7 April 2025

Dockets Management Staff (HFA-305)
Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852
via online submission to https://www.regulations.gov/

RE: Docket No. ***FDA-2024-D-4689 Draft Guidance: Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products***

Dear Sir or Madam,

The International Society for Pharmaceutical Engineering (ISPE) appreciates the opportunity to comment on the above-referenced draft guidance.

ISPE commends the FDA for its well-structured and clear guidance, which effectively outlines key considerations for AI model credibility assessment. ISPE's comments focus on enhancing clarity regarding risk assessment, third-party AI models, operational efficiencies, and specific AI model types.

ISPE is a not-for-profit organization of individual members from pharmaceutical companies, contract manufacturing organizations, suppliers and service providers, and health authorities. ISPE's 22,000+ members lead scientific, technical, and regulatory advancement throughout the entire pharmaceutical lifecycle in more than 90 countries around the world. ISPE does not take a political position or engage in lobbying activities or legislative agendas.

We appreciate the opportunity to submit these comments for your consideration. Please do not hesitate to contact me if you have any questions.

Respectfully,

Mike Martin
ISPE President and CEO
International Society for Pharmaceutical Engineering (ISPE)
North Bethesda, Maryland 20852 USA
Mmartin@ispe.org

Draft Guidance or Consultation Document title: ***FDA-2024-D-4689 Draft Guidance: Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products***

GENERAL COMMENTS ON THE DOCUMENT
It is the opinion of ISPE that the Credibility Assessment Framework guidance is clear, concise, and overall, well written. Overall, ISPE appreciates the level of detail in the consideration of manufacturing applications of AI, especially in the section “life cycle management”. The clear statement that it is not necessary to provide details of life cycle management in the dossier (and that they should instead be made available for review as part of the manufacturing sites' Pharmaceutical Quality System (PQS) is very valuable as guidance for regulatory submissions.
ISPE strongly agrees with the FDA’s statement that the guidance does not cover the use of AI models in drug discovery or when used for operational efficiencies that do not impact patient safety, product quality, or the reliability of results from a nonclinical or clinical study.
ISPE appreciates the two examples FDA included in the draft guidance to help sponsors understand how to approach the first three steps of the Risk-Based Credibility Assessment Framework. We encourage the FDA to include more examples, provide further guidance to sponsors, including for AI models that have an assessed medium risk, for example, that involve a supplier situation or the use of a dynamic, online-learning system.
FDA notes that the draft guidance focuses on “AI models more broadly,” though it does acknowledge that ML is currently the most utilized AI modelling technique used in drugs and biologics product lifecycles. However, despite suggesting that the guidance covers “AI models more broadly,” many of the recommendations are more tailored to supervised ML models and not to other models, for example, unsupervised ML models. We encourage providing clarity throughout the guidance on any specific recommendations for unsupervised models (e.g., clustering metrics) or where recommendations made for supervised models may not apply to unsupervised models (e.g., precision/positive predictive value (PPV)).
ISPE encourages the FDA to provide additional guidance for considerations when a sponsor is utilizing AI models developed, trained, maintained, or otherwise handled by a third-party vendor. In these circumstances, it may be challenging for sponsors to provide/document certain information that the FDA recommends in the draft guidance (e.g., model architecture, model parameters). ISPE recommends that the FDA identify mechanisms by which the FDA may access information the Agency considers appropriate to inform its regulatory decision-making that helps protect any third-party proprietary information. For instance, the Model Master File (MMF) may be an appropriate mechanism to share information about third-party AI models used in manufacturing.

Lines 385-386 of the draft guidance recommend that “performance estimates should be provided with confidence intervals.” Lines 456-457 include a similar recommendation for performance metrics. Moreover, for performance metrics specifically, it may be challenging to capture uncertainty, and certain assumptions about distribution would need to be made
The risk-based credibility assessment framework is well defined in each step and allows to have a good path to follow in the planning, collection, and documentation of information related to design, development, training, validation, and use of AI models.
Lastly, ISPE suggests that the agency could refer to guidance issued by CDRH on the Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions as it may apply to the use of AI systems for the development of drugs and biologics. Likewise, the recent terminology and approach proposed by the EMA Quality Innovation Group (QIG) (model maintenance protocol) could be utilized.

Specific Comments on the Text

ISPE indicates text proposed for deletion with ~~strikethrough~~ and text proposed for addition with **bold and underlining**.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
13	This guidance provides recommendations to sponsors and other interested parties…	This guidance provides recommendations to sponsors, **marketing authorization holders, manufacturers,** and other interested parties…	ISPE recommends indicating from the outset that this guidance applies to all stages of the drug product lifecycle.
19	credibility evidence, in the performance of an AI model for a particular COU.	credibility evidence in the performance **and range or area** of an AI model for a particular COU.	While the concept of credibility is helpful, ISPE suggests the definition of limitations may augment the concept, i.e., to determine the range within the context of use in which a certain degree of credibility is reached. Models may exhibit various levels of credibility across ranges of input and thus may require measures like indications of uncertainty or defining limitations on when a model result cannot be seen as being thoroughly supported by the data that it was created by and the evidence generated so far. These considerations may lead to decisions regarding the context of use, e.g., in which areas a model is applicable or in which areas a model may only be applicable with intensified human oversight and awareness of human operators for possible weaknesses in the model results.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
47-51	This guidance does not address the use of AI models (1) in drug discovery or (2) when used for operational efficiencies (e.g., internal workflows, resource allocation, drafting/writing a regulatory submission) that do not impact patient safety, drug quality or the reliability of results from a clinical or non-clinical study"	This guidance does not address the use of AI models (1) in drug discovery or (2) when used for operational efficiencies (e.g., internal workflows, resource allocation, drafting/writing a regulatory submission) that do not **directly** impact patient safety, drug quality or the reliability of results from a clinical or non-clinical study"	It is ISPE's opinion that this is a very important distinction. Particularly in the discipline of GenAI/LLM in "generating/analysing text," there are many additional use cases to the explicitly mentioned "drafting/writing a regulatory submission". Examples are used for PV applications, generating SOPs under GMP, or analysis of a text with GenAI/LLM to draw conclusions that lead to decisions. A further clarification on the "directness of impact" would be very helpful. Alternatively, the consideration of mitigation of risk on such "text applications" of Gen AI through the role of a human in the loop as a final decision-maker would be useful.
86-87	However, AI use presents some unique challenges. First, the variability in the quality, size, and representativeness of datasets for training AI models may introduce bias and raise questions….	However, **~~AI~~ ML** use presents some unique challenges. First, the variability in the quality, size, and representativeness of datasets for training AI models may introduce bias and raise questions….	Consider changing "AI" to "ML." In the introduction, the guideline calls out wider aspects of AI, not only machine learning. Here, however, the role of datasets is focused on training AI models. ISPE recommends either clarifying that the paragraph was meant to apply to the narrower range of Machine Learning or expanding the paragraph for further considerations (e.g., the use of test data for non-ML AI approaches, which would require similar properties).
92-97	Second because of the complex computational and statistical methodology underpinning these models, understanding how AI models are developed and how they arrive at their conclusions may be difficult and necessitate methodological transparency (e.g., detailing in the regulatory submission the methods and processes used to develop a particular AI model). Third uncertainty of the accuracy in the deployed models' output may be difficult to interpret, explain, or quantify.	Third uncertainty of the accuracy in the deployed models' output may be difficult to interpret, explain, or quantify. **In addition, AI model developers, AI systems end-users, testers, and auditors should follow an adequate training and should have suitable awareness of the whole AI solution within the intended use.**	In ISPE's opinion, from the second and third challenges arises another aspect (additional challenge): In addition, AI model developers, AI systems end-users, testers, and auditors should follow an adequate training and should have suitable awareness of the whole AI solution within the intended use.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
96	uncertainty of the accuracy in the deployed models' output may be difficult to interpret, explain,	Please see comments	The guidance mentioned four challenges, including "...may be difficult to interpret, explain…", but the following content discusses all four challenges except interpretability and explainability. ISPE suggests that the FDA should specifically address the terms "interpretability" and "explainability" as they pertain to the use of AI in the final guidance. ISPE recommends that "discussion for the challenges of interpretability and explainability is out of scope" is stated, or please add some guidance in this document relating to interpretability and explainability.
97-101	Finally, another challenge with some AI models is the potential for the model's performance to change over time or across deployment environments when new data inputs are introduced and these inputs differ from the data on which the model was trained (i.e., data drift) requiring life cycle maintenance of these models	Finally, another challenge with some AI models is the potential for the model's performance to change over time or across deployment environments when new data inputs are introduced and these inputs differ from the data on which the model was trained (i.e., data drift), requiring life cycle maintenance **and continuous performance monitoring of** these models. **Since data drift could happen, for example, for self-adapting models and pre-trained models used in different COU, it is recommended to give evidence of eventual re-learning and/or control processes applied and to limit as much as possible the use of just-trained models for purposes other than those for which they are created.**	Additional notes could make some concepts clearer. Continuous monitoring: In most of the self-evolving scenarios, where the model autonomously changes parameters to adapt (e.g., learning from the environment - reinforcement learning) and can converge towards non-optimal behaviors, the process of evolution should be constantly monitored and controlled directly by humans or by an external algorithm with human supervision. This same comment can also be made to Lines 436-440 and to Lines 528-532.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
121 - 136	Step 1: Define the question of interest that will be addressed by the AI model (see section IV.A.1 for details). • Step 2: Define the COU for the AI model (see section IV.A.2 for details). • Step 3: Assess the AI model risk (see section IV.A.3 for details). • Step 4: Develop a plan to establish the credibility of AI model output within the COU (see section IV.A.4 for details). • Step 5: Execute the plan (see section IV.A.5 for details). • Step 6: Document the results of the credibility assessment plan and discuss deviations from the plan (see section IV.A.6 for details). • Step 7: Determine the adequacy of the AI model for the COU (see section IV.A.7 for details).	Please see comments.	These steps appear to concentrate on the AI applications and models and do not include the validation of computer systems on which AI applications are run. The computer systems should be validated according to CFR Part 11 and associated guidance. Regarding the cadence of steps, also seen in conjunction with the detailed draft guidance provided, ISPE recommends clarifying which aspects should be considered primarily from a computerized system perspective (e.g., the context of use), and which from a model / AI function perspective. For instance, step 1 seems to handle the system aspect of the purpose within the process, while step 2 seems to be focused on the model perspective. However, further clarification could be added whether the credibility assessment as mentioned in step 6 is meant on a model or computerized system level, or both.
133 – 134	• Step 6: Document the results of the credibility assessment plan and discuss deviations from the plan (see section IV.A.6 for details). • Step 7: Determine the adequacy of the AI model for the COU (see section IV.A.7 for details).	Please see comments.	ISPE suggests that further guidance on the scope of such documentation would be helpful in the context of typical development activities. As many models may be created throughout the development process (sometimes hundreds or more), it may be overly burdensome to create a credibility assessment report for all such models. Therefore, ISPE recommends a two-step approach, including a) a selection of models that would be assessed in more detail with appropriate justification (e.g. high impact models), and b) creation of credibility assessments to derive the final model. Less

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
			impactful models would require less or no justification.
293-405	Beginning with Describe the model and the model development process ……. ……. Ending with Describe the quality assurance and control procedures of computer software (including its 403 toolboxes and packages) and how version changes were tracked.	Please see comments.	Describing model development may not be possible for some open-sourced or commercial models, of which the development process is usually unknown or unclear. It would be helpful to understand in what COUs such models can be employed. Further, as the field matures and more such models become commercially available, it would be helpful to understand how the FDA plans to advise sponsors on which open-source models are or are not meeting credibility standards
303 – 321	i Describe the model Sponsors and other interested parties should include the following information in the credibility assessment plan, as applicable, for each AI model used: An explanation of each model used, including, but not limited to, descriptions of: a) Model inputs and outputs b) Model architecture (e.g., convolutional neural network) c) Model features d) Feature selection process and any loss function(s) used for model design and optimization, as appropriate e) Model Parameters	Consider: An explanation of each model used, including, but not limited to, descriptions of: For Example: Sponsors and other interested parties should include the following information in the credibility 305 assessment plan, as applicable, for each AI model used: An explanation of each model used, including, but not limited to, descriptions of:: a) no change b) no change c) no change d) **Model input pre-processing:** Feature selection process, **Features normalization/standardization**	Including explainability methods implemented to ensure model transparency and interpretability ensures transparency, fosters trust and helps to detect biases and follow a risk-based approach to support validation. ▪ Feature selection process and any loss function(s) used for model design and optimization, as appropriate ▪ Model parameters ▪ **Explainability methods (e.g., SHAP, LIME, or other relevant approaches)** **Point d: The erased sentence is moved to point f** Point e-f: Often the parameters of a trained model are not easily explainable and transparent for NN. The only way is to list the parameters set a priori for the start of the learning process (weight initialization – usually random), define the activation function and bias of neurons and list the NN weights resulting after training. Moreover, they are not so easily

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
	A rationale for choosing the specific modeling approach	e) Model Parameters **(e.g., weights of a neural network (NN) - initialized and after training - and transition probabilities in Markov Model).** **Consider adding f** **f) Any loss function and self-update methods for parameters used for model design and optimization (e.g., backpropagation to update and optimize the weights of NN)** • A rationale for choosing the specific modelling approach	replicable because they auto-update themselves with back-propagation in a way dependent from the process and the data they met. The point of backpropagation is to improve the accuracy of the network and at the same time decrease the error through epochs using optimization techniques, like gradient descent. This is an optimization algorithm used to minimize loss function in the neural network by iteratively moving in the direction of the steepest descent of the function. The values of weights and biases indeed are adjusted and optimized during the training process in a way to minimize the difference between predicted value and the actual value. To know more about backpropagation and parameters update processes, please refers to the link below: https://www.baeldung.com/cs/deep-learning-bias-backpropagation To know more about trainable parameters and their functioning within a NN, please refer to the link below: https://medium.com/@kavita_gupta/what-are-primary-trainable-parameters-of-a-neural-network-6b99f887957c
328	including to define model weights, connections, and components. Tuning data are typically used	Change “including to define model weights, connections, and components” to “**for example, including to define weights, connections, and components in a neural network model**.”	Suggest clarifying these “model weights, connections, and components” are example terms from Neural Network models, while other AI models may have different terms.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
314 – 319	− Model features[27] − Feature selection process and any loss function(s) used for model design and optimization, as appropriate − Model parameters[28]	Please see comments.	There are challenges in the practicability of this guidance for complex AI models, e.g., when considering Large Language Models. Challenges lie in a) technical aspects to provide a meaningful description of models with millions or billions of parameters and b) proprietary knowledge of suppliers who may be unwilling to disclose model details on a parameter level. A scoping definition on the applicability of such expectations would be helpful, or a clarification of suitable expectations in the case of Large Language Models and models of similar complexity.
349-351	Describe (1) the development datasets, including how the development datasets were split into training, tuning, and any additional subsets and (2) the specification of which model development activities were performed using each dataset.	Describe (1) the development datasets, including how the development datasets were split into training, tuning, and any additional subsets and (2) the specification of which model development activities were performed using each dataset. **In this context describe in which percentage validation and test sets are split and the method for dataset splitting according to the specific COU (e.g., random choice of instances or splitting according to a rigorous sub-division – e.g., by patients in clinical context or by batches in manufacturing).**	ISPE recommends that more information is needed to understand the training and evaluation processes: • percentage of splitting for training/validation and test sets • mode of splitting data – randomly or according to a specific rule (e.g. by patients)
353-358	Describe how the development data have been or will be collected, processed, annotated, stored, controlled, and used for training and tuning of the AI model. In addition: • Provide the rationale for choosing the specific development dataset(s)	Describe how the development data have been or will be collected, **(including a detailed description of the data sources)**, processed **(e.g., data filtering, features calculation from inputs, feature normalization/standardization)**,	ISPE suggests that some additional information would be helpful to sponsors. Data sources, and their provenance, are directly related to data quality which impacts model performance, bias, and real-world applicability,

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
	• Explain how labels or annotations were established.	annotated, stored, controlled, and used for training and tuning of the AI model. In addition: • Provide the rationale for choosing the specific development dataset(s) **and features** • Explain how labels or annotations were established.	which ISPE perceives are critical factors in the risk-based Credibility Assessment Framework.
362 - 365	Explain how the development data is relevant (e.g., includes key data elements and sufficient number of representative participants or sufficient data that is representative of the manufacturing process or operation) and reliable (i.e., accurate, complete, and traceable).	Explain how the development data is relevant (e.g., includes key data elements and sufficient number of representative participants or sufficient data that is representative of the manufacturing process or operation) and reliable (i.e., accurate, complete, and traceable). **Furthermore, describe the methods used to contextualise the data aligned with COU (e.g., metadata strategies, techniques used to enhance interpretability).**	It is ISPE's opinion that documenting how data is collected and contextualized and how this is aligned with the COU helps SMEs and regulators to understand any variations in the data and how they might affect the AI model's performance and trustworthiness. Also clearly describing contextualization enhances traceability, transparency and auditability.
377-379	Describe how the model was trained including: a) Learning methodology (e.g., supervised, unsupervised)	Describe how the model was trained, including: a) Learning methodology (e.g., supervised, unsupervised, **reinforcement learning)**	Reinforcement Learning is an additional type of Learning Methodology To better understand the differences of Model Learning types cited, ISPE recommends referring to the link below: https://www.geeksforgeeks.org/supervised-vs-reinforcement-vs-unsupervised/

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
			To focus more on Reinforcement Learning functioning, please refer to: https://www.ibm.com/think/topics/reinforcement-learning
381-386	b) Performance metrics used to evaluate the model, such as the area under the receiver operating characteristic (ROC) curve, recall or sensitivity, specificity, positive/negative predictive values (PPV/NPV), true/false positive and true/false negative counts (e.g., in a confusion matrix), positive/negative diagnostic likelihood ratios (PLR/NLR), precision, and/or F1 scores. All performance estimates should be provided with confidence intervals.	Performance metrics used to evaluate the model, such as the area under the receiver operating characteristic (ROC) curve, accuracy, recall or sensitivity, specificity, positive/negative predictive values (PPV/NPV), true/false positive and true/false negative counts (e.g., in a confusion matrix), positive/negative diagnostic likelihood ratios (PLR/NLR) precision, and/or F1 **scores for classification models. Also, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE) for regression models.** All performance estimates should be provided with confidence intervals, **where possible and justified by model risk.**	ISPE suggests considering adding because mean absolute error (MAE), Mean squared error (MSE) or root mean square error (RMSE) are commonly used for regression problems when evaluating the Models' results (e.g., after cross validation - dataset folding). The metrics highlighted in the paragraph were referred to classification only. ISPE has included links that may be useful to understand the regression meaning and performance metrics (along with their mathematical computation) commonly used for regression problems below: https://machinelearningmastery.com/regression-metrics-for-machine-learning/ https://www.geeksforgeeks.org/regression-metrics/ Considering the increased rigor and resource demands for data analysis and reporting, ISPE suggests this requirement should be commensurate with the model's risk level and/or direct/indirect/immediate patient impact
395-396	If a pre-trained model was used, specify the dataset that was used for pre-training and how the pre-trained model was developed and/or obtained.	Please see comments.	For pre-trained models in commercial software or pre-trained LLM models, the dataset being used and how the models being trained may not be 100% clear or available to the sponsor. ISPE suggests that the final guidance should clarify that these details should be specified as much as possible, and rationale should be provided to explain the limitation.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
			Third-party vendors selling AI solutions may not wish to share proprietary information in relation to how the AI model was trained, which dataset was used for the pre-training and how the pre-trained model was developed and/or obtained. ISPE recommends including specific expectations for sponsors with respect to obtaining such information and the legal basis to support this. Furthermore, ISPE recommends guiding how a third-party vendor may directly share relevant information with the agency if they are unable to share with the sponsor
398	Describe the use of ensemble methods	Consider adding to line 398: Describe the use of ensemble methods**, as applicable. List especially the category (e.g., sequential or parallel), type (e.g., bagging, boosting, stacking), subcategories (bootstrapping or aggregation) and explain the functioning.**	Ensemble methods are not so often used, but if they are, they should be described in detail in terms of type and functioning To know more about ensemble methods categories, please refer to the link below: https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/
418-420	Describe how the test data have been or will be collected, processed, annotated, stored, controlled, and used for evaluating the AI model.	Please see comments	ISPE recommends including additional detail on models' evaluation: cross-validation technique to give confidence and trust in the use of the model. Consider explaining the cross-validation methods used to evaluate the model's performances (e.g., hold out, leave one out, K-fold cross validation, etc.). In the case of cross-validation use, explain in detail the modalities of the iterative evaluation on different combinations of training and test sets.
436-440	Describe the applicability of the test data to the COU. This issue is important because, for example, when prediction models are developed using historical development data, the AI model	Describe the applicability of the test data to the COU. This issue is important because, for example, when prediction models are developed using historical	ISPE recommends including additional notes to make explicit the need for control and preventive actions to avoid *data drift.*

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
	may not perform as well in the COU if the development data are different from the data encountered in the deployed environment used in the COU. This phenomenon is sometimes referred to as *data drift*.	development data, the AI model may not perform as well in the COU if the development data are different from the data encountered in the deployed environment used in the COU. This phenomenon is sometimes referred to as *data drift*. **If *data drift* is supposed to happen, provide a plan to make a re-learning process on the model or describe the control and preventive actions performed by humans or by the algorithms implemented to avoid the issue.**	
445-449	Provide the rationale for the chosen model evaluation method(s) and explain the applicability of the evaluation methods to the modeling method used and to the COU. If the COU involves a "human in the loop," ensure that the evaluation methods consider the performance of the human-AI team, rather than just the performance of the model in isolation.	Please see comments.	ISPE suggests the inclusion of the following: An explicit definition of "human in the loop" could be very helpful to understand the concepts and when applicable. ISPE's current understanding of 'human in the loop' should be addressed when, for example: - Humans are involved in the construction of labels (e.g., annotations) for building model targets -Humans are involved in the model control - Humans are the operators or end-users of a device, and the performances of the models also depend on the interaction between them In any of the cited cases, it is difficult to give an estimation of the influence of the humans within the process (at least numerically).

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
			A reference on quantitative evaluation methods commonly used to evaluate models' performances could provide useful insights to the reader.
451-457	Describe the performance metrics used to evaluate the model, such as the area under the receiver operating characteristic (ROC) curve, recall or sensitivity, specificity, positive/negative predictive values (PPV/NPV), true/false positive and true/false negative counts (e.g., in a confusion matrix), positive/negative diagnostic likelihood ratios (PLR/NLR), precision, and/or F1 scores, including the optimization methods used (e.g., use of a gradient descent). All performance estimates should be provided with confidence intervals. In addition:	Please see comments.	Please consider that the same performance metrics should be used from model development to model evaluation. It might be helpful to add acceptance criteria for model evaluation, as applicable For example, if the F1 score is used for model development, the F1 score should be used for model evaluation, not changing to another metric such as precision or recall. The performance metrics on the test dataset should pass the acceptance criteria before implementation. Consider that this addition could limit new applications, so we propose to further analyse it and add context/assumptions under which it should be applicable in order to assure expected outcomes without limiting future applications.
456-457	All performance estimates should be provided with confidence intervals.	All performance estimates should be provided with confidence intervals, **if applicable.**	Confidence intervals for performance estimates should be provided if applicable. In some use cases, the metrics can be deterministic, with no repetition/variability or lack of a proper statistical framework. In these cases, confidence intervals may not be clear, meaningful, or reliable. For instance, when we test the performance of an LLM in the task of answering real-world questions, we usually utilize a benchmark dataset with pairs of questions and answers and check whether the outputs of the LLM match the answers in the benchmark dataset. The performance can be evaluated using metrics like the proportion of

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
			outputs matching the benchmark, but the meaning of a confidence interval of the proportion is unclear. ISPE recommends the agency to note the limitations of confidence intervals in such cases.
515-517	Life cycle maintenance of AI models is a set of planned activities to monitor and ensure the model's performance and its suitability throughout its life cycle for the COU.	Please see comments.	ISPE recommends adding some text to section A Step 4 (Establish AI Credibility Plan) to detail in the credibility plan how the ongoing performance of the model will be monitored. Would also suggest that the final guidance refer to relevant guidance issued by CDRH that drug and biologics developers could use.
519 - 526	As mentioned in section III, life cycle maintenance of the credibility of AI model outputs is important because a model's performance can change over time or across deployment environments. While the use of AI to support regulatory decision-making for drugs is typically assessed on locked data and information produced by an AI model at a given point in time, there are instances where the use of AI models extends over the drug product life cycle, and life cycle maintenance of the credibility of AI model outputs is critical. For example, life cycle maintenance of the credibility of AI model outputs is important for the application of AI modeling in the pharmaceutical manufacturing phase of the drug product life cycle.	ISPE recommends some clarifying additional text after the word "critical." As mentioned in section III, life cycle maintenance of the credibility of AI model outputs is important because a model's performance can change over time or across deployment environments. While the use of AI to support regulatory decision-making for drugs is typically assessed on locked data and information produced by an AI model at a given point in time, there are instances where the use of AI models extends over the drug product life cycle, and life cycle maintenance of the credibility of AI model outputs is critical. **Ensuring credibility not only relies on continued model monitoring but also on ensuring that AI-generated outputs are based on data that is coherent**	As mentioned in section III, life cycle maintenance of the credibility of AI model outputs is important because a model's performance can change over time or across deployment environments. While the use of AI to support regulatory decision-making for drugs is typically assessed on locked data and information produced by an AI model at a given point in time, there are instances where the use of AI models extends over the drug product life cycle, and life cycle maintenance of the credibility of AI model outputs is critical.

Section or Line Number	Current Text	Proposed Change	Rationale or Comment
		with the model's expectations throughout the life cycle. For example, life cycle maintenance of the credibility of AI model outputs is important for the application of AI modelling in the pharmaceutical manufacturing phase of the drug product life cycle.	
529-536	…because they are data-driven and can be self-evolving (i.e., capable of autonomously adapting without any human intervention). Model performance metrics should be monitored on an ongoing basis to ensure that the model remains fit for use and appropriate changes are made to the model, as needed. The level of oversight for a model over its life cycle should be risk-based (i.e. commensurate with the model risk and COU). Due to the evolving nature of AI models, sponsors should anticipate inherent, model-directed changes,…"	because they are data-driven and can be self-evolving (i.e., capable of autonomously adapting without any human intervention). Model performance metrics should be monitored on an ongoing basis to ensure that the model remains fit for use and appropriate changes are made to the model, as needed. The level of oversight for a model over its life cycle should be risk-based (i.e. commensurate with the model risk and COU). Due to the evolving nature of AI models, sponsors should anticipate inherent, model-directed changes **and risks…"**	ISPE notes that CDRH has issued guidance on the Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions and would request the final guidance to refer to appropriate sections or indeed the entire guidance applies to developers of drugs and biologics. This seems to be the only section in the document referring to "autonomy" or "autonomous updating" of models. ISPE assumes that risks form use of autonomy should be considered during model development and in change management. For the life cycle management of such potentially self-updating models, ISPE suggests that additional directional guidance be added for change control approaches.

End of comments