Computing Community Consortium's Response to FDA RFI on Using AI and ML in the Development of Drug and Biological Products

July 10, 2023

Written By: David Danks (University of California, San Diego), Mona Singh (Princeton University), Kevin Fu (Northeastern University), and Haley Griffin (Computing Community Consortium)

This response to FDA's notice titled "Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products," FDA-2023-N-0743, 88 Fed. Reg. 30313 (May 11, 2023) includes answers that are divided into the three key areas identified in the solicitation: (1) Human-led governance, accountability, and transparency, (2) Quality, reliability, and representativeness of data, and (3) Model development, performance, monitoring, and validation. The questions that we chose to respond to from each area are italicized.

(1) Human-led governance, accountability, and transparency

In what specific use cases or applications of AI/ML in drug development are there the greatest need for additional regulatory clarity?: AI/ML is rapidly becoming a pervasive aspect of all drug development applications, and so we believe that it is more important to focus on conversations related to the use of AI not being siloed. In particular, there is substantial expertise at FDA in uses of AI for other medical contexts–specifically, AI experts in the Center for Devices and Radiological Health (CDRH)--and we advocate leveraging their expertise in a formal, cross-cutting manner, potentially through a focal point program dedicated to AI, as a way to improve regulatory clarity.

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): The necessity for transparency manifests in different ways depending on the application of the AI/ML. For instance, if AI is used in compound optimization and a researcher follows up with experimental verification, there isn't a compelling need for a high level of transparency due to the subsequent experimental work. However, if a researcher wants to speed up a clinical trial, there needs to be more transparency into the code and data that purports to justify the more rapid schedule.

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)?: As researchers, we find that the proprietary nature of even potential leads is the most serious barrier to collaboration or understanding. In this regard, we note that drug development processes are similar to software development, as both involve little public sharing, even when something of substantial public benefit has been achieved.

Having said that, we note that the FDA is known for having an admirable commitment to carefully protecting trade secrets, and this needs to extend with an appropriate scoping to uses of AI/ML in drug development. We believe that companies should feel secure sharing the full extent of their process with the FDA given its track record of keeping intellectual property confidential. We hope that such information sharing (in this case, about the AI) would enable oversight and assessment of algorithms (e.g., checking for tainted data, design flaws, proper testing and validation of machine learning methods, etc.).

2) Quality, reliability, and representativeness of data

What additional data considerations exist for AI/ML in the drug development process?: Data sets, at the training level and beyond, need to be thoroughly examined for bias prior to being used. Although there is now a general recognition of the need to consider bias in datasets, these checks are often limited to only a few possible biases (e.g., checking only if the dataset is gender-balanced). Types of bias that need to be scrutinized include, but are not limited to, population bias, socioeconomic status, and genetic diversity. AI/ML can easily pick up on accidental patterns, and so it is critical for the data to be as clean and carefully considered as feasible.

There has been much less discussion about the use of 3rd party data, even though they are likely to play an increasingly important role (particularly as the number of data brokers grows). 3rd party data can enable drug developers to easily and quickly acquire data about some factors that could potentially be relevant, and so clear guidelines need to be developed for the acquisition and use of these datasets. For example, we propose that companies should (by default) not mix biomedical and non-biomedical data sources.

What are some of the key practices utilized by stakeholders to help ensure data privacy and security?: There are well-established best practices for privacy and security, and it is critical that companies in these spaces ensure that they are following them. Moreover, we note that companies have significant incentives to do so, as these best practices help (1) to protect business assets for their own profits and to avoid bad publicity, and/or (2) to comply with regulatory requirements.

What are some of the key practices utilized by stakeholders to help address issues of reproducibility and replicability?: The practice in scientific journal publication of providing the code and notebooks that generate figures is a necessary level of openness for the sake of reproducibility and replicability. It is not sufficient to just claim that an AI system performed to a certain level; rather, researchers need to make their processes available so that others can potentially reproduce their findings.

What processes are developers using for bias identification and management?: Developers are completing data audits prior to training or using AI. These audits need to be multi-dimensional and sophisticated in order to measure bias across a wide range of demographic properties.

3) Model development, performance, monitoring, and validation

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

a) *Selecting model types and algorithms for a given context of use?:* Extensive testing needs to be done on any external data sets. There needs to be automated hyperparameter tuning. Researchers need to be careful to not be constrained by the background knowledge they have about the data, but rather look at the results at face value.
b) *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)?:* Validation and performance measures should be collectively determined by the FDA and the company, not by the company alone. Like with many conversations around AI/ML, there is not a long history of these kinds of discussions and negotiations when it comes to the uses of AI. We think that it is critical for the FDA and relevant standards bodies to build up sufficient expertise to engage in these precedent-setting negotiations, and this may require tapping additional or novel sources of talent and expertise as well as preparing to integrate AI performance measures into consensus standards.
c) *Evaluating transparency and explainability and increasing model transparency?:* In general, we believe that it is important for evaluation bodies such as the FDA to have full access to a model, particularly when it is playing an important role in major decisions. We recognize that some uses of AI in drug development (e.g., generating some potential leads for subsequent research) may not require full transparency. But in general, transparency for the evaluators is crucial for reproducibility and generalizability. Importantly, we do not necessarily believe that algorithms should always be explainable. Rather, the need for explainability should be assessed on a case-by-case, and use-by-use, basis.

d) *Addressing issues of accuracy and explainability (e.g., scenarios where models may provide increased accuracy, while having limitations in explainability)?:* These issues are very application-specific and challenging. Explainability can be very beneficial, but should not be prioritized over accuracy.
e) *Selecting open-source AI software for AI/ML model development? What are considerations when using open-source AI software?:* It is key for the training process to also be transparent; not just the use of the software.
f) *The use of RWD performance in monitoring AI/ML?:* There should be continuous monitoring of AI systems while it is in use. For instance, if AI is being used to decide whether a patient receives a particular drug, researchers need to collect data and re-evaluate that decision. This evaluation should determine the accuracy of the method, and might lead to it being changed. Frequent monitoring is especially important if the system is doing anything adaptive or being deployed in novel contexts.

In what context of use are stakeholders addressing explainability, and how have you balanced considerations of performance and explainability?: Explainability can be a helpful mechanism, but needs to be analyzed carefully. This is especially true when using Generative AI like Chat GPT, which is excellent at providing compelling explanations that are often wrong. Additionally, sometimes performance is much more important than explainability; for instance, when predicting adverse drug effects it is better to have higher performance/fewer false negatives than to be explainable.

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?: ML models often just come with a prediction. However, we believe that it is critical that models used in AI/health also have assessments of uncertainty for individual predictions. The techniques used to estimate model uncertainty on specific predictions can be application-specific, or can be estimated empirically by measuring performance of trained models on test data stratified by various characteristics (e.g., similarity to the training datasets).