Duke | MARGOLIS CENTER for Health Policy

# Understanding Artificial Intelligence and Machine Learning (AI/ML) in the Drug Development Lifecycle

Summary of Expert Workshop

*The Robert J. Margolis, MD, Center for Health Policy at Duke University, under a cooperative agreement with the U.S. Food and Drug Administration (FDA), hosted an expert workshop on December 13, 2022, entitled "Understanding Artificial Intelligence and Machine Learning (AI/ML) in the Drug Development Lifecycle." This workshop focused on artificial intelligence (AI), including machine learning (ML), throughout the drug development lifecycle, with an emphasis on those areas that require regulatory clarity.*

## Introduction

In recent years, there has been a significant increase in the utilization of artificial intelligence and machine learning (AI/ML) in health care and drug development. The number of medical publications involving AI/ML has grown from slightly over 200 in 2005 to over 12,500 in 2020.[1] FDA's Center for Drug Evaluation and Research (CDER) evaluated submissions with AI/ML components and found that the number of submissions has increased 2-3-fold yearly since 2016, with most of the AI/ML use occurring at the clinical research drug development stage.[2]

### Uses of AI/ML in drug development

This expert workshop discussed the landscape of AI/ML in drug development, exploring its wide-ranging applications throughout the drug development process, including drug discovery, clinical research, and postmarket safety surveillance. In early research and drug discovery, AI/ML tools can be used to identify potential drug candidates, model drug-protein interactions, and predict target protein structures. During the clinical research phase, AI/ML tools discussed played a role in patient recruitment, selection, stratification, retention, dose optimization, and site selection.

Workshop participants expressed great excitement for AI/ML tools using real-world data (RWD) or data generated from digital health technologies (DHTs) as part of novel clinical endpoints to evaluate the safety and efficacy of medical interventions in clinical trials. Other AI/ML tools could be used to assist in the timely processing of large amounts of RWD to identify and detect potential safety signals. Moreover, the workshop highlighted the significant growth in the application of AI/ML tools for imaging analysis, as imaging data are much more interoperable than other forms of clinical RWD. The participants also discussed the promising operational applications of other AI/ML tools, such as smart monitoring, which can automate tasks that are time-consuming and labor-intensive, improving system operations.

Another emerging use for AI/ML technology that garnered attention throughout the workshop was precision medicine, which involves identifying the right drug and dose at the right time for an individual patient, as well as determining the optimal treatment regimen. AI/ML tools show potential in enhancing precision medicine strategies by enabling in-depth analyses of patient characteristics and facilitating the development of predictive models that can be precisely tuned to each individual. Although there has been remarkable progress in this area, one workshop participant noted that progress may decelerate, as increasing the complexity of precision medicine models may lead to diminishing returns over time. Precision medicine holds significant value for rare or orphan diseases, as they face challenges in drug development due to small patient populations and limited opportunities for conducting large-scale dose-

[1] https://www.nature.com/articles/s41746-020-00333-z
[2] https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.2668

response clinical trials or characterizing disease progression. Although there is scant literature in this area, the integration of AI/ML with multi-omics holds promise for predicting biomarkers that are more challenging to measure, further improving precision medicine.

During the workshop, participants shared successful uses of AI/ML techniques employed in drug development. These included using AI/ML methods to support Emergency Use Authorization (EUA) applications; enhancing endpoints based on traditional disease severity scales (e.g., stroke); and improving the duration of progression-free survival of patients (e.g., AI tool trained on the morphology of tumor cells and their reaction to treatments *ex vivo*). Despite these advances in the field of AI/ML, challenges still exist for research and drug development in a rapidly evolving field.

## Challenges

The use of AI/ML tools in drug development faces several challenges related to:

1. **Data fitness for purpose (i.e., relevance and reliability):** According to one participant, as much as 80 to 90 percent of the work required for model development focuses on identifying the right data to build an AI tool. For a dataset to be fit for purpose, it must contain both the right elements and the requisite data quality attributes (e.g., accuracy, completeness) for the intended use and target population. It is widely recognized that the accuracy of an AI tool hinges on the quality of the data it is trained on, and participants cited the adage, "garbage in, garbage out."

   *Data completeness:* Existing datasets can be incomplete for a variety of reasons. One example is patient mobility, wherein individuals receiving care in multiple locations have fragmented health records dispersed among different data sources. Aggregating or linking such records and ensuring comprehensive, longitudinal follow-up can be challenging, especially because most data from electronic health records (EHR) are not standardized and important information is often in free text, which further complicates matching. Workshop participants also acknowledged that it can be difficult to determine when there is enough data to derive meaningful insights.

2. **Data access and sharing:** Much of the data from clinical trials, EHRs, and commercial claims that are desirable for use for building AI/ML models are difficult to access because they are privately owned or reside behind privacy-preserving firewalls. One of the biggest obstacles to drug development using AI/ML is related to issues with accessing and sharing training and testing data. As one participant noted, "We have been talking about [data access and sharing] for twenty-five years. What would it take to not have to wait another year?" Data owners may be reluctant to share data without adequate assurances of privacy protection. For datasets to be shared broadly or publicly, workshop participants believed there was a need for established norms, rules, privacy safeguards, and liability frameworks for data breaches. Some data owners suggested the government should provide a secure data-sharing platform. Despite company efforts to facilitate data sharing, reaching a consensus on data usage agreements can be difficult. A proposed "federated" solution to bypass data sharing concerns was proposed by a workshop participant: having the tool developer provide the data owner with a pre-specified model or tool and then transmitting the tool's output back to the developer, bypassing data sharing concerns.

3. **Security and privacy:** Although large public datasets have facilitated AI/ML tool development in non-medical sectors, privacy concerns and proprietary data rights present significant challenges in creating similar open data sets for health care and clinical pharmaceutical development. Constraints imposed by initial consent forms for clinical trials further hinder the reuse of clinical trial data sets for secondary uses, and reduce the availability of usable data sources. Workshop participants discussed novel methods to more effectively de-identify data that could be used, and

proposed building synthetic datasets using real-world data or data from clinical trials as a privacy-preserving method to grant data access. The federated solution described above could also address security and privacy concerns.

4. **Bias identification and minimization**: AI/ML tools may be susceptible to reproducing and amplifying systemic biases and inequities, which can arise from biased or unrepresentative training data. Choices made during data selection, curation, preparation, and model development can further introduce bias that can be amplified in the outputs of AI/ML tools. It is important to test tools within specific subpopulations, both during implementation and over time, to detect any potential bias or degradation of the model performance due to changes in the input or output variables.

5. **Dataset shift:** This shift occurs when the distribution of data attributes in the training dataset differs from that of the deployed model. This discrepancy can generally limit the scope and generalizability of an AI/ML tool to a broader population. During the drug development process, dataset shift can occur when the prevalence of different attributes varies over time among the clinical trial subpopulations or between site locations, compared to the model's training and testing population. Some participants proposed the idea of anticipating dataset shift by incorporating an adaptive and controlled plan into the model, similar to predetermined change control plans.[3,4]

6. **Transparency:** Certain AI/ML models and tools may lack transparency, operating as black boxes, leaving little ability to understand how the inputs are weighted and combined to generate outputs. This means that it can be entirely unclear what elements are driving the model's decisions. AI/ML models may pick up signals that may not be apparent or prioritized by human experts. An example of this is an AI/ML model trained to read electrocardiogram (ECG) data to determine adverse cardiac events. The model began to identify the statistical noise patterns from different ECG machines, impacting the model's ability to accurately determine patient outcomes. The use of black box algorithms also makes it difficult to assess how generalizable a tool is. Models trained on data from homogeneous subpopulations or clinical sites may not be generalizable to new populations of interest. While methods to determine the drivers of model performance exist, participants expressed that these methods can be labor-intensive, especially when evaluating performance differences among subgroups.

## Areas where additional regulatory perspective may be needed

Regulatory science plays a pivotal role in protecting patient safety in an innovative and rapidly evolving field. The workshop participants agreed on the necessity for greater regulatory clarity in determining standards for the use of AI/ML in drug development. They recognized the need for flexible regulatory frameworks that could effectively assess how the agency would evaluate the risks posed by different AI/ML tools used for drug development. There were concerns, however, that regulatory agencies may not currently have the necessary resources to understand and manage the broad spectrum of technologies used in healthcare and therapeutic development. For the proposed frameworks to effectively support the use of AI/ML in drug development while upholding public safety, there must be a mutual understanding

[3] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[4] https://www.fda.gov/media/166704/download

of the standards that will be used for regulatory approval and of the relevant tradeoffs of using AI/ML in drug development.

## Choosing a suitable dataset

To determine the suitability of a training dataset, a proposed framework based on three core principles was discussed. The first principle emphasized the importance of assessing a dataset based on its context of use. The second principle emphasized selecting outcomes of interest from a dataset, and providing a clear rationale for their selection. The third principle was to assess the representativeness and completeness of the data, while also considering potential biases within the included population. That is, the third principle considers who is *missing* from the dataset, and who may be *present but misrepresented*. By documenting these considerations, regulatory agencies can effectively evaluate the trade-offs made during the selection process, such as balancing completeness and representativeness, and determine whether the datasets used are appropriate for assessing the desired outcomes.

When selecting data for training or testing an AI/ML tool, it is vital to assess data quality. Several frameworks exist for documenting the content and quality of datasets, with one notable example being the FAIR data principles, which refer to the findability, accessibility, interoperability, and reusability of data. The FAIR framework was developed to ensure that researchers and other stakeholders can effectively understand the quality and usability of a dataset and to optimize its reusability. However, one workshop participant shared that adherence to these standards is not consistently practiced. As a result, there needs to be greater transparency in the documentation process, particularly regarding the tradeoffs made and how they have changed over time. Workshop participants also emphasized the importance of considering data fitness for purpose, highlighting the benefits of "building for purpose" (i.e., building datasets with the intended use in mind). In this way, AI/ML tools can be built more effectively if the datasets used for training are designed for their specific use cases.

One example of data suitability pertained to representational bias coming from the inclusion of "responsiveness to treatment" as an input variable in a model. Individuals who lacked adequate healthcare access due to being underinsured were often labeled as having low responsiveness to treatment. Therefore, the model was more likely to predict patients with low socioeconomic status as unlikely to respond favorably to treatment. Unadjusted, this AI model could inadvertently amplify disparities in health care access. Some academic frameworks have been established to understand and mitigate bias (e.g., imputing missing data, stratified analysis). Leveraging the expertise of multiple fields could help implement measures that mitigate bias and promote fair and equitable outcomes. For example, a collaboration between scientists specialized in performance analysis and bias mitigation and software engineers experienced in automating processes could prove beneficial in reducing bias.

## Risk-based framework

*"The distance from the patient allows for more stops along the way. AI used in research is much more distant than AI used by a physician to treat a patient."*

Considering the varying levels of risks among drug development applications, workshop participants suggested that regulatory agencies could categorize risk based on the direct relevance of AI/ML tools to patient care. The consensus among participants was that AI/ML applications in medical contexts for drug development pose higher risks to society compared to many non-medical applications, but there are still gradations within the medical space. Generally, an AI/ML tool used for clinical decision-making would carry more risk than one used for early-stage research or drug discovery. As one participant noted, "The distance from the patient allows for more stops along the way. AI used in research is much more distant than AI used by a physician to treat a patient."

Currently, the determination of risk levels for specific applications remains nebulous. The absence of a standardized evaluation framework makes it challenging for tool developers to assess risk. Different regulatory agencies may come to different conclusions evaluating any particular use case. Workshop participants shared their experiences in various countries, highlighting that different regulatory agencies sometimes determine AI/ML model risks in different ways. For example, one agency reviews AI/ML model risk based on two variables: model influence and decision consequence. Model influence refers to the weight placed on the output model compared to other evidence. The risk associated with a specific AI/ML model would be lower if its outputs are weighted lower than other evidence. Decision consequence refers to the negative effects of a wrong decision and depends on the context of use, particularly its impact on clinical trials and postmarket clinical decisions. It is important for the decision consequence of each different use case to be evaluated independently. Other regulatory agencies evaluate AI/ML tools in drug development similar to all other evidence used in the drug approval process. As a result of the regulatory uncertainty and variation regarding evaluating AI/ML risks in drug development, one workshop participant noted that developers have reacted in two distinct ways: some are hesitant to create any new AI/ML tools, while others interpret the substantial uncertainty to mean that there is little to no established regulation, leading them to believe they can operate without restrictions.

Several principles for risk frameworks were proposed, considering the potential harm of either a false positive or a false negative. These frameworks aim to establish an acceptable level of risk-benefit tradeoff or a balance between the positive and negative effects of AI/ML tools. Workshop participants emphasized that every AI/ML tool will carry some quantity of risk, just as the standard-of-care clinical decision frameworks currently used as comparators carry inherent risk. Additionally, participants stressed the importance of considering both the risks of action and inaction in a clinical setting. The aim is not to eliminate all risks but to create a system with an acceptable tradeoff between false positive and false negative results.

*The aim is to create a model with an acceptable tradeoff between false positive and false negative results.*

## Decision framework

Several methods for evaluating tradeoffs were discussed during the workshop. One of these methods is a decision curve analysis, which can be used to determine the cost of false positive and false negative results, ultimately identifying the optimal threshold for balancing these costs. By using such an approach, researchers can determine the threshold that maximizes the overall benefit for patients. This approach

allows researchers to test drugs at different risk thresholds to determine the cost-benefit tradeoff for that intervention. Another method discussed is the "value-of-information" framework, adopted from econometric analysis. This framework quantifies the expected monetary gain of reducing uncertainty in the input parameters of the decision model. It also estimates the expected value of perfect information, and can help assess the risk-benefit tradeoffs associated with AI tool development and use. Workshop participants shared that AI/ML tools may become riskier as inputs move away from the standard range of evaluation.

Workshop participants also noted that AI/ML tools produce outputs in the form of probabilities and that these probabilities must be tied to a predefined decision framework. Therefore, it is important to carefully establish a decision framework *a priori*, as the selection of probability thresholds retrospectively can introduce bias. Additionally, risk mitigation can be improved by establishing checks and bounds on the model's impact on clinical decisions.

Regulatory clarity is needed to ensure best practices in evaluating the appropriateness of the datasets used to build AI/ML models for drug development, as well as the impact of the model on the resulting body of evidence on the drug. Advancements in the field will be achieved through continued collaborations and discussions, aiming to align regulatory agencies and all interested parties on these frameworks.

## Standards for data access and sharing

Throughout the workshop, various strategies were proposed to address the challenges of data sharing and to ensure broad access to data. Proposed solutions included the use of biobanks and other dedicated systems designed explicitly for data sharing. These approaches can enhance data representativeness by standardizing data collection methods, increasing the dataset's sample size, and decreasing bias associated with site-based selection. However, establishing and making these databanks usable for data providers and tool developers requires concerted effort and commitment from many stakeholders. Data sharing has been promoted by successful initiatives such as C-PATH and other public-private partnerships. One participant suggested expanding the Aggregate Analysis of ClinicalTrials.gov (AACT) database to make it more effective for clinical data sharing. AACT currently contains information on all registered and publicly available clinical studies. AACT includes information from both interventional and observational studies and provides details on trial information, cohort information, and clinical results of different trial arms. An expanded AACT could be more standardized and include more detail about the clinical data collected during the trial.

It's important to note that while centralized organizations can play a valuable role in facilitating data findability, these organizations cannot share confidential data from companies without prior agreements. Some participants expressed concern that the process of obtaining approval for the secondary use of existing data is challenging. Further regulatory action could help streamline this process, for example, by providing more clarity on the use of broad consent for secondary use of clinical trial data.

In some countries, different data-sharing systems exist for researchers to access medical data. For example, in the UK there is a landing page[5] that compiles a comprehensive list of nearly all available datasets. Although granular data is usually not available, high-level summaries are provided and individual data controllers manage access to them. While this system is not necessarily very complex, it is useful for researchers to find datasets that meet their needs. Adopting a similar approach in other governments could facilitate data sharing efforts.

[5] https://www.healthdatagateway.org/

In general, data sharing relies on the coordination of multiple stakeholders who sometimes have competing interests. To enable collective action in this domain, participants at the workshop suggested that it might be helpful for the government to set up a system or a comprehensive set of rules for broader data access and sharing. One suggestion put forward was that government agencies could require that studies they fund make their data sharable, in addition to presenting clinical trial study summary information (e.g., protocol, results). By implementing such conditions, it would encourage greater data sharing and facilitate collaboration among stakeholders.

## Priorities for future development

Participants outlined several priorities and goals for the future development and use of AI/ML tools in drug development, including:

1. **Balancing the accuracy and speed of outputs and the need for explainability**. When weighing the tradeoffs between performance and explainability, the consensus was that prioritizing the reliability of performance takes precedence over explainability. While explainability is valuable for fostering trust in a tool, it holds little value if the tool does not adequately perform the tasks it is intended to. In situations where improved performance compromises the explainability of the tool, alternative forms of transparency can be beneficial, such as sharing the model's design and inputs. Providing an explanation of how a tool reaches its conclusions may help establish the boundaries of accurate performance and help with risk assessment. Therefore, defining boundaries for model use can improve the overall performance
2. **Incorporating humans in the decision-making process.** A "human in the loop" approach is a way to mitigate risk by integrating human expertise within the decision-making process. This approach may be useful for applications of AI/ML tools that are close to patient care, such as clinical practice, in silico trials, digital twins, and counterfactual simulations of placebo arms. In these cases, it is important to evaluate the performance of the collaborative human-AI team rather than solely relying on the model's performance in isolation, as real-world performance will depend on the usability and interpretability of the AI tool's output. Although this approach may simplify regulatory considerations, regulators emphasized that the use of a tool without a "human in the loop" is acceptable as long as there is sufficient evidence supporting its efficacy.
3. **Assessing and characterizing the generalizability of a model's performance**. This process involves assessing the model's performance across different clinical trials, or when applying a model to inputs that may be of differing quality or diversity than its training population. As an alternative to generalizability, regulators could establish a process similar to a predetermined change control plan, which includes outlining anticipated modifications to the tool based on the retraining and model update strategies, as well as the methodology for implementing changes to retrain algorithms for new specific uses. Participants shared the challenges of establishing predetermined plans in such a nascent field, as things can change in unexpected ways. A pre-set schedule of re-training intervals was suggested to evaluate both expected and unexpected changes. Another method to mitigate generalizability issues caused by dataset shift is to build and deploy models specifically designed for a single location. However, localizing a tool to a particular hospital or patient population increases the risk of overfitting.
4. **Ensuring model integrity through continuous monitoring for performance and bias**. When AI/ML tools are deployed, their initial performance may be satisfactory; however, changes in data and other factors can affect accuracy over time. Custom software can be used for the continuous monitoring of performance, triggering alerts users and tool developers to investigate and

implement necessary modifications. By implementing this strategy, developers can ensure the long-term effectiveness and accuracy of AI/ML models, ultimately leading to better model generalizability.

## Conclusion

Workshop participants discussed the importance of establishing trust in AI/ML tools for drug development to ensure that drugs are safe and effective while facilitating innovations in their development. The core standards of explainability and transparency are important factors to build trust in AI/ML tools among patients, healthcare providers, regulators, and tool users. As technology advances and the use of AI/ML models become more prevalent, workshop participants expected that trust in these systems would grow. While evidence of high performance can supplant the need for explainability, it can also introduce challenges in risk assessment. Over time, as these systems mature, the necessity for explainability may diminish.

Trust can also be established through the integrity of the data and how well the data reflects its intended use. Improving the reproducibility and scalability of models will contribute to building trust not only in specific models but in models more generally. Participants discussed several strategies to build trust during the workshop, with the "human in the loop" approach being repeatedly emphasized. While the presence of a human-AI/ML system can increase the trust that other stakeholders have in the outputs of the tool, it was also noted that human operators can introduce errors, variability, and uncertainty. This may make the use of such a system difficult to scale, as it would require further training of human operators on the proper incorporation of tool outputs. For this approach to be effective, human involvement and feedback should not only be integrated in the model development process, but also in the design of the operation of the tool.

For the field of AI/ML in drug development to advance, it is important not just to think about narrowly defined trust in AI/ML tools, but also about its usability and whether the tool gives good results. Participants stressed the need for flexible regulations that can adapt to the rapidly changing technological landscape. As familiarity with AI/ML technology grows, more regulatory clarity is provided, and more tools are developed and deployed in drug development, clinical trials, and other healthcare applications, trust will continue to be fostered.

Duke | MARGOLIS CENTER for Health Policy

## Acknowledgements

Duke-Margolis would like to thank the experts listed below for their participation in our expert workshop on Dec 13, 2022 and for their insightful contributions during the discussion, which have informed the development of this summary. The analysis put forward in this brief does not necessarily reflect the views, opinions, or positions of any of the individuals listed below, or their affiliated* organizations.

Participants:

**Blythe Adamson**
Flatiron Health

**Ib Alstrup**
European Medicines Agency

**Puneet Batra**
Broad Institute

**Sanjeev Bhavnani**
Center for Devices and Radiological Health,
U.S. Food and Drug Administration

**Anat Boehm-Cagan**
Ministry of Health, Israel

**Lisbeth Bregnhoj**
European Medicines Agency

**John Concato**
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Dominique Demolle**
Cognivia

**Alastair Denniston**
University of Birmingham

**Catherine Ela**
Ministry of Health, Israel

**M. Khair ElZarrad**
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Tala Fakhouri**
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Usama Fayyad**
Northeastern University

**Charles Fisher**
Unlearn.AI

**Asieh Golozar**
Odysseus Data Services, Inc

**Ben M.W. Illigens**
Novadiscovery

**Shameer Khader**
Sanofi

**Jesper Kjær**
Danish Medicines Agency

**Mike Krams**
Exscientia

**Qi Liu**
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Chris McCurdy**
Amazon Web Services

**Catherine Njue**
Health Canada

*Affiliations at the time of the workshop.

**Brendan O'Leary**
Center for Devices and Radiological Health,
U.S. Food and Drug Administration

**Anindita (Annie) Saha**
Center for Devices and Radiological Health,
U.S. Food and Drug Administration

**Johan Ordish**
Medicines and Healthcare Products Regulatory Agency

**Edgar Simard**
Verily

**Elif Ozkirimli**
Roche

**Karandeep Singh**
University of Michigan

**Allison Pearson**
Flatiron Health

**Chris Steel**
IQVIA

**Michael Pencina**
Duke Clinical Research Institute

**Adarsh Subbaswamy**
Johns Hopkins University

**Krishna Prasad**
Medicines and Healthcare Products Regulatory Agency

**Chunhua Weng**
Columbia University

## Workshop Planning and Writing Committee

**Ethan Chupp**
Duke-Margolis Center for Health Policy

**Tala Fakhouri**
Office of Medical Policy,
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Marianne Hamilton Lopez**
Duke-Margolis Center for Health Policy

**Janice Maniwang**
Office of Medical Policy,
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Christina Silcox**
Duke-Margolis Center for Health Policy

**Dianne Paraoan**
Office of Medical Policy,
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

**Erin Soule**
Duke-Margolis Center for Health Policy

**Marsha Samson**
Office of Medical Policy,
Center for Drug Evaluation and Research,
U.S. Food and Drug Administration

*This publication was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award U01FD006807 totaling $2,575,023 with 100 percent funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.*