

16 Choudhury A, Asan O. Human factors: bridging artificial intelligence and patient safety. *Proc Int Symp Hum Factors Ergon Health Care*. 2020;Oct5;9(1):211–215. https://doi.org/10.1177/2327857920091007 (Journal abstract accessed 28 April 2025)

17 Bonabeau E, Dorigo M, Theraulaz G. *Swarm intelligence: from natural to artificial systems*. New York: Oxford University Press; 1999; online edn 2020;Nov12.https://doi.org/10.1093/oso/9780195131581.001.0001 (Abstract accessed 28 April 2025)

18 Kurzweil, Ray (2005). The Singularity is Near: When Humans Transcend Biology. Viking Press.

19 Karakas F. *Imagine you are living in the age of singularity.* [Internet]. *Creative Adventures*. 2021;75 (Webpage accessed 28 April 2025)

20 Chen C, Feng X, Li Y, Lyu L, et al. Integration of large language models and federated learning. *Patterns*. 2024;5(12):101098., https://doi.org/10.1016/j.patter.2024.101098. (Journal full text accessed 19 September 2025)

21 Shama SN. *AI in pharmacovigilance: the dawn of a data-driven safety revolution.* [Internet]. Pharma Focus America; 2025. (Webpage accessed 19 September 2025)

22 Forbes. *Cybersecurity stats: facts and figures you should know.* [Internet]. New York: Forbes Media; 2024. (Webpage accessed 28 April 2025)

APPENDIX 1.
GLOSSARY

This glossary provides definitions specific to terms within the context of AI use in PV or existing definitions have been simplified for the purpose of this document e.g. technical definitions. 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 organisations and/or individuals deploying, operating and managing artificial intelligence systems to fulfil Pharmacovigilance obligations. 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, maintaining and managing artificial intelligence systems). This ensures operations remain within expected parameters throughout the artificial intelligence lifecycle while addressing 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.

Modified 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)

Agent in AI

Software program that interacts with its environment to collect data and utilize that data to perform specific tasks to meet predetermined goals. Agents can act independently or collaborate to achieve a common goal.

Modified from: Amazon Web Services (AWS). *What are AI agents?* [Internet]. Seattle (WA): Amazon Web Services; 2024. (Webpage accessed 15 October 2025)

Artificial intelligence literacy

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

Modified from: Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. *Comput Educ Artif Intell.* 2021;2:100041. https://doi.org/10.1016/j.caeai.2021.100041. (Journal full text)

Artificial intelligence system

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.

Modified from: Organisation for Economic Co-operation and Development (OECD). *Explanatory memorandum on the updated OECD definition of an AI system.* (*OECD Artificial Intelligence Papers*, No. 8.) Paris: OECD Publishing; 2024. (PDF accessed 15 October 2025) 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.

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. In: *Systems engineering for the digital age: practitioner perspectives*. Hoboken (NJ): Wiley; 2023;Oct 8. p. 375-391. https://doi/10.1002/9781394203314.ch17 (Chapter abstract)
- World Medical Association (WMA). *WMA statement on augmented intelligence in medical care.* Ferney-Voltaire (France): World Medical Association; 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 or trust 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. *Hum Factors*. 2010;Jun;52(3):381-410. https://doi.org/10.1177/0018720810376055 (Journal full text)
- Cummings ML. Automation bias in intelligent time-critical decision support systems. In: *Decision making in aviation*. Boca Raton (FL): Routledge; 2017;Jul5. p. 289-294. (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 (ISO/IEC DIS 22989). (…) 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.

Modified from: International Medical Device Regulators Forum (IMDRF). *Machine learning-enabled medical devices — a subset of artificial intelligence-enabled medical devices: key terms and definitions*. Geneva: International Medical Device Regulators Forum; 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.

Modified from: University of Saskatchewan. *Generative artificial intelligence: glossary of AI-related terms*. Saskatoon (SK): University of Saskatchewan; 2024. (Webpage accessed 4 April 2025)

Black-Box model

An AI 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 Saf.* 2022;May;45(5):419-427. https://doi.org/10.1007/s40264-022-01169-0 (Journal full text)
- Taleb NN. Black swans and the domains of statistics. *Am Stat.* 2007;Aug 1;61(3):198-200. doi:10.1198/000313007X219996. (Journal full text)

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 MAHs and competent authorities.

Modified from: Heads of Medicines Agencies (HMA), European Medicines Agency (EMA). *Guideline on good pharmacovigilance practices (GVP): Module I – pharmacovigilance systems and their quality systems.* London: European Medicines Agency; 2012.v (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.

Modified from: International Organization for Standardization (ISO). *What is change management: a quick guide.* Geneva: International Organization for Standardization; 2023. (Webpage accessed 27 October 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.* Amsterdam: European Medicines Agency; 2024. (Full text accessed 3 April 2025)

Cluster analysis

A machine learning method that partitions differing data elements into sets of data elements based on similarities to identify patterns that are not immediately evident when not combined.

Derived from: Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett.* 2010;31(8):651-666. https://doi.org/10.1016/j.patrec.2009.09.011 (Journal full text)

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.

Modified 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).* Geneva: International Council for Harmonisation; 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.

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

Cross-validation

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

Modified from: Berrar D. *Cross-validation*. Preprint submitted to *Encyclopedia of Bioinformatics and Computational Biology*, 2nd ed. Amsterdam: 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 anonymisation

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

Modified 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.

Modified from: U.S. Food and Drug Administration (FDA). *FDA digital health and artificial intelligence glossary – educational resource*. Silver Spring (MD): U.S. Food and Drug Administration; 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 emphasise the protection of sensitive and personal data (including health data).

Proposed by CIOMS Working Group XIV.

Decision tree

A model which categorizes data into various subsets to identify a potential structure, pattern and relationship among the data.

Modified from: Dikshit A, Pradhan B, Santosh M. Artificial neural networks in drought prediction in the 21st century: a scientometric analysis. *Appl Soft Comput.* 2022;114:108080. https://doi.org/10.1016/j.asoc.2021.108080 (Journal full text)

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. *Nat Med.* 2023;Aug;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8 (Journal full text)

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)

Dynamic (adaptive or continual learning) AI model

AI model that continuously learns or adapts on an ongoing basis based on exposure to new data or changing environments during the operational phase of the AI systems lifecycle.

Modified 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)

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 is the avoidance or mitigation of bias to provide comparable results considering differences for a diverse group or population. Equity is the recognition of differences in a group or population and accounting for those differences to provide a fair result. This 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

A data point incorrectly identified as not belonging to a class of interest when it does belong to a class of interest.

Proposed by CIOMS Working Group XIV.

False positive

A data point incorrectly identified as belonging to a class of interest when it does not belong to a class in interest.

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.

Generative artificial intelligence application

A computerised application using artificial intelligence methods trained on data sets that can be used to generate new content, such as text, images, video or conduct discriminative tasks (e.g. classification) based on prompts provided by the user.

Proposed by CIOMS Working Group XIV.

Generative Large Language Models

Probabilistic models trained on a large number of parameters that enable the processing of natural language through algorithms specifically designed to generate text.

Modified from: Chiarello F, Giordano V, Spada I, Barandoni S, Fantoni G. Future applications of generative large language models: a data-driven case study on ChatGPT. *Technovation.* 2024;133:103002. https://doi.org/10.1016/j.technovation.2024.103002 (Journal full text)

Governance (for AI)

Governance refers to the human management system 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.

Proposed by CIOMS Working Group XIV.

Hallucination

In generative AI, hallucinations are generated content that is presented as authoritative but in actuality the information is incorrect or misleading.

Proposed by CIOMS Working Group XIV

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.

Modified from: European Commission. *Ethics guidelines for trustworthy AI.* Brussels: European Commission; 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.* Brussels: European Commission; 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.

Modified from: European Commission. *Ethics guidelines for trustworthy AI.* Brussels: European Commission; 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

The complete information provided by a reporter at a certain point in time to describe an event or incident of interest. The report can include information about a case involving one subject or group of subjects.

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)

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.

Modified from: Hauben M, Rafi M. Knowledge graphs in pharmacovigilance: a step-by-step guide. *Clin Ther.* 2024;46(7):538-543. https://doi.org/10.1016/j.clinthera.2024.03.006 (Journal full text)

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.

Derived 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)

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 to provide an assigned output.

Modified 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 (Concept Drift)

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

Modified from: Wang S, Schlobach S, Klein M. Concept drift and how to identify it. *J Web Semant.* 2011;9(3):247–265. doi:10.1016/j.websem.2011.05.003 (Journal full text)

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. *Nat Med.* 2023;Aug;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8 (Journal full text)

Negative control

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 'perceptrons' (nodes), usually arranged in layers, communicating with one another and performing transformations upon input data.

Adopted from: Thirunavukkarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. 2023;Aug;29(8):1930-1940. (Full text accessed 13 November 2025)

Non-deterministic AI

An AI system in which the same input is not guaranteed to produce the same output, due largely to the inherent incorporation of randomness, probabilistic decision-making, or underlying stochastic algorithms in its design.

Proposed by CIOMS Working Group XIV.

Open and Closed Large Language Models

Closed models do not release the model weights to the public and access to these weights is restricted under proprietary licenses.

Open models provide access to model weights and are governed by non-proprietary license enabling adaptation and ability to further investigation the model.

Modified from: Xu J, Ding Y, Bu Y. Position: open and closed large language models in healthcare [preprint]. *arXiv.* 2025;Jan17. doi:10.48550/arXiv.2501.09906. (Journal full text)

Overfitting

Learning details from training data that reflect noise and will not generalize to new data.

Modified 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. *Nat Med.* 2023 Aug;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8 (Journal full text)

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.

Modified from Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*. 2020;Nov20;415:295-316. https://doi.org/10.1016/j.neucom.2020.07.061 (Journal full text accessed 15 October 2025)

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.

Modified 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). Off J Eur Union.* 2016;L 119:1–88. (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: Council for International Organizations of Medical Sciences (CIOMS). *CIOMS cumulative glossary, with a focus on pharmacovigilance*. Version 2.1. Geneva: Council for International Organizations of Medical Sciences; 2024. https://doi.org/10.56759/ocef1297 (Full text)

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*. London: European Medicines Agency; 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).

Modified from: Hicks SA, Strümke I, Thambawita V, Hammou M, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;Apr 8;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8 (Journal full text)

Positive control

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 model

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.

Modified from: De Hond AA, Leeuwenberg AM, Hooft L, Kant IM,et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ digital medicine*. 2022;Jan10;5(1):2. https://doi.org/10.1038/s41746-021-00549-7 (Journal full text)

Quality management system

Part of the pharmacovigilance system utilizing a framework of polices, processes and resources to maintain and improve safety and efficacy of any product or system.

Derived from: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *Pharmaceutical quality system Q10*. Geneva: ICH; 2008. (PDF accessed 15 October 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).

Modified from: Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;Apr8;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8 (Journal full text)

Red team

A group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture. The Red Team's objective is to improve enterprise cybersecurity by demonstrating the impacts of successful attacks and by demonstrating what works for the defenders (i.e., the Blue Team) in an operational environment. Also known as Cyber Red Team.

Adopted from: National Institute of Standards and Technology (NIST). *Glossary* [Internet]. Gaithersburg (MD): NIST Computer Security Resource Center; 2025. (Webpage accessed 23 October 2025)

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.

Derived 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;May 7. Chapter 3, Understanding reproducibility and replicability. (Chapter full text accessed 27 October 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 in the execution of core PV tasks. 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 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, which should generally be commensurate with the identified risk.

Proposed by CIOMS Working Group XIV.

Robustness

A system reliably achieves its intended objectives while accounting for variations in data.

Proposed by CIOMS Working Group XIV.

Secondary Use of Data

Use of existing data for a different purpose than the one for which they were originally collected. In the setting of AI this could include data used for the purposes of training or validating a model.

Modified 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)

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.

Derived from: Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. *J Biomed Inform*.2009;Apr1;42(2):390-405. https://doi.org/10.1016/j.jbi.2009.02.002 (Journal full text)

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.

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.

Modified from: Council for International Organizations of Medical Sciences (CIOMS). *Practical aspects of signal detection in pharmacovigilance*. Geneva: Council for International Organizations of Medical Sciences; 2010. (Full text accessed 15 October 2025)

Static AI model

AI model that remains unchanged once deployed.

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.* Geneva: International Medical Device Regulators Forum; 2021. (Full text accessed 3 April 2025)

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.* Geneva: International Medical Device Regulators Forum; 2021. (Full text accessed 3 April 2025)

Traceability (AI)

The ability to track and document the data, processes, classifications used to create an artificial intelligence model and derived output.

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.* Geneva: International Medical Device Regulators Forum; 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 organisations 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, perceived benefits 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.* Geneva: International Medical Device Regulators Forum; 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 assurance process.

Proposed by CIOMS Working Group XIV.

Validation dataset

Data used to tune hyperparameters or to validate some algorithmic choices (rule design, etc.).

Derived from: International Organization for Standardization (ISO). *ISO/IEC DIS 22989. Information technology — artificial intelligence - artificial intelligence concepts and terminology.* Geneva: International Organization for Standardization; 2022. (Webpage accessed 4 April 2025)

Zero-shot learning

Artificial intelligence developed to complete tasks without exposure to any previous examples of the task.

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

APPENDIX 2. **COMPARISON TABLE OF GUIDING PRINCIPLES**

Table 10: Comparison of CIOMS Working Group XIV guiding principles for artificial intelligence across regional and country government institutions, and international organisations – Extracted description of principles

Source: CIOMS Working Group XIV

	Examples of regional - and country government institutions', and international organisations' principles									
Principle	EU[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
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 lifecycle, 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 and should thus include effective, transparent monitoring of human values and moral considerations.	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[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
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 with the AI system in an adversarial manner.	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 oversight of such tests and evaluation to ensure that they are conducted safely and effectively.	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[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
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 protects privacy.	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 lifecycle. 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[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
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 Medical Devices (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.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	EU[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
Account-ability	The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems.	People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.	Accountability means that organisations 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. This includes the proactive documentation of policies, processes, and measures implemented.	While developers should be responsible for the proper design of algorithms used in the AI-MD, organisations using AI-MD to deliver care will be responsible for the decision to implement the AI-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 medical device, 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.	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 appeal or contest its impacts on you."	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 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.	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[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
Societal well-being	…the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's lifecycle. Ubiquitous exposure to social AI systems in all 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.	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 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.	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[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
Environmental well-being	See Societal well-being.	See Societal well-being.						AI systems should be designed to minimise 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 environment, ecosystems and climate.	See Societal well-being.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	EU[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
Fairness & Equity	…enable inclusion and diversity throughout the entire AI system's lifecycle. 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 mean 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 and gender).	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 lifecycle 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.

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	EU[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
Explainability	…the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes.	See Transparency		The decisions or recommendations from an AI-MD should endeavour to be explainable 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.	See Transparency	"Automated systems should provide explanations that are technically 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."		See Transparency	See Transparency

	Examples of regional - and country government institutions', and international organisations' principles								
Principle	EU[1,2]	Australia[3]	Canada[4]	Singapore[5]	UK[6]	US[7]	PAHO[8]	WHO[9]	OECD[10]
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 and acting upon risks.	"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.
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 lifecycle.			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.	

Appendix 2 – References

1 European Medicines Agency (EMA), Committee for Medicinal Products for Human Use (CHMP). *Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle*. Amsterdam: European Medicines Agency; 2024;Sep 9 (PDF accessed 27 April 2025)

2 European Commission, Directorate-General for Communications Networks, Content and Technology. *The assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*. Luxembourg: Publications Office of the European Union; 2020., https://data.europa.eu/doi/10.2759/002360 (PDF accessed 27 April 2025)

3 Australian Government, Department of Industry, Science and Resources. *Australia's artificial intelligence ethics principles*. Canberra: Australian Government; [year unknown] (Webpage accessed 27 April 2025)

4 Government of Canada. *The Artificial Intelligence and Data Act (AIDA) – companion document*. Ottawa: Government of Canada; 2022. (Webpage accessed 27 April 2025)

5 Ministry of Health, Singapore. *Artificial intelligence in healthcare guidelines*. Singapore: Ministry of Health; 2021. (PDF accessed 27 April 2025)

6 Department for Science, Innovation and Technology (DSIT). *Implementing the UK's AI regulatory principles: initial guidance for regulators*. London: UK Government; 2024 Feb. (PDF accessed 27 April 2025)

7 White House Office of Science and Technology Policy (OSTP). About this document: Blueprint for an AI Bill of Rights [Internet]. © 2025; (Webpage accessed 15 October 2025)

8 Pan American Health Organization (PAHO). *Artificial intelligence in public health: digital transformation toolkit*. Washington (DC): Pan American Health Organization; 2021. (PAHO/EIH/IS/21-011). (PDF accessed 27 April 2025)

9 World Health Organization (WHO). *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)

10 Organisation for Economic Co-operation and Development (OECD). *OECD AI principles overview*. © 2025. (Website accessed 27 April 2025)

APPENDIX 3.
USE CASES

The use cases presented in this appendix serve as practical illustrations for applying AI within PV. These cases provide insight into the potential applications of AI across various subdomains and highlight the methodologies, limitations, and associated performance with the integration of AI methodologies. Each use case exemplifies specific AI solutions across the current PV lifecycle offering readers a practical example of how AI can transform workflows, enhance efficiency, and ultimately contribute to improved patient safety.

Utilising these use cases appropriately requires an understanding of the principles that have been described in this guidance, and how some of these have guided the development and implementation of the use cases. It is important to note that the use cases pre-date this guidance and therefore not all principles may have been considered.

Readers are encouraged to analyse the context, objectives, and outcomes of each case study to derive meaningful insights for their organisational needs. By systematically evaluating the alignment of each AI solution with the governance framework outlined in the main report, stakeholders can identify best practices and potential pitfalls in AI integration, fostering a responsible approach to adopting AI technologies in PV.

Moreover, the use cases highlight the importance of adhering to the key guiding principles. Through consideration of these examples, organisations can gain valuable perspectives that drive innovation while safeguarding the integrity of their PV systems.

Use Case A: Large Language Models data extraction for case processing

Source:[1]
Area of PV: ICSR Processing

A1. Business rational and challenges

The GVP refer to the set of guidelines and standards established by EMA to ensure the safety and efficacy of pharmaceutical products throughout their lifecycle. These practices are essential for the systematic monitoring, assessment, and management of adverse drug reactions. Pharmaceutical companies must act on reports of potential adverse reactions to drugs to protect public health by ensuring that potential risks are identified and addressed promptly. 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, that identify and categorise key information using pre-defined annotations from unstructured sources, such as name(s) of reported AE, reporter qualification, countries, and specific terms, to analyse and extract relevant data, were unsuccessful in consistently improving case intake/processing operations under real-world circumstances. The use of LLMs in case intake/processing provides potential to advance processes without compromising quality. However, LLM-based tools should undergo periodic re-evaluation to monitor model drift effectively. Additionally, training and calibration processes must ensure that all identifiable data handling complies with GDPR and HIPAA regulations to safeguard data privacy and maintain compliance.

A2. Solution

A pharmaceutical company executed a proof-of-concept (PoC) study to assess the feasibility as well as the quantitative and qualitative business impact of utilising 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 three-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[i]-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. Redacted copies of 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 of 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.

[i] *JSON* = JavaScript Object Notation

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. Challenges and Lessons Learned

The learnings of this PoC converged into five key Points to Consider (PtC), which can be used as a springboard to support future research. Taking a practical industry perspective as well as relating the observations to scientific work in the field, the authors reflect on enabling innovative technologies and the experience shared, while preliminary, should aid others working in this space.

1. The Return of Investment (RoI) needs to be measurable in a business context:

   The RoI for implementing LLMs must be quantifiable, as they can yield significant efficiency gains (in this PoC up to 39%), translating into financial benefits and increased team productivity.

2. Early involvement of SMEs increases RoI:

   Engaging SMEs early is essential for optimising model performance, enhancing process understanding, and enables effective prompt engineering, ultimately leading to improved reliability and resource efficiency while addressing limitations of LLMs to increase RoI.

3. Regulatory uncertainty remains a significant hurdle:

   Regulatory uncertainty poses a significant hurdle for compliance with GxP standards in AI technologies, as the evolving regulatory landscape from major health authorities like the EMA and US FDA creates challenges that necessitate proactive risk management

and practical, solution-oriented approaches to ensure validation and accuracy in real-world applications.

4. System integration needs to be contextualised in the operational environment:

   System integration of LLMs must be contextualised within the operational environment, taking into account existing system limitations and user requirements to derive meaningful study results, while emphasising the importance of a prompting strategy and dedicated pre-processing and post-processing for effective embedding in established safety solutions.

5. Organisational readiness goes beyond technology:

   Organisational readiness for adopting new technology extends beyond mere technological capabilities, requiring human involvement, sufficient trust, and robust oversight, which can be fostered through early engagement with operational teams, mindset shifts, awareness, training, and process readiness to mitigate potential inhibiting factors and facilitate effective study conduct.

To effectively implement these key Points to Consider, a risk-based approach can serve as a strategic framework that aligns assessment of potential impact of inaccuracies on patient safety, development of effective mitigation strategies for false positives and ensure compliance; continuous monitoring and evaluation of the LLM's performance and optimisation of the integration of this technology into existing system.

A4. Compliance with the governance framework

Table 11: Use case A: Alignment with the governance framework (detail)

Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	To categorise risks associated with the implementation of Large Language Models for data extraction in case processing, it is essential to assess the potential impact of inaccuracies on patient safety and pharmacovigilance outcomes. Additionally, stakeholders should be engaged to develop effective mitigation strategies for false positives and ensure compliance with regulatory requirements, including GDPR and HIPAA. Continuous monitoring and evaluation of the LLM's performance, alongside robust training for users, will be critical to managing risks and optimising the integration of this technology into existing pharmacovigilance systems.	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

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Validity & 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 & 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. The case intake and processing team takes over the accountability and performs the 100% human QC process. The ultimate accountability remains with the MAH.	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

Use Case B: Case deduplication

Source:[2]
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 AE 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 AEs 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 AE reports. The regulatory agency collaborated with an academic partner to address this issue by developing a deduplication pipeline relying on modern technologies (mainly, NLP, network analysis, and cloud computing) and utilising 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% and 13% of known duplicate pairs, respectively, with (nearly) perfect precision in both cases (95% and 100%, respectively).[3] This algorithm was refined in subsequent steps to reach acceptable levels of performance that, in some cases and based on new evaluations using drug AE 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.[4] 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 and was approved for processing all historical reports and incoming live data in an ETL process (extract, transform, and load process). As of July 30, 2025, 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. Challenges and lessons learned

The deduplication pipeline was developed through a multi-year investigation, which involved investing various human and other resources to achieve the desired performance and facilitate the migration of this solution into the production environment. Still, a validated and transparent AI-based solution that outperforms existing ones and is freely available to the regulatory agency along with its underlying code, opens several opportunities to leverage the deduplication output and integrate the pipeline into existing systems. This maximises the benefits and eliminates the considerable costs associated with proprietary tools.

As this deduplication actively processes large AERS data daily and the output cannot be reviewed and confirmed by humans, it is essential to develop a strategy to ensure that the performance demonstrated in all evaluation rounds remains consistently high. A solid QA plan is not yet in place and presents a significant challenge to building more trust within the user community. On the other hand, the decision-support tool mentioned above and described in Figure 8 enables review of groups of duplicate reports and confirmation of the case that best represents the reported AE, namely the reference case. Although this process occurs in a case series analysis setting and cannot be done for all data, it may support a QA plan through this more limited evaluation of smaller data sets. This approach will indicate whether performance remains at the same level and if any correction strategies are necessary. The regulatory agency carefully reviews these aspects and plans to conduct periodic audits through this or other mechanisms as part of a QA strategy.

B5. Compliance with the governance framework

Table 12: 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. Periodic audits during the routine operation of this deduplication pipeline are essential to ensure the performance shown in the pre-deployment phase remains consistently high.	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

Principle	Activities	SPEC	DEV	PreD	PstD	RU
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. 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.	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 C: Artificial intelligence translation assistant

Source:[5]
Area of PV: ICSR reporting

C1. Business rational and challenges

Processing of ICSRs starts with the collection of information from the worldwide markets and the intake into the electronic database system/workflow. Typically, this involves many languages which require translation into English for the further processing steps by the global functions. Translation plays a crucial role as errors in the translation can lead to misunderstandings and wrong conclusions downstream. Furthermore, manual translation requires time and effort and coverage of all markets/languages by well-trained translators can be a challenge.

In the current example the pharmaceutical company had engaged with a vendor to consolidate and streamline the global case intake and translation process. The vendor had established two hubs in Europe and Asia to cover 16 languages across 32 countries replacing a distributed network of multiple local country organisations 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 centralised case management in English and subsequent submission to authorities. The pharmaceutical 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 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. Until today, the 100% QC by the HITL has been kept.

C3. Results

The translation's quality was assessed by BLEU scores. BLEU 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 six months, the AI translation assistant mimicked the quality of a human translator (i.e. BLEU equal or greater than 0.6).[6]

The results of the AI Translation Assistant pilot for the first language (Portuguese) were leading to a reduction of translation efforts by approximately 30%. Hence the solution

was extended to five further languages (Chinese, Dutch, French, German, and Spanish). The pharmaceutical 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.

C4. Challenges & Lessons Learned

Since the launch of the AI Assistant for Translation in 2021, translation quality has further improved, with less than 10% of outputs needing human corrections. The team is considering shifting from full human quality checks to sample-based monitoring, adjusting sample sizes as needed by language.

With the rise of GenAI since 2022, these tools now offer multiple features, such as extracting structured data, translating information, preparing case summaries, and translating reports for non-English regions, all within a single platform. Currently, several projects are under way to implement GenAI into the (commercial) applications available for ICSR management. These platforms again have the potential to increase the efficiency in ICSR management drastically. For now, the HITL will play a crucial role to ensure high quality.

C4. Compliance with the governance framework

Table 13: Use case C: Alignment with the governance framework (detail)

Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	Translation of incoming information bears the risk of mistakes, which may lead to wrong conclusions or assessments downstream. Therefore, a risk-based approach has been followed thoroughly and consequently measures have been taken to ensure full and continuous human oversight.	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 & 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

Principle	Activities	SPEC	DEV	PreD	PstD	RU
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 at 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	The application is not providing any interpretation of data, consolidation, or decision support. In the event that certain words or expressions (e.g. popular language) are not known to the AI assistant, the human translator steps in during the QC. The 1:1 match of the translation is always verified by the human QC.	A	A	A	A	A
Governance & Accountability	The translation assistant is a standalone tool owned by the vendor company. Hence, the regular lifecycle governance is executed by the vendor and available on request to the pharmaceutical 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 pharmaceutical company. Hence, in addition to the 100% human QC process by the vendor, the pharmaceutical 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

Use case D: Large Language Models for context-aware Structured Query Language

Source Article:[7]
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 SQL queries, leveraging a business context document.

D2. Solution

A sandboxed version of OpenAI's GPT-4 model was utilised within a 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.

The RAG framework facilitates the transformation of NLQs into SQL queries by harnessing its retrieval mechanism to access relevant information from the business context document. This enriched contextual data is then provided as input to the GPT-4 model, enabling it to generate SQL queries that are aligned with the specific schema and operational requirements.

D3. Results

Results showed significant improvements in query generation accuracy across three experimental phases. In Phase 1, using only the database schema, the LLM achieved a pass rate of 8.3% with 78.3% failing to generate valid SQL queries, highlighting the challenges of generating accurate SQL queries without contextual information. In Phase 2, the addition of a business context document increased the pass rate to 78.3%, and achieved a statistically significant improvement (P-value: 0.0006) compared to Phase 1. In Phase 3, which used a narrowed schema without the business context document, showed modest improvements reducing the failure rate from 78% to 50% compared to Phase 1, but did not match the performance achieved in Phase 2.

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. Challenges & Lessons Learned

The study highlighted challenges in automating SQL query generation for PV databases using LLMs. One limitation was the difficulty in handling high-complexity queries, as the LLM struggled to generate accurate SQL code when faced with intricate database relationships and ambiguous user intents. This challenge was particularly evident in Phase 1, where the absence of contextual knowledge resulted in a failure rate of 78.3%. While the introduction of a business context document in Phase 2 significantly improved performance, achieving a pass rate of 78.3%. Another consideration for implementation would be the need for

domain experts to construct and maintain the business context document, as updates to the database schema could impact its utility.

D5. Compliance with the governance framework

Table 14: 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, controls and processes needed to ensure its trusted use.	A	A	N/A	N/A	N/A
Human oversight	Within the PoC human experts reviewed the relevance of the outputs against a reference standard. Within a product setting consideration would need to be given to how human oversight will ensure robustness of the outputs, including confirming if the correct data has been extracted. As the database schema may change, thought should also be given to monitoring performance over time and at defined intervals.	A	A	N/A	N/A	N/A
Validity & Robustness	The tool has been evaluated against a curated reference standard. Beyond the PoC, consideration would need to be given to generalisability in production use including ensuring outputs are correct based on the user's requirement.	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 that does not use individual patient data to generate SQL outputs.	N/A	N/A	N/A	N/A	N/A

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Governance & Accountability	During the 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 scientific integrity principles are adhered to, while future product use governance should cover how the tool fits into the overall PV system and 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

Use Case E: Causality assessment of adverse drug reactions

Source:[8]
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, determining the expedited reporting requirements for each ICSR. 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 15: Use case E: Modelling Data

Source: Modified from Cherkas Y, et al, 2022 [9] Table reproduced with permission

Modelling 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 utilised 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 PPV 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. Challenges and lessons learned

Data availability presents several challenges but also opportunities for improving the model through addition of new features and allowing further validation of model performance.

The lack of well-annotated causality data from additional sources limited the exploration of the model's performance for drug-event combinations not included in the internal data set. Creation of a public reference set, while itself likely to be challenging, introduces opportunities to validate such models and compare their performance with other methodologies across a wider spectrum of drugs and events.

The study used a limited set of related clinical trial cases to establish an estimate of plausible time-to-onset between exposure to the drug and the event onset. More comprehensive datasets (e.g. EHRs) containing such data could provide potential for improving this feature of the model design. Use of drug mechanism of action data in the time to onset feature may also help improve model performance.

Access to drug label data would support the addition of features to identify whether events are labelled for any drugs in the case or drugs from the same class as the drug under review. Similarly, incorporation of data on medical conditions and drug indications might be used to identify confounders, including whether the reported reaction is a symptom of an existing medical condition or associated with a concomitant medication's indication.

Variability in the causality assessments for drug-event pairs is well documented and presents a potential challenge in ensuring transparency when designing models to support this activity. The development of a validated decision-tree tool (CASCADE) provided a structured, consistent, and transparent approach that helped inform the topology of the resulting model and demonstrated the value of integrating expert clinical knowledge into ML models although interpretability of the model remains a challenge that needs to be addressed.

E5. Compliance with the governance framework

Table 16: Use case E: Alignment with the governance framework (detail)

Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based Approach	The scope of the work and resulting model was limited to solicited, post-marketing cases. The 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. There is 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 used in this study were limited to a specific period where causality assessment methods were consistently applied across a broad range of product and event categories to increase the reliability of the resulting model. Model training and testing activities are described in detail, as is the approach used for performance assessment. The authors consider areas for investigation that could be used to further demonstrate the model's validity and improve robustness including the availability of a public reference set of drug-event causality assessments.	A	A	N/A	N/A	N/A
Transparency	There is a focus on transparency throughout the paper. Information about intended use of the model and its design are provided. A decision tree tool (CASCADE) designed to provide clear rationale for the resulting causality assessment was created and informs the design of the resulting model. 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

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Data Privacy	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 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

Use Case F: Process efficiencies supporting signal detection

Source Article:[10]
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 summarised in the flowchart below.

Figure 6: Flowchart summarising the implementation of the automatic selection of controls and the dismissal of false positive signals when using a conditional inference tree

A dual approach combining the ATC classification system code and the approved therapeutic indication in the US Prescribing Information (USPI) of galcanezumab was used for automatising 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 analysed by using conditional inference trees to identify alternative cause co-reported drugs. The USPI of disproportionally co-reported drugs was screened to identify those drugs that listed in the USPI 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 USPI as these cases had alternative causes for the AE.

F3. Results

By using conditional inference trees, the framework was able to dismiss 20.00% of erenumab, 14.29% of topiramate, and 13.33% of amitriptyline disproportionality signals on the basis of purely alternative causes identified in cases, from within the control group. Furthermore, of the disproportionality signals that could not be dismissed purely on the basis of the alternative causes identified, the authors estimated a 15.32%, 25.39%, and 26.41% reduction in the number of galcanezumab cases to undergo manual validation in comparison with erenumab, topiramate, and amitriptyline, respectively.

The authors concluded that 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. Challenges & Lessons Learned

The study highlighted specific challenges in automating signal detection within the FAERS database using AI. One limitation was the lack of clear guidelines for control selection in disproportionality analysis. This is in part due to disproportionality analyses being frequently conducted against a background of the rest of the databases. Also, the choice of an appropriate control can sometimes be nearly impossible in a given dataset and the subjectivity associated with selection of controls. Nevertheless, this approach seems promising in some circumstances, in particular when the automatic process proposed in this study for the selection of controls within and outside the chemical subgroup of galcanezumab showed an 86% success rate.

Another challenge emerged from the manual process needed to verify alternative causes for adverse events through screening of co-reported drugs, as the AI framework did not fully address this step. Although the conditional inference trees could identify statistically significant differences in co-reported drug proportions, enhancing the dismissal of false-positive signals, there remains a need for ad-hoc tools to automate Summary of Product Characteristics (SmPC) checks.

In addition, the framework did not establish a clear cutoff for the number of drug classes to identify viable controls, nor did it determine the optimal number of controls for disproportionality analysis, highlighting a need for further research. Also, while the number of alerts changes proportionally with the number of controls, developing systematic criteria to effectively manage this multiplicity issue in practice remains challenging, necessitating further validation across different drugs and databases.

F5. Compliance with the governance framework

Table 17: Use case F: Alignment with the governance framework (detail)

Source: CIOMS Working Group XIV

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Risk-based approach	This study adopted a systematic approach to identify and mitigate the impact of false positives in signal detection processes. 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 may be required to support broader deployment.	A	A	N/A	N/A	N/A
Human oversight	In the study, medical experts played a crucial role in manually validating AI-generated outputs to ensure accurate signal detection. The article emphasised the need for human review due to the complexity and variability in each case, underscoring the importance of involving domain experts in interpreting AI findings. The oversight involved identifying alternative causes for adverse events that AI might flag, ensuring alignment between algorithmic predictions and clinical knowledge.	A	A	N/A	N/A	N/A
Validity & Robustness	The article noted the need for validation and testing using FAERS data and simulations. Implementing controlled tests and optimising control selection addressed stability and prediction reliability across diverse drugs and spontaneous reporting databases.	A	A	N/A	N/A	N/A
Transparency	Transparency was considered through documentation of the methodologies employed and the rationale for control selection, addressing variability impacts.	A	A	N/A	N/A	N/A
Data Privacy	The method uses publicly available information on labelling alongside FAERs data which required limited consideration for data privacy.	N/A	N/A	N/A	N/A	N/A

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Fairness & Equity	The study emphasised the necessity of ensuring comparability and inclusivity in control selections to prevent bias. It underscored assessing data variability and comparability among drugs to ensure fair representation, addressing inequities possibly introduced by inadequate data collection or control dynamics. This was important for balancing comparability complexity and broader drug class inclusion.	A	A	N/A	N/A	N/A
Governance & Accountability	During the 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 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

Use Case G: Generative Artificial Intelligence: synthesis and summary from a large unstructured safety document repository for facilitating pharmacovigilance evaluations

Source: Internal to CIOMS Working Group XIV member organisation
Area of PV: PV document retrieval
Note: This use case outlines the implementation of a GenAI solution by one CIOMS Working Group XIV member organisation to enhance its PV-related work processes. It is important to note that many MAHs and Regulatory Authorities (e.g. US FDA's ELSA) globally are exploring GenAI technologies for potential work enhancement.

This use case reflects an approach that was developed and implemented prior to the release of this CIOMS guidance. As a result, the practices described within this use case may not fully align with the best practices or recommendations.

G1. Business rational and challenges

The curation of data to support PV evaluations – from safety analyses and signal assessments to aggregate reports and regulatory authority safety requests – is time- and labour-intensive, often requiring search, retrieval, review, and summarisation of vast amounts of unstructured safety data, from text-heavy clinical study reports and dossiers submitted to health authorities to extensive legacy safety analyses, and to signal assessments and scientific literature. LLMs, such as the GPT models can facilitate this effort for PV professionals with the capability to summarise and synthesise unstructured safety data from broad and/or varied repositories.

G2. Solution

With hundreds of thousands of documents rich in safety-related data, LLMs were employed to optimise the search, retrieval, review and summarisation of unstructured safety data in a manner specific to the parameters and requirements of a human PV professional.

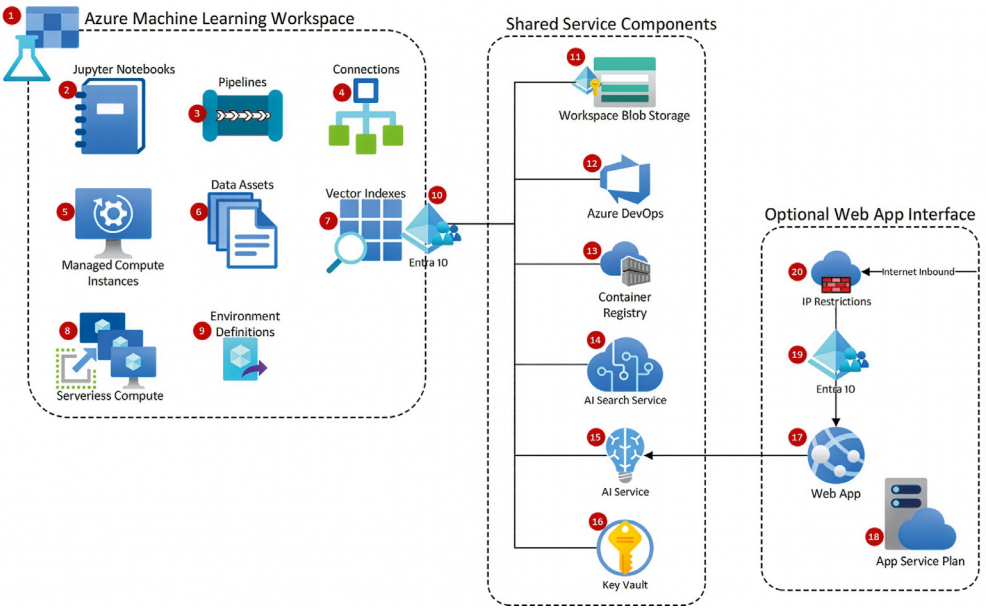
A Custom AI Search engine tool was developed using C# & .Net Framework and leverages Microsoft's semantic kernel SDK and RAG pattern to allow PV professionals to interact with unstructured safety data within private vector stores using a large language model with custom systems instructions. PV professionals submit safety data-related queries via a web front end which are routed to a private deployment of LLM along with context (data retrieved in-line from the vector store) and system instructions. Azure OpenAI provides rEpresentational State Transfer (REST) application programming interface (API) access to a powerful and diverse set of models (OpenAI Chat, OpenAI text embedding, GPT- 4.1 Series) and integrates these models with a large and diverse repository of unstructured safety data.

With access to the "text-in, text out" interface of the OpenAI model, PV professionals can provide an input prompt and the model generates text with usage of OpenAI and GPT-4.1 Series models which facilitate interactive conversation with text-based inputs and responses; this also leverages Open AIs embedding models which converts text into dense vector representations for NLP tasks.

By automating the process of retrieving relevant information, PV professionals can redirect their time towards value-added endeavours rather than manual data sifting. The figure below outlines Azure LLM architecture and interface with data assets.

Figure 7: An outline of our initial artificial intelligence architecture

Source: Internal to CIOMS Working Group member organisation. Figure reproduced with permission.



Development and deployment of a meticulously crafted vector index of relevant company file structures allows the organisation to leverage an LLM to easily and efficiently navigate documents, files and data available within those company files. By transitioning to newer models as they release, the organisation 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 minimises errors. Guidelines were also developed for framing questions to the model effectively, further enhancing the tool's usability.

G3. Results

The GenAI 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.

For the current status of this project, the PV users are training to better understand and apply GenAI tool capabilities, particularly for summarising and retrieving safety data. Although no formal metrics were collected, QC was performed by the users and shared with the GenAI developers in the project team. The use of the GenAI tool by multiple PV users has been instrumental in evaluating the tool's accuracy and performance. The manual verification process allowed the users to assess that the tool provides relevant search results, retrieves the correct information, and summarises the source materials accurately. Based on the feedback, the GenAI tool is being refined to deliver efficient and faster responses and include downloadable files of the safety data references. Enhancements include the integration of clearer instructions and contextual prompts to support more precise and relevant answers.

Looking ahead, there are plans to expand this application of GenAI, focusing on areas such as extraction and summarisation of safety literature for signal detection purposes and outputs for case reporting. In addition, there may be potential to use GenAI applications to extract and create summary information for aggregate safety reports, support audits and inspections or support tasks related to benefit-risk assessment. For example, to the question on "EMA expedited reporting", the GenAI tool was able to review and locate the appropriate documents from the very the large document repository and instantaneously provide extracted outputs in the form of tabular summaries of the expedited reporting requirements along with references of the source documents to support PV with case processing and safety database configuration.

Ultimately, leveraging AI-driven document retrieval and summarisation from large document repositories may help PV professionals in performing critical medical and scientific evaluations of safety information more efficiently, thereby enhancing product safety and efficacy.

G4. Challenges and Lessons Learned

The main challenges of OpenAI were high and unpredictable cost, a gap in required AI expertise, concerns relating to data privacy and security. Once these challenges were identified the company quickly developed methodologies to monitor and control API costs, optimising prompt design, deployed indexing, and established data security and privacy protocols to protect potentially sensitive information.

For implementation of GenAI projects cost control / cost capping at project onset to control project budget is highly recommended. Availability of AI experts / vendor at the early stages of the project would facilitate project development, while creating an indexing system improves efficiency of users in asking the targeted questions.

As part of further development of the program, it is envisaged that user quality assessments would be systematically collected and evaluated with feedback sessions that, over time, will build up a knowledge / experience base for developers to continually improve and enhance the GenAI tool.

G5. Compliance with the governance framework

During the development and pre-deployment phases, the GenAI project was carefully developed and managed with a limited scope, ensuring alignment with the applicable guiding principles as indicated in Table 18.

As the project transitions into production and its use 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. This is explained in more detail in Table 18.

Table 18: 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 use, risks will be regularly reassessed. Therefore, all phases are considered partially aligned with this guiding principle. As described above specific data privacy and security protocols were developed.	A	A	A	N/A	N/A
Human oversight	Fully aligned in development phase, as there is human oversight from the user. Moving into production use, consideration will need to be given to the level of human oversight required to mitigate against known risks of GenAI e.g. hallucinations and automation bias. In addition to individual accountability of the output.	A	A	A	N/A	N/A
Validity & Robustness	Validation and testing were conducted based on the appropriateness of the results. Once in post deployment and routine use, the data sets are very large; however, expansion of use cases will follow a similar trajectory of human testing. Any inaccuracies in information retrieval will serve as valuable feedback for the GenAI developers to further refine and update the tool. Whilst no formal metrics were collected, quality control from the perspective of the users to collectively review and assess the results and GenAI outputs is expected.	A	A	A	N/A	N/A
Transparency	As the GenAI solution expands its scope and complexity during post-deployment and routine use, further realignment is anticipated to support post hoc transparency to the end user of the system. Transparency in relation to the public is not applicable as this is a closed system.	A	A	A	N/A	N/A

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Data Privacy	Fully in alignment during all phases. All data remains internal within the company. Also, role-based access and restrictions are applied. For example, individuals from the organisation's safety department would not have access to unblinded clinical trial safety data in the document repository, i.e. search outputs would remain blinded.	A	A	A	N/A	N/A
Fairness & Equity	Inherent limitations and biases exist within safety data which may manifest themselves within GenAI based outputs. PV professionals are aware of the limitations of safety data to limit the impact of bias.	A	A	A	N/A	N/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	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

Use Case H: Artificial intelligence to support diagnosis and prediction of (hydroxy)chloroquine retinopathy

Source:[11,12,13,14,15]
Area of PV: PV 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 of progression, therapy duration and attendant therapeutic benefits might be maximised. 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 as age-related macular degeneration (AMD) and diabetic retinopathy (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, nine minimal retinopathy, and nine manifest retinopathy) monitored with mfERGs and perimetry for a 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) on mfERGs to identify hydroxychloroquine retinopathy in 1463 eligible eyes (748 predominantly female

patients), of which 95 eyes (48 patients) were eligible for inclusion as controls. Very recently, Woodward-court et al reported the development and application of a convolutional neural network (CNN) to detect the presence and predict future development of hydroxychloroquine retinopathy from SD-OCT by calculating a Likelihood of Retinopathy Score (LRS). The study is notable for a larger and more diverse dataset involving 409 patients (171 positive for hydroxychloroquine retinopathy and 238 negative) and 8251 SD-OCT b-scans (1988 volumes) from five independent international clinical locations representing relatively diverse self-reported racial or ethnic groups, as well as two different SD-OCT technologies.

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. Woodland-Court reported that their CNN-based algorithm was able to detect hydroxychloroquine retinopathy at the time of clinical diagnosis, and with a substantial lead-time before clinical diagnosis (mean: 220.8 days before clinical diagnosis; accuracy: 0.987 [95% CI: 0.962—1.00]; sensitivity: 1.00 [95% CI: 0.833—1.00]; specificity: 0.983 [95% CI: 0.952—1.00]; PPV: 0.944 [95% CI: 0.836—1.00]; negative predictive value: 1.00 [95% CI: 0.937—1.00]). For eyes that developed retinopathy, the average lead time relative to clinical diagnosis was 2.74 years. The algorithm also demonstrated face validity based on the high coefficient of determination (0.93) for LRS between left and right eyes and the temporal evolution of LRS consistent with the known clinical trajectory of this retinopathy.

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.

Challenges and Lessons Learned.

Although the most recent cited study by Woodward-Court was an advancement relative to previous studies in several respects, the study populations were more/less small and limited or imbalanced in various respects according to the study. Because the clinical scenarios for drug use often involve autoimmune disorders, the subjects were predominantly female. Asians were under-represented in study samples and there was a need for further assessment in larger and more diverse populations. Nonetheless, the most recent study by Woordward-Court was larger and more diverse than previous studies, and also included and an assessment using two different SD-OCT instruments. Over the multiple geographically distinct data sets, instrumental variability normally presents a potential generalisability challenge that is not always accommodated. As is often the case in AI diagnostic applications involving retinal pathology, retinal comorbidities were excluded, or under-represented, which limits generalisability to more diverse patient populations that have multiple retinal comorbidities (e.g. diabetic retinopathy and drug-induced retinopathy). 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.

Importantly, the most contemporary of the cited studies strongly emphasised their solution within the context of the challenges and corresponding desirable features of diagnosis from health-care systems perspective, namely, limiting the patient and the health care system burden using a single, widely available, automatable diagnostic solution that could "democratise" diagnosis to clinicians of various specialisation levels.

The rarity of the disease would require large patient cohorts for such a clinical study, a significant hurdle to prospective validation. Validation of the prediction of future clinical retinopathy requires very lengthy patient surveillance. Prospective deployment in the clinic remains challenging due to the time and financial resources required for seeking regulatory approval for software as a medical device in many areas. Further work may include a financial assessment of deployment of the algorithm in the ophthalmology clinic to support decisions on future development.

Table 19: 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. 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 are not discussed. Discussed is the fact that the available human oversight in some locations may be provided by generalists with less experience and expertise than retinal specialists, affording more opportunity for incremental benefits in underserved settings.	A	A	N/A	N/A	N/A

Principle	Activities	SPEC	DEV	PreD	PstD	RU
Validity & Robustness	Partial alignment. Reference standards defined. One or more studies note the limitation of the imbalanced data sets used that impair generalisability. Also, in one/ more studies patients with other ocular pathology excluded so the two classes were HCQ retinopathy present versus normal retina, which limits generalisability 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.	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

Appendix 3 – References

1 Roemming H-J, Hauben M, Wannhoff W, et al. How LLMs can advance safety case intake—points to consider and insights from a proof of concept. Therapeutic Advances in Drug Safety. 2025;16. doi:10.1177/20420986251386222 (Webpage accessed 12 December 2025)

2 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. (Journal full text)

3 Kreimeyer K, Menschik D, Winiecki S, Paul W, et al. Using Probabilistic Record Linkage of Structured and Unstructured Data to Identify Duplicate Cases in Spontaneous Adverse Event Reporting Systems. Drug Saf. 2017;Jul;40(7):571-582. doi: 10.1007/s40264-017-0523-4. (Journal full text)

4 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. *Front Drug Saf Regul.* 2022;2:918897. doi:10.3389/fdsfr.2022.918897. (Journal full text)

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

6 Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL); 2002 Jul; Philadelphia, PA.* Stroudsburg (PA): Association for Computational Linguistics; 2002. p. 311–318. (Webpage accessed 23 September 2024)

7 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. https://doi.org/10.1093/jamiaopen/ooaf003 (Journal full text)

8 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. https://doi.org/10.1007/s40264-022-01163-6 (Journal abstract)

9 Modified from 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. https://doi.org/10.1007/s40264-022-01163-6 (Journal abstract)

10 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. https://doi.org/10.1007/s40264-023-01317-0 (Journal full text)

11 Fan WS, Nguyen HT, Wang CY, Liang SW, Tsao YM, Lin FC, Wang HC. Detection of Hydroxychloroquine Retinopathy via Hyperspectral and Deep Learning through Ophthalmoscope Images. Diagnostics (Basel). 2023;Jul14;13(14):2373. https://doi.org/10.3390/diagnostics13142373 (Journal full text)

12 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. Sci Rep. 2024;Oct22;14(1);24853. https://doi.org/10.1038/s41598-024-76943-4 (Journal full text)

13 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. Ophthalmol Retina. 2022;Dec1;6(12):1241-1252. https://doi.org/10.1016/j.oret.2022.05.031 (Journal abstract)

14 Habib F, Huang H, Gupta A, Wright T. MERCI: a machine learning approach to identifying hydroxychloroquine retinopathy using mfERG. Doc Ophthalmol. 2022;Aug;145(1):53-63. https://doi:10.1007/s10633-022-09879-7. (Journal abstract)

15 Woodward-Court P, Hogg J, Lee T, Taribagil P, Zhao CS, et al. Deep learning algorithm for the diagnosis and prediction of hydroxychloroquine retinopathy: an international, multi-institutional study. Ophthalmol Retina. 2025;9(1):1-10. https://doi.org/10.1016/j.oret.2025.06.003. (Journal full text)

APPENDIX 4.
CONTENT RELATED TO EXPLAINABILITY AND TO FAIRNESS & EQUITY

Illustrative examples related to Explainability

As stated in Chapter 6 on Transparency, it is essential to disclose why, when and how AI is being used in different PV tasks. This is to maintain trust, awareness, and responsibility among stakeholders, including developers, PV professionals and decision makers, regulatory authorities, HCPs, and patients. However, the requirements for explainability, and the manner in which it is employed, differ according to the context, for example, who is seeking the explanation, for what purpose, the nature of the task, and the stage of the system's lifecycle.[1] In the following sections, illustrative examples are presented to demonstrate the range of scenarios. The associated benefits of explainability as highlighted in the examples are summarised and subsequently synthesised at the end of this section.

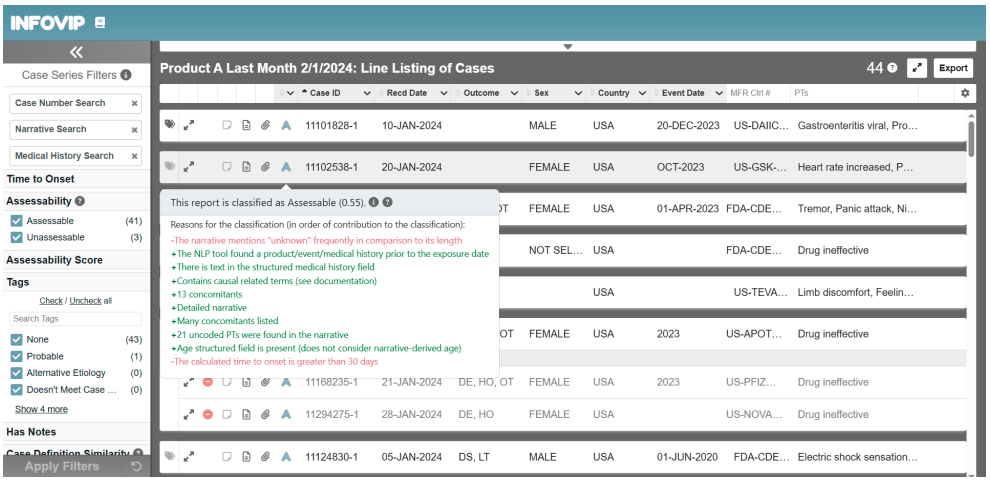
Examples of explainability in artificial intelligence-supported pharmacovigilance tasks

Consider a setting in which a PV officer is reviewing a case selected by an AI system as of interest, yet the rationale for this classification is not immediately apparent. In such situations, the reviewer may benefit from access to information indicating which text in the case data contributed to the AI's recommendation. An actual example of this is described below.

The Information Visualization Platform (InfoViP), developed for the US FDA's Center for Drug Evaluation and Research (CDER) is an example of how explainability may benefit the human experts engaged in signal detection and assessment supported by an AI system.[2] InfoViP uses NLP and several other components to process post-marketing data from multiple sources (FAERS, product labels, and biomedical literature) and provides a visualisation of the information, i.e. explanations, to support medical reviewers who detect and evaluate potential signals from the millions of adverse event reports submitted to the US 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 assessing 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.[3]

Figure 8: US FDA's Information Visualization Platform user interface illustrates the system capabilities, focusing on the features that positively contribute to classification for accessibility

Source: Botsis T et al, 2024[3]



The example above illustrates a core benefit of explainability described by Albahri et al (2023).[4] Explainability can facilitate human experts in making "sound and reliable" decisions. Ultimately, when the human decision and the accompanying explanation are retained, this information would nurture trust of the system owner and QA 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, post hoc explainability has resulted in increased trust and the perception of fairness in AI-supported decision making.[4]

Examples of pharmacovigilance stakeholders benefitting from explainability

While the likelihood of an individual from the general public requiring explainability in a PV setting may be small, the possibility cannot be excluded entirely as the use of AI becomes more commonplace. Some conceivable scenarios are described below:

— If a reporter (HCP or patient) reports a serious AE and the report is processed as a non-serious case by an AI triage system, the reporter may request an explanation from the MAH. Traditionally, the reporter could receive an explanation from the PV officer who has made the final triage decision. However, when this takes place in an automated AI triage process, a lack of explainability may impact trust and acceptance of the result by the reporter.

— If GenAI would 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.[5] Here, questions concerning the AI recommendations could be raised by both parties.

Examples of explainability in system development

A data scientist or ML engineer 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 explainability, supporting tools could provide explanations that facilitate troubleshooting by revealing what to change or exclude in order to "flip" the outcome.[6] However, in most cases, tweaking the system architecture of a deep neural network or specific features based on such insights can be quite challenging. These explainability methods are more likely to identify hidden biases in the training data which can be corrected as illustrated in the example below.

Ribeiro et al (2016)[7] demonstrates how Local Interpretable Model-Agnostic Explanations (LIME) could be leveraged to support explainability 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, demonstrating how post-hoc explainability methods 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, PV SMEs benefited from seeing which terms in the case were considered by the AI in its seriousness predictions. In this case, a LIME analysis revealed a focus on the drug name. Combined with the fact that the missed serious cases concerned Over the Counter (OTC) drugs, the PV SMEs discovered that the AI was basing decisions on the drug name and had 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 cases associated with OTC drugs or when there is no bias, and 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 artificial intelligence-systems interacting with health care professionals and patients

A hypothetical example can be the case of a HCP who is requesting product-specific information via a chatbot provided by a MAH. Such a chatbot could have multiple objectives ranging from the provision of drug product information to the collection of AE and quality defect reports. When the HCP notices that the chatbot response is inadequate, i.e. not considering key medical terms or AEs, or providing questionable information, the HCP may contact the MAH for an explanation.

Whilst the scenario above is a fictive example, one example of a chatbot that is currently available is the Smart Artificial Intelligence Resource Assistant for Health (SARAH) on the WHO 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.[8] 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.[9] 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 solution that interacts directly with the public, a lack of explainability may be an important consideration.

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.[10] Without the medical background, a layperson is at increased risk of harm by not being able to recognise incorrect information. In a recent systematic review and meta-analysis of 83 studies comparing GenAI models to physicians in diagnostic tasks, AI models achieved an overall diagnostic accuracy of about 52% (95% CI: 47.0-57.1 %) and performed comparably to non-expert physicians, but significantly worse than expert physicians.[11]

Thus, aside from an inability 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 solution 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.

Examples 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 does not require detailed explanations into how the AI system translated a specific piece of 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. In some such circumstances, not only will the human

responsible for the quality check be able to spot errors, but also understand the underlying reasons for machine translation errors. 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.[12] 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. In this setting, consideration may also need to be given to the risk of automation bias, where repeated exposure to seemingly correct outputs can cause the reviewer to become less vigilant and overly reliant on the machine translation. See also chapters on Risk-based approach, Human oversight, Data Privacy, and Fairness & Equity.

In PV, another example of AI use, which may not require explainability would be the automatic de-identification of case narratives presented by Meldau et al 2024.[13] In this case, a system using an LLM was trained to automatically detect likely person names and initials in case narratives for the purpose of redaction. While a HITL will not be able to know why a specific piece of text was highlighted as a likely person name by the LLM, this is not required for them to decide whether it should be redacted. More important is that the AI system has good enough performance, especially with respect to false negatives (missed names or initials), that human operators can rely on its output. At the same time, the lack of explainability of this method did present a challenge in assessing fairness and equity, when the only full name in the test set that was not redacted by the method was found to be of Indian origin, as discussed in the above-cited paper.

Example of how explainability may improve human processes and decision making

A final example illustrates how inherently explainable AI models may help improve human processes and decision making. In the evaluation of a statistical method for duplicate detection in AE reports,[14] a pair of Norwegian reports were identified as suspected duplicates, and ranked above all other pairs in the data set by the AI model. However, these two reports were not labelled as known duplicates and did not look like obvious duplicates to the human assessors: onset dates and ages that were close but not matching, and there were no exact matches on AE terms (although they were clinically similar). Inspection of the AI model's output revealed that its classification of this pair as likely duplicates by the AI model was driven by an exact match on six identical drug substances, which were not commonly co-reported. The cases were subsequently confirmed by the national regulator to be previously unknown duplicates that concerned the same incident but had been reported by two different physicians in the same hospital, thus accounting for the differences.[14] In this example (and in general), human assessors did not fully appreciate how unlikely it was for two independent reports to match on six distinct drug substances and therefore failed to lend this piece of evidence the appropriate weight in their assessment. Insights like this could be used to improve evaluation of suspected duplicates by human operators going forward.

Examples of methods supporting explainability

Some methods for post-hoc explainability in use at the time of writing this report include:

- LIME - see the example of Ribeiro et al (2016)[7] 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).[15]

- Trust scores that indicate the model's uncertainty for the output.[16]
- Confidence scores are a metric that is usually available and can be used to flag output that is uncertain for human review.[1]
- 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 to complete the view of the current landscape. See examples in Plass et al (2023).[16]

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 US FDA InfoViP described above as an example of explainability benefits.[2,3]

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. Although the tool is intended to support medical decisions, there have been observations that patients who are high health care utilisers with complex medical conditions may be discriminated against and underserved for pain management because of a "high risk score".[17] 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 and found that the NarxCare tool had a 17.2% false positive and 13.4% false negative.[18]

In this example, bias was introduced into the NarxCare tool because of inadequate data that did not account for subgroup factors (e.g. patients with complex medical conditions, healthcare models that result in multiple prescribers, lack of context for patients who require prolonged opioid use, lack of diagnostic information) potentially resulting in inappropriate

and misleading high patient opioid risk score predictions. The threat to fairness and equity for patients within subgroups who have a high score assigned because of bias, is that they 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 utilisers" were applied.[17]

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 CNN. The International Skin Imaging Collaboration (ISIC) database was used to train the neural network. Approximately 25 thousand images of skin lesions from the ISIC database were applied to the CNN. 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 for detecting skin lesions was less than 40% when the tests were carried out with Brazilian patients as compared to the higher accuracy of 90% with Caucasian patients. Thus, the potential for inequity in proactive treatment of skin lesions in Brazilian patients would be higher as a result of the low CNN accuracy detection rate.[19]

Example of Explicit negative bias

In Appendix 4, "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 AEs 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. In addition to the inadequate training data set, there was an explicit bias that it was not likely the OTC products in question would have serious AEs 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 HCP 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 potentially impacting safety risk identification and assessment could be seen as a threat to fairness and equity.

Appendix 4 – References

1 Royal Society. Explainable AI. Roy Soc [Internet]. 2024. (Website accessed 11 August 2024)

2 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 Saf. 2020;Sep;43:905-915. https://doi.org/10.1007/s40264-020-00945-0 (Journal abstract)

3 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)

4 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. Inf Fusion. 2023;Aug1;96:156-191. https://doi.org/10.1016/j.inffus.2023.03.008 (Journal full text)

5 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. https://doi.org/10.1016/j.japh.2023.11.023 (Journal full text)

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

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

8 World Health Organization (WHO). Using AI to lead a healthier lifestyle. WHO [Internet]. 2024. (Website accessed 21 March 2025).

9 Organisation for Economic Co-operation and Development (OECD). OECD legal instruments. Recommendation of the Council on Artificial Intelligence. OECD [Internet]. 2019. (Webpage accessed 21 March 2025)

10 Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. JMIR Medical Education. 2023;Jul10;9:e46939. https://doi:10.2196/46939 (Journal full text)

11 Takita H, Kabata D, Walston SL, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *npj Digit Med.* 2025;8:175 https://doi.org/10.1038/s41746-025-01543-z (Journal full text)

12 Savoldi B, Gaido M, Bentivogli L, Negri M, Turchi M. Gender bias in machine translation. Trans Assoc Comput Linguist. 2021;Aug;18;9:845-874. https://doi.org/10.1162/tacl_a_00401 (Journal full text)

13 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)

14 Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;14:305-328. https://doi.org/10.1007/s10618-006-0052-8 (Journal abstract)

15 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 Saf. 2022;May;45(5):583-596. https://doi.org/10.1007/s40264-022-01159-2 (Journal full text)

16 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. J Pathol Clin Res. 2023;Jul;9(4):251-260. https://doi.org/10.1002/cjp2.322 (Journal full text)

17 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 https://doi.org/10.1016/j.annemergmed.2022.04.006 (Journal full text)

18 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 https://doi:10.1016/j.drugalcdep.2021.109067 (Journal full text)

19 Artificial Neural Network Protocol in Dermatology. I-Workshop on Artificial Intelligence Applied to Health 2021 (I-WAIAH), 6th and 7th October 2021. (Feedback Report). (PDF accessed 29 September 2025)

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. The meeting minutes that document the report writing process can be found on the CIOMS website at www.cioms.ch.

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, Luisa	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
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

Pharmaceutical companies		
Name	**Company/Organisation**	**Country**
Lorenz, Denny	Formerly Bayer AG	Germany
MacEntee Pileggi, Elizabeth	Johnson & Johnson	USA
Patel, Ravi	United Therapeutics	USA
Reinhard Pietzsch, John	Bayer	Germany
Römming, Hans-Jörg	Merck KGaA, Darmstadt, Germany	Germany
Straus, Walter	Moderna	USA
Whitehead, James	AstraZeneca	UK

Regulatory authorities		
Name	**Company/Organisation**	**Country**
Buch, Brian	Medicines and Healthcare products Regulatory Agency (MHRA)	UK
Durand, Julie	European Medicines Agency (EMA)	The Netherlands
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
Kjær, Jesper	Formerly Danish Medicines Agency (DKMA)	Denmark
Ling, Benny	Health Canada	Canada
Da Luz Carvalho Soares, Monica	Brazilian Health Regulatory Agency (ANVISA)	Brazil
Matsunaga, Yusuke	Pharmaceuticals and Medical Devices Agency (PMDA)	Japan
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

Regulatory authorities		
Name	**Company/Organisation**	**Country**
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 met 11 times from 2022 to 2025, as below, and most of the meetings were hybrid in nature.

Chapter 1: Geneva, Switzerland	18-19 May 2022
Chapter 2: Geneva, Switzerland	10-11 October 2022
Chapter 3: Virtual meeting	19 January 2023
Chapter 4: Virtual meeting	12 April 2023
Chapter 5: Zurich, Switzerland	6-7 June 2023
Chapter 6: Virtual meeting	8 November 2023
Chapter 7: Virtual meeting	11 January 2023
Chapter 8: Geneva, Switzerland	7-8 March 2024
Chapter 9: Darmstadt, Germany	24-25 September 2024
Chapter 10: Geneva, Switzerland	25-26 June 2025
Chapter 11: Virtual meeting	8 September 2025

The CIOMS Working Group XIV Editorial Team met 28 times from March 2024 to October 2025.

APPENDIX 6.
PUBLIC CONSULTATION COMMENTATORS

Name	Company/Organisation	Country/ region
Adeleye, Qadri	National Postgraduate Medical College of Nigeria	Nigeria
Audibert, Francois	Vitrana Inc.	USA
Biswas, Aditya	Johnson and Johnson	USA
Bryant, Jason	ArisGlobal	UK
Burns, Ashley	PIC/S Good Pharmacovigilance Practices Expert Circle Working Group on AI and Machine Learning Deputy Chairperson, FDA	USA
Chaturvedi, Tanvi	Soterius, Inc.	India
Cho, Sylvia	US FDA	USA
Chouiyakh, Maria	Mohammed V University	Morocco
Dave, Jay	COD Research PVT LTD	USA
de la Peña Solís, Francisco José	Novo Nordisk	Mexico
El Hussien, Amira	RAY CRO	Egypt
Ezzeldin, Hussein	US FDA	USA
Freixas, Elisabet	Bristol Myers Squibb	Switzerland
Ghimire, Namita	Nepal Health Research Council	Nepal
Grigolo, Sabrina	EUPATI	Italy
Gutierrez, Israel	Caparna Inc.	USA
Hapani, Kalindi	COD Research PVT LTD	India
Heitmann, Martin	The Triality Group, LLC	Germany
Ho, Jeffrey	Perigent	UK
Iyer, Anand	Johnson and Johnson	USA
Jakubczyk, Jan	PIC/S Good Pharmacovigilance Practices Working Group on AI and Machine Learning, Polish Chief Pharmaceutical Inspectorate	Poland
Josephson, Aaron	Teva Pharmaceuticals	USA
Kessi, Una	HDI Safety, Oracle Health and Life Sciences	UK
Klueglich, Matthias	DGPharMed	Germany
Layton, Debbie	Lane Clark & Peacock LLP	UK
McAteer, Kaitlyn	Merck Animal Health	USA

Name	Company/Organisation	Country/ region
Nedog, Katarina	European Federation of Pharmaceutical Industries and Associations	Belgium
Nilaus Præstegaard, Søren	Novo Nordisk A/S	Denmark
Nishizawa, Claudio	ANVISA	Brazil
Orriss, Andrew	Kenvue	USA
Patel, Jiggar	Kenvue	USA
Pharmacovigilance Working Group	The European CRO Federation, EUCROF	Europe
Prajapati, Vatsal	COD Research PVT LTD	India
Prendergast, Christine	PIC/S Good Pharmacovigilance Practices Working Group on AI and Machine Learning, HPRA	Ireland
Radicke, Sophie	PIC/S Good Pharmacovigilance Practices Working Group on AI and Machine Learning, MHRA	UK
Sahu, Aneesha	US FDA	USA
Salem, Myriam	PIC/S Good Pharmacovigilance Practices Expert Circle Working Group on AI and Machine Learning Chairperson, Health Canada	Canada
Santana-Quintero, Luis	US FDA	USA
Schaeffer, Brian	Johnson & Johnson	USA
Scheerlinck, Rudi	Merck KGaA	Germany
Shee, Angela	Johnson & Johnson	USA
Singh Bedi, Simranjeet	Accenture Solutions Private Limited	India
Smith, Sean	US FDA	USA
Stockton, Brandi	The Triality Group, LLC	Germany
Thomas, Michael	American Society of Pharmacovigilance Physicians	USA
Trevett, Kiernan	Roche	USA
Tsvetanova, Antonia	Lane Clark & Peacock LLP	UK
van Hunsel, Florence	Pharmacovigilance Centre Lareb	The Netherlands
Viñas, Norbert	Vigintake SL	Spain
WENG, Xinyu	PIC/S Good Pharmacovigilance Practices Working Group on AI and Machine Learning, WHO	Switzerland
Wilson, Marie-Claire	Bristol Myers Squibb	Switzerland
Yuen, Alexander	Bristol Myers Squibb	Switzerland

Name	Company/Organisation	Country/ region
Zdorovtsova, Natalia	Lane Clark & Peacock LLP	UK
Zhou, Jessica	US FDA	USA
Anonymous	Sanofi	USA
Anonymous	Jazz Pharmaceuticals	Italy

This report on *Artificial intelligence in pharmacovigilance* addresses a rapidly emerging cross-disciplinary field that is at the intersection of pharmacovigilance, computer science, mathematics, 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 artificial intelligence in pharmacovigilance. The latter must be clearly defined and understood by many people from different backgrounds to propel research and practical implementation in an effective, safe and responsible manner. The diverse pool includes professionals, researchers, and decision makers working in pharmacovigilance in biopharmaceutical industry, regulatory authorities, and academia. It also includes software vendors that develop artificial intelligence solutions for pharmacovigilance, including signal management and all aspects of Individual Case Safety Report 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 one's organisation.

Artificial intelligence in pharmacovigilance. Report of the CIOMS Working Group XIV. Geneva: Council for International Organizations of Medical Sciences (CIOMS), 2025.

This publication is freely available on the CIOMS website.

CIOMS publications may be obtained through the publications emodule at https://cioms.ch/publications/. CIOMS, P.O. Box 2100, CH1211 Geneva 2, Switzerland, www.cioms.ch, email: info@cioms.ch.

ISBN 978-92-9036-110-7