



7 October 2025

Submission of
comments on

Annex 22 Artificial Intelligence

*Please note that these comments and the identity of the sender will be published unless a specific justified objection is received.
When completed, this form should be sent to the European Medicines Agency via the EU survey, in Excel format (not PDF).
Columns A to E should mandatorily be filled in prior to completing the columns "Comment" and "Rationale" and/or "Proposed wording".
For more details on how to use this template please refer to the tab "Manual for commenter" below.*

Country	Organisation raising comment (if no organisation, name of individual)	Line from	Line to	Comment (only one topic per comment) (max 600 characters)	Rationale (must be included when proposing a change) (max 600 characters)	Proposed wording (must be included when proposing a change) (max 600 characters)
USA	ISPE	0	0	General Comment 1: ISPE suggests that Annex 22 should be harmonized with other health authorities for terminology, definitions, and risk-based approach.	Lack of harmonisation will create unnecessary barriers for pharmaceutical manufacturers to adopt AI technologies and will hamper efforts to align on harmonised dossier content, such as through ICH M4Q efforts.	n/a
USA	ISPE	0	0	General Comment 2: The use of artificial intelligence (AI) and machine learning (ML) is new in the pharmaceutical industry and advancing rapidly. The development of these technologies should leave a high level of flexibility for industry to support the implementation of innovation in Europe and global development and manufacturing. ISPE suggests allowing and supporting the use of advanced and innovative AI/ML models in critical GMP applications in manufacturing. These models and supporting procedures should remain under control; however, industry should be supported to use them if rigorously evaluated and justified using risk-based approaches. These AI/ML models include, but should not be restricted to, dynamic models, those with probabilistic outputs, Generative AI, and Large Language Models.	Restricting the Annex to only the use of static models in critical GMP applications will restrict innovation, which is one of the four pillars of the European Commission's pharmaceutical strategy for Europe. ISPE recommends that Annex 22 allow the use of advanced models in critical GMP applications when justified based on science and quality risk management. There are metrics that provide a more accurate view of reliability and impact on quality and safety than traditional classification metrics like recall or F1, which do not apply to generative outputs. For example, key metrics include Brier score (calibration), relevance (context alignment), coherence (logical consistency), and uncertainty quantification (confidence in outputs).	n/a

USA	ISPE	0	0	<p>General Comment 3: ISPE suggests clarifying in the Principles that QRM applies to AI. Annex 22 should not repeat the details of QRM implementation but should refer to Chapter 1 and GMP-related documents, Q9, Quality Risk Management in Eudralex Volume 4, Parts I and III.</p> <p>Additionally and very importantly, ISPE considers that Annex 22 may not reflect a true risk-based framework as advocated in ICH Q9, despite establishing QRM as a principle. The document appears to introduce a binary classification of AI applications into “critical” and “non-critical” GMP applications based on whether the AI system has a direct or indirect impact on patient safety, product quality, or data integrity. ISPE considers that a binary classification oversimplifies the spectrum of risk associated with AI systems.</p> <p>ISPE recommends that Annex 22 should build on the foundation of model categorisation given in ICH Q8/Q9/Q10 Points to Consider guidance.</p>	<p>Attempting to repeat requirements of other Chapters or Annexes in Volume 4 has the strong potential over time to lead to inconsistencies between the documents.</p> <p>Risk is not a binary concept and this is clearly described in ICH Q9(R1), the revision of which was led by EMA. As an example, ICH Q9(R1) says "Formality in quality risk management is not a binary concept (i.e. formal/informal);.....". Additionally, ICH Points consider (R2) ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation in Section 5.1 provides an example of categorisation of models into 3 levels of impact on product quality.</p>	n/a
USA	ISPE	0	0	<p>General Comment 4: ISPE recommends that this Annex 22 contain only WHAT regulators want to see and avoid describing HOW industry should implement these requirements.</p> <p>We suggest making a guidance leaving flexibility for improvement to industry, with no HOW description.</p> <p>As an example, for algorithm development, we suggest having high-level requirements such as transparency, diversity, no subjectivity, verification of the model training and testing. We suggest not giving too descriptive requirements to industry and allowing industry to make proposals in a controlled way supported by appropriate justification.</p>	<p>Making HOW as requirement could have a negative impact as the annex may discourage the use of modern best practice and current and innovative technology. This could lead to earlier pressure to update the guidance and has the potential to discourage innovation.</p>	n/a
USA	ISPE	5	8	<p>ISPE considers that care should be taken in the language linking Annexes 11 and 22, for example the word "embedded" in this context is hard to understand.. Alternative text is suggested.</p>	<p>ISPE considers that some elements of computerised systems apply to AI; however, it is difficult to understand what "embedded" means.</p>	<p>ISPE suggests modification to the text: "The document provides additional guidance to Annex 11 for computerised systems" much of which applies to AI applications.</p>
USA	ISPE	12	21	<p>ISPE suggests that the limitation of scope to deterministic / fixed models should be changed, please. Restriction of use should be loosened to allow for more flexible use of the technology, such as dynamic continuous/automatic learning models and probabilistic models for critical applications. Rather than excluding these models, the guideline should provide the expectations for their use.</p>	<p>To support innovation and advancement of pharmaceutical manufacturing, the Annex should be written with the potential to introduce future new technologies, such as appropriately self-validating models, with consideration to how and where human-in-the-loop interventions should occur.</p>	<p>ISPE suggests that the text in lines 12 to 21 is amended to provide high level expectations for more advanced than static models, for example such as expectation for levels of control based on quality risk management principles.</p>

USA	ISPE	29	29	ISPE considers that the term "validation" in this context may be a confusing term, that points to model performance evaluation by means of the validation data set. ISPE suggests a different term when applying the model to the validation data set	Validation in this context may lead to confusion between the data science interpretation of validation data sets and the computer system validation (CSV) interpretation as used in Annex 11 (Annex 11 being referenced in this draft Annex 22).	Suggest renaming the term as "evaluation of the model against the validation data set", and making the same change in line 90 also.
USA	ISPE	33	37	ISPE considers that sections 2.2 and 2.3 should be revised to avoid redundancy with existing GMP guidance and to focus more specifically on AI-related quality risk management. Please refer to General Comment 3 which requests to describe QRM requirements mainly in the scope of the document and as little as possible in the clauses. We suggest that the high-level description should come here.	Both sections repeat general GMP expectations already covered in Annex 11 and ICH Q9. To add value, the document should highlight AI-specific documentation needs and risk-based considerations.	Sections should be omitted or refer to other relevant documents e.g. Volume 4, Parts I and III and only include AI-specific elements .
USA	ISPE	39	46	ISPE recommends that AI models could also be applied to new processes, provided that adequate process understanding and risk controls are in place. ISPE suggests that the scope is amended.	The phrase "in-depth knowledge of the process" suggests that only existing, fully understood processes are eligible for AI support. This could significantly limit applicability. It should be clear whether AI can also be applied to new processes, provided sufficient process understanding is ensured. Restriction of application of AI to existing, fully understood processes would limit innovation and ISPE considers this restriction of innovation may not be in line with the European Commission's pharmaceutical strategy.	ISPE considers rephrasing to indicate that AI models can also be applied to new processes, provided that adequate process understanding and risk controls are in place. Limitations in the model selection, capability, and inputs that could potentially result in erroneous and biased results should be identified with mitigation measures justified.
USA	ISPE	47	51	ISPE suggests that explaining why subgrouping is necessary would be beneficial here.	Explanation would provide context and rationale for the reader.	Suggested additional sentence at the start of the para: "Defining subgroups of inputs helps the user to ensure fairness, robustness, and generalisation by making them aware of potential performance differences across distinct groups."
USA	ISPE	59	61	4.1 ISPE suggests removing the examples since other metrics could be considered such as MAE (mean absolute error), Mean squared error (MSE) or RMSE (root mean square error) – commonly used for regression problems when evaluating the Models' results.	The choice of metrics is highly specific to the use case (e.g., classification or regression). Even if only meant as an example, such guidance may be perceived as prescriptive, while other options also exist in the case of classification models.	4.1 <i>Test Metrics</i> . Suitable, case dependent test metrics, should be defined to measure the performance of the model according to the intended use. As an example, suitable test metrics for a model used to classify products (e.g. 'accept' or 'reject') may include, but may not be limited to, a confusion matrix, sensitivity, specificity, accuracy, precision and/or F1 score.
USA	ISPE	64	67	ISPE recommends that clause 4.2 refers to the WHAT is required and not HOW it could be achieved. Deletion of the last sentence is suggested.	Rationale is given in General Comment 2 above.	Starting at line 65: process subject matter expert (SME) should be responsible for the definition of the acceptance criteria, which should be documented and approved before the start of acceptance testing. When setting acceptance criteria, data drifts also have to be taken into account, if applicable. This means to have, prior to the test phase, a full plan for the specific model.

USA	ISPE	68	69	ISPE recommends that the text in clause 4.3 could be clearer to indicate that the acceptance criteria for a model should be the same or equivalent to the process it replaces. Alternative text is suggested.	Acceptance criteria are likely to be a mix of (manufacturing) process performance and capability of the "method", Acceptance criteria, however, may not be exactly the same between the model and process it replaces. Acceptance criteria should be equivalent or higher. The acceptance criteria of the model should be to achieve performance at least as good as that of the process it replaces, where a direct comparison is possible. It is noted however that the use of AI may drive changes within the process that render a direct comparison impossible	4.3 <i>No decrease</i> . The acceptance criteria of a model, should be at least as high as , equivalent or higher to the performance of the process it replaces . procedure or tests that the model replaces .
USA	ISPE	71	124	ISPE suggests new titles for the paragraphs 5., 6., 7.. They are currently called respectively 'Test Data', 'Test Data Independency', 'Test Execution'	If the word 'Test' is used in its general meaning to test computerized systems, and it does not indicate the 'test set' used to test the performances of a model, the paragraphs should be revised in naming and concepts to avoid ambiguity. Explanation should be given (e.g. in glossary) in the meaning of the used term 'Test'.	New titles are proposed as: 5. Data management 6. Data independency 7. Model evaluation and documentation
USA	ISPE	72	72	"Full sample space" is a data science/statistical term which can be misunderstood by non-technical stakeholders or auditors. ISPE suggests rephrasing.	It may be interpreted as exhaustive testing. In addition, "Full sample space" implies testing every possible input scenario within the model's intended use — which is often not feasible in practice.	5.1 <i>Selection</i> <i>Current text:</i> Test data should be representative of and expand the full sample space of the intended use. Suggest rephrasing as "Test data must be representative of the full range of the intended use and must include relevant common, rare, and edge case scenarios. "
USA	ISPE	84	84	Suggest replacing "fully justified" with "justified using a risk-based approach".	It is unclear what "fully" means in this context.	Suggested rewording: "Any cleaning or exclusion of test data should be documented and justified using a risk-based approach."
USA	ISPE	87	87	6. Test Data Independency ISPE suggests that the term "Test Data" is explained as the data for the final performance verification of the final selected model. ISPE also suggests that ideally the name of this dataset should be changed so that it does not align with the GMP terminology.	When using the term "test" here it should be clear that it should be understood as less the formal term "test" used in data science and not the "test" as used in GMP and ICH guidances. ISPE recommends distinguishing the term "Test Data" used in model validation from its traditional interpretation in GMP and ICH contexts. In data science and machine learning, "test data" refers to a subset of data reserved for the final performance evaluation of a trained model. However, in GMP-regulated environments, the term "test" often implies formal, regulated testing procedures with specific compliance implications.	To avoid confusion and ensure regulatory clarity, ISPE suggests renaming this dataset to something more neutral to emphasize its role in model validation rather than regulated product testing. This distinction helps maintain data independency by clearly separating model development activities from GMP-compliant testing.

USA	ISPE	88	92	ISPE suggests that for Test Data Independency, describing the statistical principle is sufficient. While the principle is well accepted, ISPE suggests that text relating to HOW to implement test data independence should be removed since the HOW detail described may be very difficult to apply in practice.	The scientific principles underlying the section on test data independency for training-based models of all algorithm types are well accepted and are worth mentioning to also apply for AI, or rather, machine learning type models.	We suggest removing all sections and text after first sentence, i.e. end with "...is not used during development, training or validation of the model" on line 90.
USA	ISPE	93	97	ISPE suggests deleting paragraph 6.2 "Data Split" and merging the concept into the "Staff independency" paragraph (line 103-108)	Moreover the sentence "If test data is split from a complete pool of data before training of the model" has been avoided. Indeed this concept can be also applicable, for collectors of test data with respect to collectors of training data, in the case of test collection after the training and validation. A double-blinded process should be applied in any case.	This paragraph 'Data Split' could be linked to the 'Staff independency' paragraph (line 103-108) Staff independency. It is essential that employees involved in the development and training of the model have never had access to the test data and viceversa. This is also valid for people involved in effective procedural and/or technical controls of test. The test data should be protected by access control and audit trail functionality logging accesses and changes to these. There should be no copies of test data outside this repository.
USA	ISPE	121	124	For clause 7.4 ISPE suggests removing the example. There is no need to have too much detail relating to retention of test documentation, which is evolving very quickly and would benefit from flexibility.	Test documentation and its retention is a field that is moving quickly, and flexibility in requirements is desired.	Test documentation. All test documentation should be retained along with the description of the intended use. the characterisation of test data, the actual test data, and where relevant, physical test objects. In addition, documentation for access control to test data and related audit trail records should be retained similarly to other GMP documentation.
USA	ISPE	135	137	The current text is ambiguous... a classification is a prediction, and ISPE suggests it should be clarified.	When testing a model used to predict or classify data, the system should, where applicable, log the confidence score of the model for each prediction or classification outcome.	ISPE recommends the text is changed to: "When testing a model used for classification, the system should, where applicable, log the confidence score of the model for each outcome."
USA	ISPE	144	149	ISPE suggests avoiding the HOW in this clause and only including the requirement for configuration.	Rationale is given in General Comment 4 above.	ISPE suggests revised text as follows: <i>Change control.</i> A tested model, the system it is implemented in, and the whole process it is automating or assisting should be put under change control before it is deployed in operation. Any change to the model itself, the system, or the process in which it is used, 146 including any change to physical objects the model is using as input, should be documented 147 and evaluated to determine if the model needs to be retested. Any decision not to conduct 148 such retest should be fully justified.

USA	ISPE	153	158	ISPE recommends merging section 10.3 with 10.4 to make it more dedicated to AI. Alternative text is recommended.	Simplification is recommended to make the text more relevant to AI.	ISPE recommends the following text: Performance Monitoring The performance of a model should be continually monitored using appropriate metrics to detect changes in system behaviour or environmental conditions. Where appropriate, this should include monitoring the input sample space to ensure it remains within the bounds of the model's training data and intended use.
USA	ISPE	159	163	ISPE considers the current text of clause 10.5 contains too much HOW and is considered too prescriptive. ISPE recommends that the text should describe that the level of review should depend more on QRM and criticality of the process. Alternative text is suggested.	Such requirements may discourage the use of ML. Instead of requiring a consistent manual review or testing of every model output. There are automated monitoring systems that flag anomalies or performance deviations to SMEs based on predefined thresholds. This would enable focused reviews only when necessary, reducing the time burden while maintaining oversight according to the procedure.	ISPE suggests revised text as: Human review. When a model is used to give an input to a decision made by a human operator, (human-in-the-loop), and where the effort to test such model has been diminished, records should be kept from this process. The level of criticality (QRM) of the process and the level of testing of the model, should inform the level of review of the output and/or test of every output from the model, according to a procedure.
USA	ISPE	164		ISPE considers that the guidance lacks a sufficiently precise definition of "critical GMP applications." Although it attempts to differentiate between "critical" and "non-critical" applications based on their potential to have a "direct impact on patient safety, product quality, or data integrity," it does not provide a clear explanation of what constitutes a "direct impact." This omission leaves the definition ambiguous and open to interpretation.	As the scope of the draft guidance hinges on the interpretation of AI models used in "critical GMP applications," it is essential that the definition of this term be clearly articulated. In the absence of a precise definition, companies may adopt an overly cautious interpretation of the guideline's applicability. This could lead to unnecessary regulatory burden and resource allocation. More importantly, the resulting uncertainty may discourage or delay the adoption of AI technologies in manufacturing, potentially hindering innovation and continual improvement in GMP environments.	"Critical" should be defined in the Glossary as it applies to the scope of this Annex, perhaps considering the text in ICH Q8/9/10 Points to Consider document, section 5.1.
USA	ISPE	164	164	ISPE suggests aligning the glossary to be more in line with ISO and AI Act.	Clear definitions will ensure consistent interpretation across the industry and regulatory authorities.	n/a
USA	ISPE	164	164	ISPE recommends adding the term "Subject Matter Expert (SME)" to the glossary and including text to clarify who would be considered an SME.	This additional text will help clarify requirements/expectations for an SME.	n/a
USA	ISPE	164	164	ISPE suggests that the term "critical applications" should be defined in the glossary. ISPE recommends that the EMA partner with the US FDA to come up with a common risk-based framework.	ISPE considers that further elaboration of the spectrum of risk is required. ISPE suggests that EMA should consider a graduated risk model, such as the one proposed by US FDA.	n/a

USA	ISPE	165	168	ISPE proposes a rewording of the definition of Artificial Intelligence, taking cue from other official definitions considered applicable to pharmaceutical industry applications. The proposed definition expands the concept.	ISPE considers that the proposed definition of AI does not take account of other published definitions used in the pharmaceutical industry. It is recognised that the definition from the EU Artificial Intelligence Act has been used. The following is a list of references which has been used to draft the suggested proposed text: https://artificialintelligenceact.eu/ (EU AI ACT) https://interoperable-europe.ec.europa.eu/collection/digital-ready-policy-making/glossary/term/artificial-intelligence https://csrc.nist.gov/glossary/term/artificial_intelligence Source PHG Foundation "Synthetic data for development of AI alaMDs" - regulatory consideration (2025) FDA – AI/ML Software as a Medical Device Action Plan https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai	Artificial Intelligence: refers to machine-based systems designed to perform tasks that typically require human intelligence, such as learning, reasoning, or decision-making. These systems operate with varying levels of autonomy and may adapt their behavior after deployment. Based on input data, they generate outputs—such as predictions, recommendations, or decisions—that can influence physical or virtual environments.
USA	ISPE	169	170	The term “Deep Learning” is included in the glossary but is not referenced in the main text. ISPE suggests either removing it or integrating it meaningfully into the document if relevant.	Glossary entries should reflect terminology used in the document. If “Deep Learning” is not mentioned or required for understanding the content, its inclusion may be unnecessary or confusing.	ISPE suggests either removing the glossary entry or referring to “Deep Learning” in the main text where appropriate (e.g. in the Scope or Explainability sections).
USA	ISPE	171	171	ISPE suggests that the definition of Feature should be changed. An alternative definition is proposed.	The current definition is abstract and ambiguous.	ISPE suggests the following definition: Feature: An individual measurable property or characteristic of a dataset used as input for a model, often derived or transformed to represent meaningful information that helps the model learn patterns.
USA	ISPE	178	179	ISPE suggests simplification of "patterns/features" in the definition of "Models" is suggested.	Patterns can involve combinations of features.	Model – Mathematical algorithms with parameters (weights) arranged in an architecture that allows learning of patterns from training data
USA	ISPE	180	180	The term “Overfitting” is defined in the glossary but not referenced in the main text. ISPE suggests either removing it or integrating it into relevant sections, such as test execution or model validation.	Glossary entries should reflect terminology used in the document. If “Overfitting” is not mentioned or discussed, its inclusion may be unnecessary or confusing for the reader.	ISPE suggests either removing the glossary entry or referring to “Overfitting” in the section where model performance is discussed.
USA	ISPE	183	183	We recommend revising the term "Static" to "Static Model" for completeness. Additional text provided for clarification.	"Static Model" is considered a better title for the definition description.	Please consider changing to "Static Model"

USA	ISPE	185	185	ISPE suggests that 'Test Dataset' is used once in the Annex in section 5.2 and appears to indicate "test data", which is used in other sections of section 5. ISPE recommends that "test data" be defined, and a suggestion is given. Additionally, there is similar confusion between "training data" and "training dataset." ISPE considers that clarification would be useful. Definitions of "test data" and "training data" are suggested.	<p>Since it appears that there is confusion between "test data" and "test dataset", ISPE recommends that "Test data" term is used and defined. A suggestion is given. Additional ISPE considers that "training data" terms should be defined and not confused with "training dataset", which is currently in the Glossary but not used in the text. ISPE has given a suggestions for the definition of "training data".</p> <p>Source: The PHG Foundation's July 2025 report titled "Synthetic Data for Development of AI as a Medical Device (AIaMDs): Regulatory Considerations" provides a detailed framework for understanding how synthetic data can be used in the development and regulatory evaluation of AI-based medical devices.</p> <p>Eventually also Validation/Tuning data can be defined: Tuning Data: "Data typically used by the manufacturer of an AI system to evaluate a small number of trained models. This process involves exploring various aspects, including different architectures or hyperparameters (i.e., parameters used to tune the model for the task). The tuning phase happens before the testing phase of the AI system and is part of the training process.</p>	<p>Test Data: "Data used to characterize the performance of an AI system. These data are never shown to the algorithm during training and are used to estimate the AI model's performance after training. Testing is conducted to generate evidence to establish the performance of an AI system before the system is deployed."</p> <p>Training Data: "These data are used by the developer of an AI system in procedures and training algorithms to build an AI model, including to define model weights, connections, and components."</p>
USA	ISPE	185	188	The terms "training dataset", "validation dataset", and "test dataset" are defined in the glossary but are not explicitly used or explained in the main text.	ISPE considers that these terms may be important to understanding and implementing the requirements for AI models in GMP environments. Including them in the main text would improve clarity and consistency.	The terms "training dataset", "validation dataset", and "test dataset" should be explicitly referenced and explained in Section 5 "Test Data", in the context of their respective roles in model development.
USA	ISPE	187	188	ISPE requests clarification of "size smaller than the training set" in the definition of "validation dataset."	The current draft does not express an expectation or clear guidance.	ISPE recommends that clarification is required for the unfinished sentence in the definition of "Validation dataset".
USA	ISPE	187	188	ISPE suggests that the definition of "validation dataset" requires clarification. A revised definition is proposed, which includes reference to other validation approaches.	Other techniques that can effectively substitute the traditional fixed train/validation set approach are commonly used in industry, such as cross-validation techniques.	Validation dataset – The dataset used during model development, to inform on how to optimally train the model from training data; its size is smaller than the training set. Use of a "validation dataset" approach could be substituted by cross-validation approaches.
	END					