EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

PIC/S

<06 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
When completed, this form should be sent to the European Medicines Agency via the EU survey, in Excel format
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 organisati	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)
Germany	ECA Foundation and European QP Association	0	0	We are developing AI at a rate that is difficult to describe. We already (almost) have Artificial General Intelligence ("AGI", e.g. around the level of human intelligence). It is more likely than not that we will have "Artificial SuperIntelligence" at some point in the (near) future. As AGI and ASI will be able to process loads of data at a way faster rate than humans can, this may kickstart a whole set of technological advances. Under these circumstances today's "static vs dynamic" distinction may soon be obsolete. The Annex 22 should not only regulate the current generation of AI, but also anticipate the next. Otherwise, the framework may quickly lose relevance. The time it might take to amend Annex 22 again may be too long in order to catch up with the speed of the future developments.	Rule-based models can make the correct determinations with adequate accuracy, and costs and efforts to design and test AI-models outweigh those of rule-based models, it might be more sensible for companies to not invest in AI.	It might be worthwhile to compare the draft Annex 22vto the current developments in the United States of America. Whilst the FDA currently has only allowed 'locked algorithms" (FDA terminology, EMA uses "frozen models"), more complex systems are not inherently excluded. Rather, they are/will be subject to a 'Total Products Life Cycle' ("TPLC") approach, combined with a 'Predetermined Change Control Plan' ("PCCP"). This allows for predefined, validated updates to learning systems post-deployment, without undermining control or safety.
Germany	ECA Foundation and European QP Association	0	0	No comment in the document about mixture of models/pipelines, keyword RAG. In RAG the original extraction can be made by a static model (similarity scoring on static embeddings for example). Then the extracted results are given to an LLM to select the most fitting candidate. How is this handled by the Annex? In the given form, you are not allowed to use the LLM even though the selection is made by a fitting model and only the refinement on fitting candidates is done by the LLM.	I think our comment is along the same lines as -> Monika Hupfauf 0 0. We agree with the statements.	I think our comment is along the same lines as -> Monika Hupfauf 0 0. We agree with the statements.
Germany	ECA Foundation and European QP Association	16	19	Probabilistic decisions should be valid in the right circumstances. There needs to be a defined acceptance interval or range of acceptance. This range should be defined by a metric derived from an applicable reputable source from the field that this model is to be used in.	An AI cannot be deterministic, it is probabilistic. Item #9 - Confidence - emphasises the necessity to verify the reliability of the trained model, since it is probabilistic. An AI model should at less equally or better perform than the existing approach.	The document applies to models with a probabilistic output which, when given identical inputs, provide similar outputs.
Germany	ECA Foundation and European QP Association	20	25	LLMs can be successfully used for critical GMP applications as long as a human is in the loop	Not allowing the use of LLMs in GMP is creating a huge disadvantage for the pharmaceutical industry compared to other industries	LLMs should be used with great precaution and a risk assessment needs to be conducted for any proposed application of a LLM based solution ~~in a critical GMP application. The use of LLMs may be justified if a humain-in-the-loop approach is followed.~~
Germany	ECA Foundation and European QP Association	26	37	Since large language models are currently the focus of public debate, the specific risks associated with the training, use, and continuous monitoring of traditional classification models, e.g., DL algorithms, should be covered by the supplementary paragraph	The proposed paragraph should be understood as a supplement to Chapter "Principles" in order to cover the needs of traceability and specific requirements for traditional classification models. The traceability of model creation in general is essential. In order to draw conclusions about potential weaknesses throughout the model's life cycle in the event of problems and thus promote continuous learning, the identified risks, the choice of data for training and testing, the training parameters, etc. should be documented during the development of the model.	Traceability: Regardless of the task, such as the generation of new data or the classification of data into categories, the development of a model should always be traceable, and the risks should be properly identified and documented. For classification models, such as deep learning algorithms, training data plays a particularly important role. In addition to the area of application and the requirements specification, particular attention should be paid to the selection and composition of the training data, the definition of the classes, and the training and model parameters.
Germany	ECA Foundation and European QP Association	27	32	Point 2.1 refers to the need for personnel to have a "adequate and sufficient understanding" of the AI model used. To align with the terminology of the EU AI Act (AI literacy) and reduce regulatory burden, it may be good to have overlapping definitions. This term is increasingly standardised across EU legislation.	Reduce regulatory burden and allign with other harmonized legislation.	All personnel should have adequate qualifications **in AI literacy**, defined responsibilities and appropriate level of access.
Germany	ECA Foundation and European QP Association	30	31	Add "User" (user may be represented through SME but it is important to include the actual users in the design and development)	Ensuring a workable and compliant solution	This includes but may not be limited to process subject matter experts (SMEs), users, QA, data scientists, IT, and consultants.
Germany	ECA Foundation and European QP Association	39	46	Point 3.1 assigns responsibility to a process Subject Matter Expert ("SME") for documenting the intended use of the AI model, which must be "documented and approved" before acceptance testing. It would be helpful to clarify who is expected to approve this documentation and the necessary qualifications such person should have including any necessary follow up training, etc.	Clarifying whose responsibility it is and which qualifications and necessary follow-up training, etc. such person should have.	**An AI** process subject matter expert (**AI**SME) [...]. The AISME should have the following qualifications [...] and follow-up trainings [...] and [...] it should be documented and approved by the **AISME** [...]
Germany	ECA Foundation and European QP Association	68	70	Point 4.3 states that AI models may not reduce acceptance criteria and implicates that the performance of the replaced process must be known. While sensible in most cases, this requirement may (unintentionally) exclude the use of AI in areas where current systems are inadequate to gather data. Specifically in these areas, AI could help and give more insights. A too strict interpretation of this provision could mean that certain data cannot be gathered. For example, a somewhat more flexible (e.g. risk-based) approach could increase both innovation and long-term safety.	The currently proposed text potentially hinders further development of GMP processes, as elaborated in the comment.	The acceptance criteria of a model should be at least as high as the performance of the process it replaces.

Germany	ECA Foundation and European QP Association	72	75	In some cases it will be impossible to include "all common and rare variations". Propose to remove "all" from sentence. Taking automated VI as an example: samples that are showing a defect that do not fall into any of the pre-trained defect categories bzt still show a defect should not be classified as reject but as "to be inspected manually". This also alignes with what is written in 9.2	Avoid limiting the use of AI-based solutions to cases where all rare variations are known and prevent incompliance to Annex 22	It should be stratified, include all subgroups, and reflect the limitations, complexity and common and rare variations within the intended use of the model.
Germany	ECA Foundation and European QP Association	72	86	Since large language models are currently the focus of public debate, the specific risks associated with the training, use, and continuous monitoring of traditional classification models, e.g., DL algorithms, should be covered by the supplementary paragraph	The proposed paragraph should be understood as a supplement to Chapter "Test data" in order to cover the risks, which do apply specifically for traditional classification models. Specifically for classification modells, as they are typically used in automation prcoesses or production, test data needs to be collected under the same circumstances as training and validation data, e.g., same lighting conditions, same physical product characteristics, equipement settings etc.	Consistency: Test data needs to be collected under the same circumstances as training and validation data, e.g., same lighting conditions, same physical product characteristics, equipement settings etc. In addition, it should be ensured that the data used for training and testing is representative of the application domain.
Germany	ECA Foundation and European QP Association	72	86	Especially for large language models the training data is difficult or impossilbe to control, the specific risks associated with the collection and composition of test data should be covered by the supplementary paragraph. Parts of the paragraph are covered by 5.1 "Selection". Perhaps	The proposed paragraph should be understood as a supplement to Chapter ""Test data"" in order to cover the risks, which do apply specifically to generative models, like LLMs, stressing the importance of representativeness, completness and variants of test data. It is IMPORTANT to convey that it is not the quantity but the quality of the test data that is essential. The proposed paragraph "General" for the section "Test data" is already partially included in paragraph 5.1 "Selection," but opens alternative approaches dealing with applications where this requirement cannot be met.	General: For AI models, especially generative models such as LMMs, where training data is typically difficult to control, the testing is of essential importance. To limit the risks of application to an acceptable level, test data must be representative, complete, balanced and rich in variants, and it should be ensured that data outside the test population does not reach the model. Alternatively, the expected behavior of the model must be taken into account during development for previously unseen data (e.g., anomalies).
Germany	ECA Foundation and European QP Association	79	79	The term "very high degree of correctness" should be defined	In order to be able to determine "very high" a reference number needs to be introduced	provide a reference number and define "very high" as a percentage of this reference number
Germany	ECA Foundation and European QP Association	111	111	The terms "generalising well" and "satisfactory performance" should be defined	In order to be able to determine "satifactory performance" the requirement for performance in general needs to be defined	the term "generalising well" should be deleted and the term "satisfactory performance" should be defined as a performance equally or better than any performance based on manually controlled process
Germany	ECA Foundation and European QP Association	126	128	Feature attribution should not be limited for testing of models but even more relevant for model selection and model training and validation, which is then followed up by the testing of models. Further more, feature attribution should be also considered to be available during operation.	In the process of the model selection it is most important to choose the most suitable (pre-trained) model for the intended use. During model training and validation it is important to e.g. determine the training progress in order to detect under- and overfitting. Further more, feature attribution is a part of explainability visualization.	During model selection, training, validation and testing of models used in critical GMP applications, systems should capture and record the features in the test data that have contributed to a particular classification or decision (e.g. rejection).
Germany	ECA Foundation and European QP Association	128	130	Feature attribution is a necessary part of the ongoing AI explainability and should be a mandatory part of the training, validation and test of a AI model. Considerations should be made if the feature attribution is required in the operation phase (e.g. for review activities) to improve confidence.	The feature attribution should be a mandatory factor in the full model development phase starting after model selection from training, validation and test. The feature attribution is the foundation of the explainability visualization which is also needed during operation to allow production as well as QA to review the ongoing AI decisions during production.	Where applicable, techniques like feature attribution (e.g. SHAP values or LIME) or visual tools like heat maps should be used to highlight or log key factors contributing to the outcome for performance monitoring and review purposes during model training and operation.
Germany	ECA Foundation and European QP	138	139	The proposed regulation is using the term 'suitable' which should be defined as critical applications with direct impact on patient safety, product quality or data integrity are at hand.	A clear definition of 'suitable' in a mathematical / statistical sense is essential for consistent implementation across the industry.	Add a definition for 'suitable' into the glossary.
Germany	ECA Foundation and European QP Association	139	142	The proposed regulation is using the term 'very low' which is not suitable in particular ther should be definition or guidance as critical applications with direct impact on patient safety, product quality or data integrity are at hand.	A clear definition of 'very low' in a mathematical / statistical sense is essential for consistent implementation across the industry.	Replace the term 'very low' with a statistical/mathematical requirement
Germany	ECA Foundation and European QP Association	150	152	Annex 22 will be read by many people that are not SMEs for AI, suggest clarifying what configuration control exactly refers to	Aligned implementation across industry, meeting of agency expectations	Configuration control. A tested model should be put under configuration control (according to Annex 11) before being deployed in operation, and effective measures should be used to detect any unauthorised change.
Germany	ECA Foundation and European QP Association	153	155	The term "regularly monitored" should be defined	a minimum interval for monitoring should be defined	10.3. System performance monitoring. The performance of a model as defined by its metrics should be continuously monitored and regularly reviewed comensurate to the risk to patient, process and product.
Germany	ECA Foundation and European QP Association	164	164	the term "risk" should be defined	In order to implement a definition for "risk" either the definition of the AI Act or a new "GMP definition" should be introduced in order to define the term "risk". In this case the definition of "risk" from the AI-Act could be used. According to ICH Q9(R1) Risk definition: The combination of the probability of occurrence of harm and the severity of that harm (ISO/IEC Guide 51:2014).	Art 3 (2) AI Act: 'risk' means the combination of the probability of an occurrence of harm and the severity of that harm;
Germany	ECA Foundation and European QP Association	164	164	The term "intended use" should be specified	In order to define the term "intended use" either the definition of the AI Act or a new GMP definition should be introduced. As the AI Act already defines the term "intended purpose" it would be favourable to use the term "intended purpose" as defined in the AI Act instead of "intended use".	Replace "intended use" by "intended prupose" and add the definition of the AI Act Art 3 (12): "intended purpose" means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation.
Germany	ECA Foundation and European QP Association	164	164	The definitions for "training data" and "validation data" should be changed and replaced by the definitions used in the AI Act	These terms have been defined in a different way than the identical terms used and defined in the AI Act. This is confusing – either completely new GMP definitions should be introduced or the terms defined in the AI Act should be used. As the AI Act already defines these terms, it is recommended to use the definitions as set out in the AI Act.	Use the definitions of Art 3 (29) – (30) of the AI Act: "training data" means data used for training an AI system through fitting its learnable parameters; " validation data" means data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process in order, inter alia, to prevent underfitting or overfitting;
Germany	ECA Foundation and European QP Association	164	164	The definitions for "validation data set", "testing data" and "input data" should be changed and replaced by the definitions used in the AI Act	These terms have been defined in a different way than the identical terms used and defined in the AI Act. This is confusing – either completely new GMP definitions should be introduced or the terms defined in the AI Act should be used. As the AI Act already defines these terms, it is recommended to use the definitions as set out in the AI Act.	Use the definitions of Art 3 (31) – (33) of the AI Act: "validation data set" means a separate data set or part of the training data set, either as a fixed or variable split; "testing data" means data used for providing an independent evaluation of the AI system in order to confirm the expected performance of that system before its placing on the market or putting into service; "input data" means data provided to or directly aquired by an AI system on the basis of which the system produces an output;
Germany	ECA Foundation and European QP Association	183	183	It refers to "static – frozen models", something called in the US as locked algorithms. It is best to allign terminology.	See comment.	The term "locked algorithm" should be used instead of the term "static – frozen model"