August 3, 2023

Dockets Management Staff (HFA-305)
Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852

Re: Docket No. FDA-2023-N-0743-0002 Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products

Dear Sir/Madam:

The Biotechnology Innovation Organization (BIO) thanks the Food and Drug Administration (FDA) for the opportunity to submit comments regarding the request for information and comments on the Agency's discussion paper entitled "Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products."

BIO is the world's largest trade association representing biotechnology companies, academic institutions, state biotechnology centers and related organizations across the United States and in more than 30 other nations. BIO's members develop medical products and technologies to treat patients afflicted with serious diseases, to delay the onset of these diseases, or to prevent them in the first place.

Sincerely,

/s/
Sam Gunter
Director, Science & Regulatory Affairs
Biotechnology Innovation Organization

General Comments/Questions:

BIO applauds FDA for the issuance of this comprehensive and clear discussion paper on the use of AI/ML in drug development and supports FDA's risk-based approach to the review, assessment, and regulation. We also welcome the Agency's decision to build on existing guidance and regulatory tools with respect to the use of AI in drug development, including CDRH guidance and the 2021 Good Machine Learning Practice for Medical Device Development Guiding Principles.[1] As the Agency continues to gain experience in this space, BIO notes that this discussion document is a good first step and more insights on different algorithms or statistical methods supporting various aspects of drug development, such as signal detection and risk assessment would be useful.

As a general matter, BIO supports the National Institute of Standards and Technology in this space. AI/ML is a cross-sector topic, as such the overarching high-level principles and standards should apply to all sectors, including AI/ML use in drug development. All parts of the federal government should be aligned in this space. With AI/ML being a dynamic and nascent field, BIO supports an incremental approach to guidance development that would allow for sufficient learning opportunities from both sponsors and FDA.

We encourage FDA to support further stakeholder conversation on the broad concept of good data science practice, which is focused on bringing the best tools, skills, and relevant data to answer a well scoped research question. AI/ ML approaches might be considered for many uses in drug development but require multidisciplinary consideration to ensure they are fit-for-purpose for different uses. We look forward to further discussions with the agency on the appropriate use and quality control (QC) in drug development. Given the increased awareness and use of AI/ML applications in many sectors, BIO also recommends that the Agency reference the White House Blueprint for an AI Bill of Rights and consider providing additional context on how the Agency is engaged in these broader efforts and how they will impact the Agency.

Specific Questions/Comments:

1. **Human-led governance, accountability, and transparency**

a. *In what specific use cases or applications of AI/ML in drug development are there the greatest need for additional regulatory clarity?*

In general, BIO notes that there is a need for greater regulatory clarity regarding the use of AI/ML across the entire drug development continuum from the use of nonclinical models to support target identification, to the design of clinical trials, to the analysis of clinical trial data to postmarket surveillance. Broadly speaking, BIO believes that the greatest need for regulatory clarity would be around the application of AI/ML models that directly impact the benefit-risk assessment or efficacy assessment of a medical product. We also believe that stakeholders would benefit from continued discussion regarding best practices for validation of AI/ML algorithms that are fit for different uses across drug development. BIO provides more specific suggestions below:

[1] FDA, Health Canada, UK MHRA, "Good Machine Learning Practice for Medical Device Development: Guiding Principles," October 2021, https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

Nonclinical Studies

Regulatory clarity is needed on the qualification, validation, appropriate use, and documentation of AI/ML tools in Good Laboratory Practice (GLP)-regulated nonclinical studies supporting dosing of humans in clinical trials. If the FDA adopts a risk-based regulatory framework, it is important for FDA to establish a clear criterion for assessing that risk. Additional clarity is also needed on the characteristics to be considered in assessing risk such as the probability of occurrence, the distinction between systematic and local risks, and the timeframe.

Clinical Trial Design

Use cases of AI/ML in clinical development have the most pressing need for regulatory clarity as they potentially have the highest impact on the benefit-risk assessment of a drug or biological product. Specific uses cases requiring greater clarity include AI/ML models for patient identification, selection, and stratification as well as endpoint selection and evaluation. Clarification is also needed on how risk assessments will vary across the different phases of clinical development (e.g., how might requirements differ for Phase 1 exploratory studies vs. Phase 3 registrational studies). More specific to efficacy assessments, additional guidance is needed on the type of evidence FDA requires to show that an endpoint is valid in a future study if the AI/ML is trained on legacy trial data. Further discussion would be welcomed regarding potential need for labelling if a sponsor uses an AI/ML model to assess efficacy.

Clinical Trial Conduct

BIO suggests FDA provide additional information on the totality of considerations on the use of digital health technologies that rely on AI/ML components for evaluating safety and efficacy.

Data Analysis

BIO notes that when integrating diverse datasets, trade-offs are typically made between explainability and performance. Sponsors would benefit from further guidance on how to optimize this trade off. Guidance is also needed on the use of “wide data” in predictive models where the number of features/parameters tend to be more than the number of subjects. This includes meta-analysis from fusion of multi-modal data from electronic medical records, claims data, trial data, and PROs in AI/ML models.

Regulatory Decision-Making

More clarity would be welcomed on the level of transparency and documentation required for applications in which AI/ML are key components of a registrational study or drug substance/product quality data, for example, specific discussion on the circumstances that AI/ML analysis can replace or complement statistical or predictive approaches.

Postmarket Surveillance

It would be helpful to have additional regulatory clarity on the extent AI/ML can be applied for decision support as opposed to decision-making for post-market surveillance, specifically related to predictive safety signal detection. It would also be helpful to have additional regulatory clarity on the performance metrics (e.g. AUC, accuracy) for AI/ML used in decision support and decision-making for signal detection. Additional guidance on how to manage post-market surveillance efforts (e.g., when to treat as a modification versus issuance of medical device report) would also be appreciated. AI/ML is currently used to assess known risks in labels and new risks that arise once the products are marketed. As discussed in several publications, AI/ML has been used to support signal detection.[2]

[2] Colilla et al, Drug Safety 2017; Kurzinger et al, JMIR 2018; Gavrielov – Yusim at al, PDS, 2019)

Lifecycle Management
BIO suggests the Agency develop a robust model of lifecycle management for AI/ML models. This would involve defining metrics for assessing the reliability of a model's bias and variance and ensuring consistent and unbiased performance over a defined period. This approach would also be bolstered by guidance on addressing data heterogeneity (e.g., due to variations in equipment and raw materials), considerations for transfer learning, the validation and maintenance of models, and methods for establishing conformance as more data is collected over time.

Application of Medical Device Regulations
Clarity is needed on how the medical device regulatory framework, specifically SaMD regulations, apply to AI/ML applications. Specific examples of the different types of AI/ML used in clinical research and whether they are classified as medical devices would be most beneficial. For example, if an AI/ML application is used in a clinical trial for a drug or biological product development but is not intended to be commercialized and does not impact patient treatment or management, BIO seeks to understand which specific device regulations should be followed. The verification and validation requirements for these different use cases would also be useful. Similarly, BIO seeks to understand when an algorithm used as a diagnostic would be considered a companion diagnostic.

b. *What does transparency mean in the use of AI/ML in drug development (for example, transparency could be considered as the degree to which appropriate information about the AI/ML model—including its use, development, performance, and, when available, logic—is clearly communicated to regulators and/or other stakeholders)?*

Transparency, in the context of AI/ML in drug development, refers to the degree to which relevant information about the AI/ML model is clearly communicated to regulators, healthcare professionals, patients, and other relevant stakeholders. Some key aspects of transparency in AI/ML in drug development include model architecture and development, data sources and quality, model performance and validation, ethical considerations and safeguards. BIO notes that different levels of transparency are needed for different stakeholders and for different use cases. For example, a higher level of transparency may be required for FDA, while a different level of transparency may be appropriate for patients. To support transparency between sponsor and regulator, BIO believes a comprehensive set of documentation should be communicated to the regulators to serve as an audit trail through the AI/ML lifecycle. This documentation should include:

- the application's intended use scenario, such as the targeted user population,
- details regarding the training data, including its source, version, and processing history, along with the persistence of intermediate data artifacts such as pre- and post-cleaning datasets;
- information on the infrastructure employed, encompassing both hardware and software, such as code libraries,
- insights into the model learning processes, such as ranges for hyperparameter search, seeding strategy, and any fine-tuning methods employed when leveraging foundational models and transfer learning approaches,
- validation steps undertaken, especially if multiple validation methods are explored (although not necessarily encouraged), along with their outcomes,
- methodologies employed for model deployment,

- a log documenting real-world application performance with predefined model metrics
- a planned schedule and approach for model assessment and retraining.

Conversely, to support transparent communications from the regulator to sponsor, BIO suggests FDA and industry work in concert to develop a standard set of performance metrics for different uses cases. For example, for AI/ML in pharmacovigilance activities such as signal detection, Area Under the Curve (AUC) and accuracy may be important performance metrics. For binary classifiers, however, performance metrics could include Positive Predictive Value (PPV) and Negative Predictive Value (NPV), while AUC, accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) are less relevant.

To develop a transparent, open, and collaborative ecosystem regarding AI/ML, BIO encourages FDA to continue to inform sponsors and stakeholders on existing FDA AI/ML-related projects and initiatives. Last, BIO also recommends FDA develop a decision tree or quick start guide specifying which center or team within FDA (e.g., Digital Health Center of Excellence, CDRH, etc) to contact with which questions.

c. *In your experience, what are the main barriers and facilitators of transparency with AI/ML used during the drug development process (and in what context)?*

One of the primary barriers to transparency between sponsors and regulators regarding AI/ML models is the “black box” concept. The “black box” concept refers to an AI/ML models where even its designers cannot explain why that model arrived at a specific decision. Many deep learning models use a black box strategy. While black box models are appropriate in some circumstances, they can potentially introduce unacceptable risks in specific situations. For that reason, BIO encourages FDA to foster discussion on when such approaches may be acceptable within the context of drug development. In addition, BIO supports FDA’s statement that in cases where models have similar performance, preference should be for explainable models, especially in circumstances that impact the benefit-risk ratio.

The proprietary nature of many algorithms can sometimes represent a barrier to transparency. This issue is compounded by the fact that sponsors often work with third parties during the development phase who consider their underlying code and training data a trade secret. Clear expectations from FDA regarding the expected level of transparency required to support regulatory decisions would increase stakeholder understanding and facilitate collaborations to advance the use of AI/ML in drug development.

BIO encourages FDA to define an acceptable level of qualification and validation for algorithms and datasets based on the context of use and regulatory classification. For AI/ML models developed to match human performance, FDA should define the minimal congruence that must be achieved. While for AI/ML models where no human expert measure equivalent is available, the minimal sample size to validate the model should be defined. Lastly, clear expectations from FDA regarding the expected level of transparency required to support regulatory decisions would also increase stakeholder understanding and advance the use of AI/ML in drug development.

d. *What are some of the good practices utilized by stakeholders for providing risk-based, meaningful human involvement when AI/ML is being utilized in drug development?*

The application of AI/ML to drug development is a multi-disciplinary project, requiring the involvement of a wide range of skills and expertise. Thus, BIO recommends that efforts to use

AI/ML in drug development be multi-disciplinary, involving individuals with expertise in medicine, epidemiology, computer science, data science, risk management, regulatory compliance, and ethics, a principle reflected in the CDRH good machine learning principles.

In considering when to use an AI/ML application in drug development, stakeholders have used AI/ML as a screening tool prior to involving human experts, especially in situations of high processing volumes and sparse expert resources. Such uses for extracting or organizing information essentially augment human expertise, freeing up experts to focus on more complex, high impact tasks. Examples include the use of AI/ML for screening high-throughput toxicology imaging, product quality inspection, safety monitoring, and more. Decision thresholds for the preliminary AI/ML step can be set to conservative values, while still enabling large efficiency gains by reducing the volume which is to be processed by humans and retaining human decision power for any cases that are not self-evident.

We also highlight the below specific use cases of incorporating meaningful human involvement with AI/ML applications:

Nonclinical / Early Phase Studies
While AI/ML are frequently used to prioritize molecules in compound screening, nonclinical research, and the lead optimization phase, meaningful human involvement is carried out by actively assessing model applicability, prospective performance, and model updating when data drifting is observed. Many of these activities are carried out in an iterative learning cycle with extensive human involvement.

Lifecycle Management
Many stakeholders employ a combination of automated validation and manual review of a model's performance, assessing its, robustness, and potential biases. This involves human experts examining the outputs and evaluating their alignment with expected results. Manual review allows for human judgment and intervention, providing an opportunity to identify and address potential issues that may not be captured by automated processes. Complementing automated validation with manual review strikes a nice balance between leveraging the power of AI/ML and maintaining human oversight.

e. What processes are in place to enhance and enable traceability and auditability?

To ensure traceability and auditability, sponsors must adopt good practice in data documentation – specifically documenting data curation methods, algorithm annotation, model version control, and model performance reporting. Additional recommended processes to improve traceability and auditability include the adoption of an AI quality management system, establishing and defining criteria for AI Transparency and security, the development and explanation of rules for AI/ML supervised or unsupervised learning, establishing criteria for the prevention of bias, standards for human testing and training, and standards for the automated validation of AI/ML model drifts.

Member companies currently have in place a number of processes and employ a number of programs to ensure auditability and traceability including:

- Use of a GxP compliant cloud environment;
- ML operations platforms that manage the entire lifecycle of AI/ML development and enable reproducibility;

f. How are pre-specification activities managed, and changes captured and monitored, to ensure the safe and effective use of AI/ML in drug development?

In managing pre-specification activities, BIO suggests all decisions, assumptions, and justifications made in the development of the AI/ML model be documented without disclosing trade secrets or proprietary information. This documentation should serve as a reference for the entire model development process, ensuring transparency and reproducibility.

When developing a reference document pre-specifying the model development process, BIO recommends the documents address the following questions:

- Define the research question or problem. What is the machine learning model trying to do? What are the objectives and intended outcomes of the analysis?
- Specify the data collection process. Where will the data come from? What variables will be collected? How will the data be cleaned and processed?
- Determine the appropriate machine learning algorithm or model architecture. What type of model is best suited for the research question and data characteristics? What are the model parameters and hyperparameters?
- Define the metrics or criteria to evaluate the model's performance. How will the model's performance be measured? What are the primary and secondary metrics of interest?
- Specify approaches for handling missing data and outliers.
- Specify the method for cross-validation. How will the data be divided into training, validation, and test sets? How will imbalanced data or model validation be handled?
- Define the statistical analysis plan. What statistical tests will be used? What is the significance level? Will multiple comparisons be adjusted for? What other analyses or sensitivity analyses will be conducted?
- Specify the software packages or tools to be used. What software packages or tools will be used for data analysis, model development, optimization, hyperparameter tuning and evaluation? Document the versions or configurations of these tools for reproducibility purposes.
- Identify and address ethical considerations. What ethical considerations are related to data privacy, bias, fairness, or potential unintended consequences of the model's application?
- How to factor in changes to the data distribution over time that are consumed by the models for updating and/or predictions?
- Clearly document amendments to specifications.

2. Quality, reliability, and representativeness of data

When considering the use of AI/ML in the drug development process, there are several data considerations that need to be considered. These considerations help ensure the reliability, validity, and ethical use of AI/ML models. Here are some key data considerations:

- Ensure that the data used for training and validation is of high quality and integrity
- Ensure that the training data used for AI/ML models is representative of the target population or clinical scenario
- Protect patient privacy and ensure compliance with relevant data protection regulations

- Consider opportunities for data sharing and collaboration to enhance the development and validation of AI/ML models
- Consider the inclusion of real-world data, such as electronic health records or claims data, to complement clinical trial data in the training and validation of AI/ML models
- Account for the longitudinal nature of patient data and the temporal dynamics of disease progression or treatment effects
- Be aware of potential biases in the data that can lead to unfair or discriminatory outcomes
- Provide clarity on data transformation and harmonization rules, if applicable
- Establish data governance frameworks that outline ethical considerations, data access policies, and guidelines for data use in AI/ML models

a. What additional data considerations exist for AI/ML in the drug development process?

BIO notes that many of the data quality, relevancy, bias, and reliability considerations included in FDA's existing guidances are foundational and apply to the use of AI/ML in drug development. Nonetheless, we highlight the below data considerations associated with AI/ML in drug development.

Data Flow

The flow of data from device to device involves multiple partners, requiring clearly articulated roles and responsibilities with respect to hand-offs. There also remains a lack of clarity on how a data flow system is defined. Guidance on how end-to-end testing should be performed is also needed.

Data Availability

The success of applying AI/ML to drug development relies upon the availability and accessibility of large databases to train systems. Thus, access to government-sponsored databases will enable rapid progress from proof-of-concept prototypes to real-world technology thereby accelerating biomedical research. BIO supports the establishment of public-private partnerships and other collaboratives to would advance the creation and sharing of machine-readable data sets used for drug development. Given the global drug development ecosystem, BIO also recommends that the Agency consider highlighting opportunities to include data from outside of the US in existing or future collaborative efforts.

Data Quality

Data quality attributes should include the presence of rich metadata or other technical aspects that may affect model performance. This metadata will help make models more generalizable and will avoid issues due to differences in data generation with respect to the training set(s).

Data Size

In terms of the datasets themselves, datasets need to be large enough to capture the complexity and variability in the drug development process. Moreover, datasets that are too small also risk the re-identification of patients. Special consideration needs to be given to specific cases utilizing small data, where the potential for biased prediction or inference is much higher. Similarly, data collected as a sample of a particular population versus the total population is also an important consideration.

Data Type

Longitudinal or time series data creates special issues for AI/ML as the models are currently not naturally adapted to this type of data.

Data Transformation
The process for transforming and transmitting data to constitute meaningful information manufacturers use to run their operations must be demonstrated.

Data Heterogeneity
Use of heterogeneous data sources or combining of different data sources within AI/ML solutions needs special consideration as it is often difficult to understand all the compatibility issues in using different data sources.
Bias
Data coverage of all relevant variables or features is particularly important as data that does not contain all relevant variables or features can lead to biases or inaccurate outcomes that may not be fully interpretable or generalizable.

Imbalance
Imbalances within the dataset, such as variations in the distribution of classes or rare events, can impact the performance of AI/ML models. Addressing data imbalance requires employing techniques such as oversampling, under sampling, or synthetic data generation to ensure fair representation of all classes. Mitigating data imbalance is crucial to prevent biased predictions and ensure robust model performance.

b. What practices are developers, manufacturers, and other stakeholders currently utilizing to help assure the integrity of AI/ML or to address issues, such as bias, missing data, and other data quality considerations, for the use of AI/ML in drug development?

BIO member companies suggest FDA, industry, and stakeholders develop and adopt a consensus-based framework to assure integrity of AI/ML in drug development, similar to the Digital Medicine Society's V3 Framework[3] for determining whether digital health technologies are fit-for-purpose. Some current frameworks and methods to consider include the use of FAIR (findable, accessible, interoperable, and reusable) data during clinical studies.

In addition to the above, BIO also highlights the below practices being used by member companies:

- **Continuous Feedback Loops**: Once deployed, every ML model should be periodically tested and data that could be leading to bias should be removed from the model. This continuous feedback loop can help assure the integrity of the model over time and increase transparency. This practice can also help address the inadequate representation of certain populations in health care or drug development datasets, which could lead to bias if used to train the algorithm.
- **Cross Validation**: Cross validation is a common practice to assess and mitigate bias in AI/ML models. By splitting the dataset into multiple subsets and iteratively training and evaluating the model on different partitions, cross validation helps detect and address biases and overfit that may arise due to imbalanced or limited data, or bad choice of training algorithm.

[3] Digital Medicine Society (DiME), V3 Framework, April 2020, https://dimesociety.org/access-resources/v3/.

- **Model Monitoring and Retraining**: Continuous monitoring of AI/ML models to detect and address issues related to bias, missing data, or evolving patterns. Regularly updating and retraining models using new data helps maintain model integrity and adapt to changing conditions, improving overall performance, and addressing data quality concerns. This of course implies utilization of adaptive models which requires clear guidance on validation and suitability.
- **Exception Handling for Input Data**: To incorporate exception handling mechanisms to handle issues with input data, such as missing or inconsistent data. Exception handling helps maintain the integrity of AI/ML models by mitigating the impact of data quality issues on model performance.
- **Data Preprocessing and Cleaning**: Rigorous data preprocessing and cleaning techniques as part of the highly recommended "Data-Centric" approach in machine learning. This includes removing outliers, addressing missing values through imputation, handling data inconsistencies, ensuring data uniformity, and most importantly custom data transformations to avoid the need to absorb complexity through large model trainings. Thorough data preprocessing helps improve the quality and reliability of the input data, reducing the potential for biases and enhancing model performance.
- **Prespecified Approach to Assess Model Generalizability**: Building quality assurance into dataset generation, for example, in a human labeled dataset, labels are generated by two separate annotators, discrepancies are resolved via a third arbitrator. AI/ML for data adjudication and promoting standardization of clinical interpretation.

Missing data and bias remain a continual challenge for BIO member companies. To address bias, member companies collate data from a myriad number of sources in order to handle bias that may occur at the data collection side from one vendor. As diversity is often less represented in historical clinical trial data, sponsors augment model output by providing downstream teams contextual information from the diversity point of view. For testing sponsors have used numerous techniques including cross-fold validation, bootstrapping techniques, and country-level analysis.

c. *What are some of the key practices utilized by stakeholders to help ensure data privacy and security?*

The need to have in-depth knowledge of data privacy laws in the jurisdictions where AI/ML might be applied is critical. Ensuring that the proper assessments, such as Data Protection Impact Assessments (DPIAs) and Transfer Impact Assessments (TIAs) are conducted is critical to understand the use of the AI/ML and its impact subjects. Transparency to data subjects on how AI will process data and for what purpose is required under some laws and requires an in depth understanding of where consent needs to be obtained and maintained to ensure that data subject rights are exercisable. The legal basis needs to be considered as well in order to ensure that data being used has been obtained and is processed in a lawful manner and proper notice has been provided to the individuals. If anonymization is required to process data in this matter, the anonymization of the data needs to be maintained and the risk of re-identification of the data needs to be assessed for every situation. Unfortunately, data protection laws are not consistent, which makes it a challenge to implement a clear framework addressing the use of AI/ML.

Practices that are being employed by stakeholders to ensure data privacy and security include:

- Engaging in and upholding data usage agreements with data providers and collaborators.
- Implementing training and education campaigns to promote responsible and secure usage of relevant systems.
- Continuously monitoring the data to ensure it is properly stored in appropriate systems without incident.
- Using potent encryption keys for database access, and other protective measures to ensure data security.
- Using data loss prevention capabilities to control sensitive information and prevent it from leaving the drug developer's environment, when using a public model hosted by outside providers.
- Data-at-rest and data-in-transit encryption is used commensurate with how the data are classified for privacy and security purposes.
- Hosting algorithms and models within the corporate infrastructure of the drug developer to reduce privacy and data loss risks.
- Ensuring adequate procedures and processes are in place to comply with applicable data privacy laws and regulations.
- Using secure third-party servers that anonymize data and build in blinding and unblinding procedures.
- Sharing "fingerprints" of molecules in early phase development.

d. What are some of the key practices utilized by stakeholders to help address issues of reproducibility and replicability?

To ensure reproducibility and replicability, transparent access to comprehensive documentation, including the audit trail of the workflow and parameter selection, is crucial throughout the entire model development and validation processes. This documentation should also extend to the training and derivative data sets, enabling thorough examination and verification.

The population on which the model is applied in production can evolve away from the training data over time, which can lead to model drift; this requires risk-based monitoring of the function loss and performance metrics for the AI model.

Depending on the Intended use and risk profile of the AI application, an appropriate involvement by human beings (Human Oversight) must be ensured (e.g., "human in the loop" technical or procedural risk mitigations). The performance of the Human-AI team must be assessed to ensure they can perform meaningful oversight of the AI (i.e. they are an appropriate Subject Matter Expert on the business process.)

Options for Human in the Loop can take various forms, e.g. AI provides proposed output and human must review/edit/approve output, AI is autonomous but human can intervene, AI is fully autonomous, (this is not appropriate for high risk AI applications).

In addition, BIO member companies have also utilized the below practices to ensure reproducibility and replicability.

- **Documenting Code and Dependencies**: Thorough documentation of code, including all necessary dependencies and libraries, to provide clear instructions on how to set up and run the code. Documenting the code enables other developers and stakeholders to understand the implementation details, reproduce the results, and build upon the work effectively. Promote code sharing and utilization of R Markdown and Jupyter notebooks.

- **Use of Containers**: Utilizing containerization technologies, such as Docker, to encapsulate the entire software environment required to reproduce and replicate AI/ML experiments. Containers provide a consistent and isolated runtime environment that includes the necessary software dependencies, libraries, and configurations. By sharing the containerized environment, AI/ML developers and regulators can ensure that others can replicate their experiments precisely, regardless of differences in underlying system setup.
- **Solid Data Management**: Robust data management practices contribute to reproducibility and replicability. Documenting information about data collection, preprocessing steps, data augmentations, and any data transformations applied. This includes specifying details like data sources, data collection protocols, preprocessing scripts, and data augmentation techniques used.
- **Modular Programming**: Emphasize on modular programming approaches to enhance reusability of code and models. By breaking down complex AI/ML systems into modular components, AI/ML developers can create building blocks that can be easily reused and integrated into different projects. Modular programming also facilitates code maintenance, collaboration, and version control.
- **Boundary Conditions**: Stakeholders diligently test and document the boundary conditions, as well as assess the sensitivity of the assumptions made.

e. *What processes are developers using for bias identification and management?*

Developers can reduce the risk of bias by training models on diverse data sets and leveraging real world data sets to estimate the expected distributions for accurate representation. Prior to launching developers should employ algorithmic bias mitigation techniques during the training and model selection process, such as recursive feature engineering, cross-validation and regularization techniques. Similarly, rules should be established for how ML models will be monitored to detect whether precision and accuracy drift over time and what conditions will trigger the need for re-tuning the model. Once a model is deployed developers use multi-dimensional approach monitoring several factors - input data, model performance, and application performance - as well as different methods (e.g., direct/indirect, manual/automatic). Post launch developers also continually compare the model's prediction across different demographic groups and use metrics such as Demographic Parity and Equalized Odds to assess the model's fairness.[4]

3. Model development, performance, monitoring, and validation

a. *What are some examples of current tools, processes, approaches, and best practices being used by stakeholders for:*

1. *Documenting the development and performance of AI/ML models that can be applied in the context of drug development (e.g., CONSORT-AI (Liu et al., 2020) and SPIRIT-AI (Cruz Rivera et al., 2020))?*

[4] Fairlearn, "Common Fairness Metrics," https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html - :~:text=While demographic parity assesses the,by the positive target variable)..

To measure model efficiency, stakeholders routinely monitor metrics like accuracy, precision, recall and other metrics suitable for a particular suite of AI/ML models. Model metrics are often derived based on a business use case, problem definition and outcomes like its regression-based or classification-based problem.

Performance metrics like r-squared (R2) and root mean square error (RMSE) are tracked for AI/ML models, and few business driven metrics can be tracked from model to model using tools like MLFlow, which allows for the capability of tracking performance over time. In the context of signal detection, the performance (sensitivity, specificity, accuracy) of AI/ML could be measured against the known risks as shown in the labels and new risks observed when the products are in the market. The existing drug development Quality Management System should address or include the requirements for establishing the essential elements of a quality framework for Machine Learning models through the model lifecycle to ensure continued suitability and effectiveness. The framework can include three primary components required for effective model life-cycle management:

- The management of the people who perform the lifecycle activities;
- The procedures or company standards which articulate what will be done and how to do it in accordance with the framework;
- The methodology used to perform the activities.

To enhance the drug discovery and development process, there is a requirement for an expanded domain-specific documentation framework that goes beyond the clinical trials CONSORT-AI guidelines. This endeavor could be expedited by leveraging insights from diverse industries, such as adopting practices like Google's Model Card, Meta's Methods Cards, TRIPOD from healthcare, and Stanford's XAI (Explainable Artificial Intelligence).

GAMP 5 version 2. Appendix D11 has a good approach for the validation of AI Models. It complements traditional Computer System Validation but adds new Data Acquisition and Monitor and Evaluate sections.

2. *Selecting model types and algorithms for a given context of use?*

In the selection of models, both black-box and others, stakeholders consider various factors such as the amount of available data, level of domain knowledge and potential to incorporate mechanistic elements into the model, the desired level of accuracy and its trade-off with model complexity, and computational and infrastructure requirements to meet the required speed of inference in low latency applications. Determination of how ML will be applied in solving the business problem and the intended level of ML application (e.g., whether it will inform, reduce, or replace human decision-makers) are also key criteria for model selection.

3. *Determining when to use specific approaches for validating models and measuring performance in a given context of use (e.g., selecting relevant success criteria and performance measures)?*

Traditional machine learning methods often benefit from well-established validation guidance, which pays particular attention to ensuring the alignment between validation criteria and the intended use scenarios. Validating models are based on available data, and size of data. Team discussions involving experts help reach a consensus on success criteria, and performance measures and often need to account for current practices and field standards.

Context of use needs to be specified clearly and the place in the workflow the AI application will be deployed. The question is whether the testing population matches the training population. If a model can accurately predict under assumptions, the accuracy on similar assumptions should generally hold. Performance targets can be relatively low depending on context of use. For example, a false positive does not lead to a significant increase in risk for the patient, but a true positive may yield a significant benefit. As another example, the absolute performance of a given measure is in a lower range but changes in that measure over time is still indicative of clinically meaningful change.

4. *Evaluating transparency and explainability and increasing model transparency?*

Stakeholders elucidate the explainability of models by employing various methodologies and evaluating the significance of features after the model has been built. Models based on SHAP (Shapley Additive Explanations) can be used for elucidating the attribution of features. Black box models, such as those based on deep learning and gradient methods, often present a challenge, however, in that while providing higher accuracy, there is less explainability and thus reduced ability to interpret results. Stakeholders must then balance transparency and accuracy, and may have to make tradeoffs, based on the relative importance of these attributes for the various use cases.

5. *Addressing issues of accuracy and explainability (e.g., scenarios where models may provide increased accuracy, while having limitations in explainability)?*

Although increasing the size and complexity of the ML model structure may help reduce the fitting accuracy, it often raises the risk of overfitting and the need for larger training datasets. In certain applications where domain knowledge and expertise are available, model developers can leverage mechanistic knowledge and hybrid modeling techniques to enhance accuracy and model explainability simultaneously, while mitigating data scarcity. Generalizability and explainability should be preferred over insignificant improvement in accuracy. Modern techniques, including Physics-Informed Neural Networks, provide a systematic framework to train models that leverage governing principles and measured experimental data.

6. *Selecting open-source AI software for AI/ML model development? What are considerations when using open-source AI software?*

Generally, stakeholders ensure that the models utilized in drug development have been thoroughly vetted on platforms like CRAN and internally build validation platforms and other reliable sources before integrating them into their practices. Additionally, stakeholders often seek to utilize software packages originating from their own tested environments, to ensure they are utilizing stable versions widely accepted and used within the community. The scrutiny and additional validation of packages are also performed based on the specific use case they are intended for.

The use of open-source code may have advantages and remove complexities around the use of proprietary algorithms. Open-source packages with sufficient and transparent documentation, rigorously tested, and active support community should be considered. The stability and maintenance of the open-source package are important factors to consider. Developers should assess whether the package receives regular updates, bug fixes, and security patches. Active development and maintenance indicate a healthy and reliable software ecosystem.

Open-source licenses may have varying restrictions on usage, modification, and distribution. It is crucial to understand the terms of the license and ensure compliance with any requirements or obligations. Moreover, security considerations should not be overlooked when using open-source AI software. Developers need to assess the security practices and protocols to ensure the protection of sensitive data and guard against potential vulnerabilities. In addition, compatibility with existing infrastructure, systems and codes is essential to ensure seamless integration of open-source AI software into the development environment.

7. *The use of RWD performance in monitoring AI/ML?*

AI/ML model development and validation is only the first step of AI/Model life cycle. Once a model is deployed, there needs to be continuous monitoring of model performance using real world data as well as any data distribution shift. Most of the AI/ML models should not be static, they need to be retrained and updated using RWD continuously. Guidance from FDA on model performance monitoring, updating, and re-validating using RWD would be extremely beneficial. When considering the use of RWD, certain considerations should be addressed to ensure performance. These include the RWD collection period, geographical/population sample, collection methods, version control, and quality control.

b. What practices and documentation are being used to inform and record data source selection and inclusion or exclusion criteria?

AI/ML should be treated similarly to any other technical or statistical analysis. The same approaches should be used as are applied to other scientific projects and statistical analyses. These center around study synopses, protocols, statistical analysis plans, study reports, peer review and reproducible research practice.

Some common considerations while selecting data sources include:

- Is the data representative (i.e., is it Structured vs. Unstructured, Balanced vs Imbalanced, Genomic biomarkers vs Proxy biomarkers, text, numbers, images, tables, etc.)
- Does the data form align with the chosen algorithm?
- Does the data in scope align with the system of interest, question of interest, and context of use?
- Is contextual data already available or are activities requiring some form of study design used (sampling and/or measurement)?
- How was the source data collected?
- How was any data splitting achieved (i.e, are there different datasets required for training, testing, and validating the machine learning model)? What is the procedure for version control of datasets after splitting?
- Were the models exposed to too much (overfitting) or too less (underfitting) to the training data? What are the ML training strategies for mitigating underfitting or overfitting of the resultant ML model?

c. In what context of use are stakeholders addressing explainability, and how have you balanced considerations of performance and explainability?

Explainability and model performance do not necessarily contradict to each other, and in many scenarios, they are independent from each other. The first step of developing a Machine

Learning model is to begin with a simple model that is easy to explain. However, in case of sub-optimal model performance, complex models are assessed for any performance improvements. Explainer libraries are used to get an understanding which factors have led to the improved performance of the model. Explainability should be by design and address the issue of trust and safety (e.g., use AI/ML model as part of the decision-making process to augment human capabilities rather than replacing the human decision maker). An emphasis on validating AI/ML models rather than forcing explainability should be considered.

d. What approaches are being used to document the assessment of uncertainty in model predictions, and how is uncertainty being communicated? What methods and standards should be developed to help support the assessment of uncertainty?

Uncertainty quantification and reporting is the key to AI/ML model safety. In general, there are two types of uncertainty. Aleatoric uncertainty (also known as data uncertainty) which is an inherent property of the data distribution, this cannot be reduced but can be estimated from the training data. In contrast, epistemic uncertainty (also known as knowledge uncertainty) occurs due to inadequate knowledge, this is when data used at inference time is from a different data distribution used for model training. While epistemic uncertainty can be reduced by designing and implementing a responsible data collection strategy, it cannot be completely avoided and must be addressed for AI/ML model safety (e.g., an out of distribution detection model should be combined with a prediction model to allow the final model to be inconclusive). Standardization of the ways we communicate uncertainty will be important in certain contexts.

Some approaches to document uncertainty include:

- <u>Probabilistic Modeling</u>: Probabilistic modeling approaches, such as Bayesian neural networks or Gaussian processes, enable the estimation of uncertainty by providing probability distributions over model predictions. These methods capture inherent uncertainty and allow for more reliable uncertainty quantification.
- <u>Uncertainty Quantification Metrics</u>: Developing standardized metrics for quantifying uncertainty can aid the assessment and comparison of different AI/ML models. Metrics such as predictive variance, mutual information, or calibration measures like expected calibration error can help evaluate the accuracy and calibration of uncertainty estimates