U.S. FOOD & DRUG ADMINISTRATION
CENTER FOR DRUG EVALUATION & RESEARCH
OFFICE OF CLINICAL PHARMACOLOGY

The Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products

Ruihao Huang, Ph.D.,
AI Review Lead
FDA/CDER/OTS/OCS
13th ASA NJ Chapter/Bayer Workshop

Disclaimer: This presentation reflects the views of the presenter and should not be construed to represent those of the FDA or the United States Government.

Proprietary

Outline

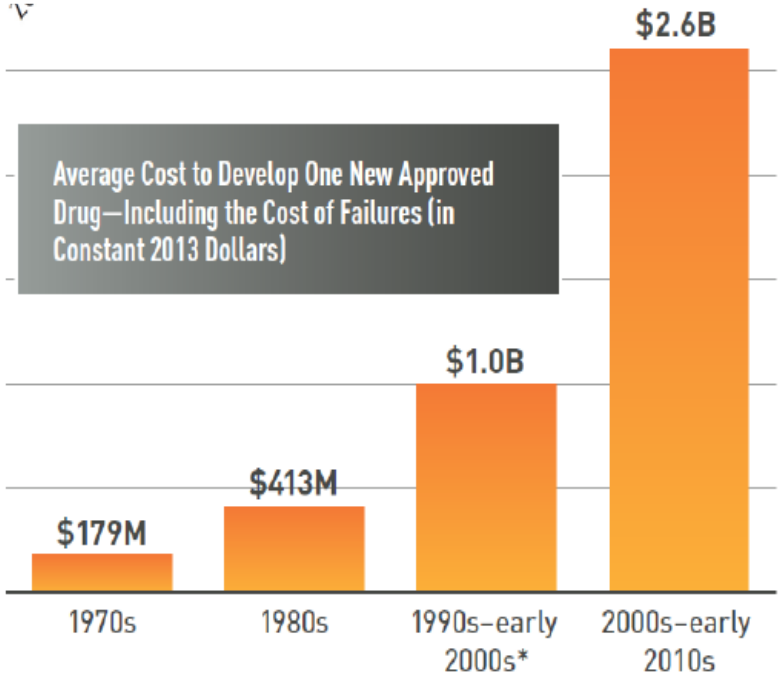
- Introduction
  - Landscape Analysis for AI/ML Related Submissions at the FDA
  - Considerations for AI/ML in Regulatory Decision Making
  - Risk-based Credibility Assessment Framework



- Case Examples
  - LLM-Based Clinical Endpoint Assessments
  - ML for Patients Risk Stratification
  - ML-based Enrichment Trial Design

Proprietary

Why are we talking about AI/ML today?



Tufts Center for the Study of Drug Development (CSDD) (2014).

Challenges in new drug development:

- **Decreasing productivity**
- **Increasing development cost**
- **High late phase attrition**

- The application of AI/ML in drug development is expanding rapidly.
- AI/ML have the potential to improve the efficiency of drug development and advance precision medicine, but they also have unique challenges.

Proprietary

Increasing AI Related Submissions at CDER

PERSPECTIVES

PERSPECTIVE

Landscape Analysis of the Application of Artificial Intelligence and Machine Learning in Regulatory Submissions for Drug Development From 2016 to 2021

Qi Liu[1,†], Ruihao Huang[1,†], Julie Hsieh[1,†], Hao Zhu[1,*,†], Mo Tiwari[1], Guansheng Liu[1], Daphney Jean[1], M. Khair ElZarrad[2], Tala Fakhouri[2], Steven Berman[3], Billy Dunn[3], Matthew C. Diamond[4] and Shiew-Mei Huang[1]

An analysis of regulatory submissions of drug and biological products to the US Food and Drug Administration from 2016 to 2021 demonstrated an increasing number of submissions that included artificial intelligence/machine learning (AI/ML). AI/ML was used to perform a variety of tasks, such as informing drug discovery/repurposing, enhancing clinical trial design elements, dose optimization, enhancing adherence to drug regimen, endpoint/biomarker assessment, and postmarketing surveillance. AI/ML is being increasingly explored to facilitate drug development.

BACKGROUND

Over the past decade, there has been a rapid expansion of artificial intelligence/machine learning (AI/ML) applications in biomedical research and therapeutic development. In 2019, Liu *et al.* provided an overview of how AI/ML was used to support drug development and regulatory submissions to the US Food and Drug Administration (FDA). The authors envisioned that AI/ML would play an increasingly important role in drug development.[1] That prediction has now been confirmed by this landscape analysis based on drug and biologic regulatory submissions to the FDA from 2016 to 2021.

THE TREND OF INCREASING AI/ML–RELATED SUBMISSIONS AT THE FDA'S CENTER FOR DRUG EVALUATION AND RESEARCH

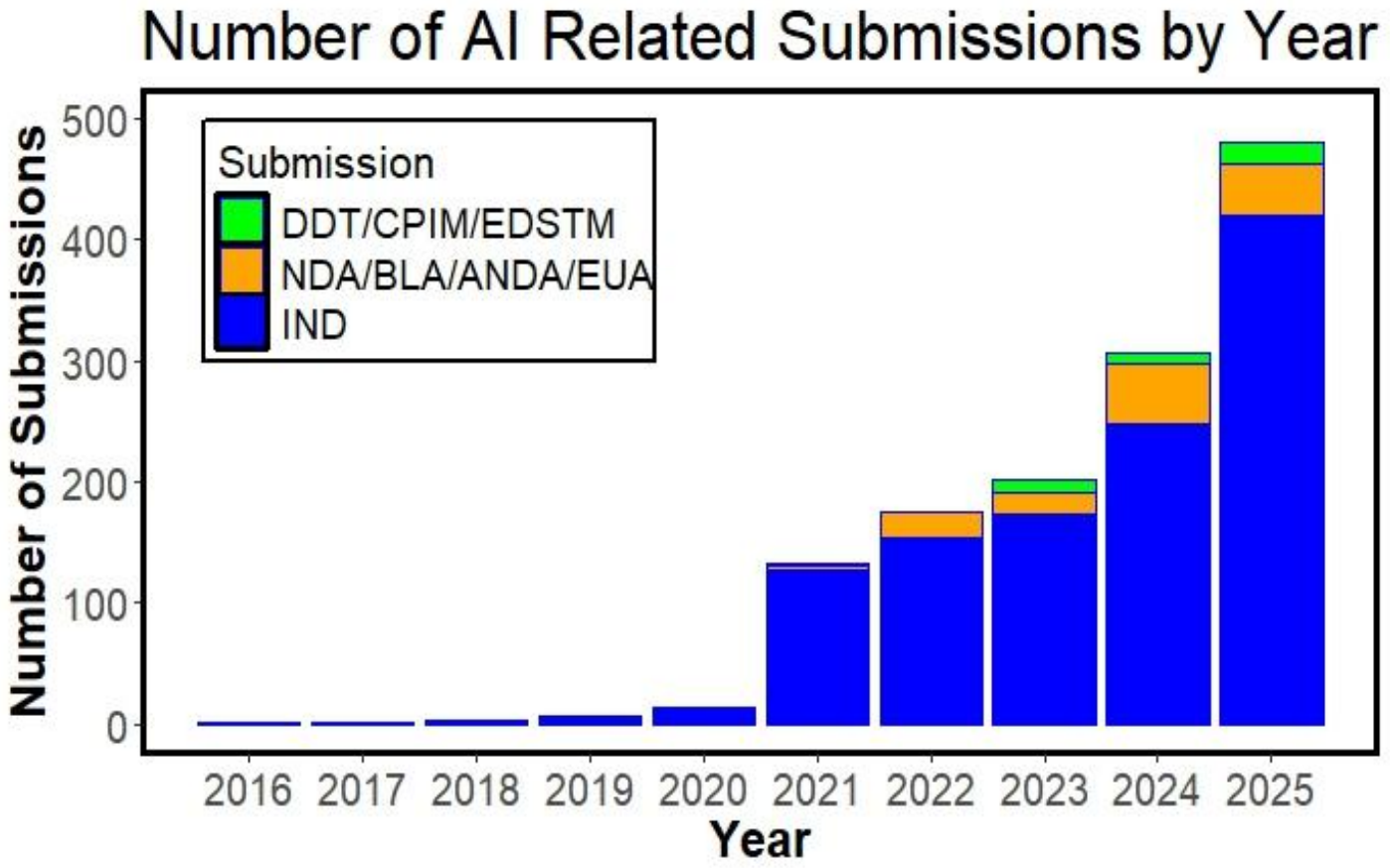
This analysis was performed by searching for submissions with key terms "machine learning" or "artificial intelligence" in Center for Drug Evaluation and Research (CDER) internal databases for Investigational New Drug applications, New Drug Applications, Abbreviated New Drug Applications, and Biologic License Applications, as well as submissions for Critical Path Innovation Meeting and the Drug Development Tools Program. We evaluated all data from 2016 to 2021. Figure 1a demonstrates that submissions with AI/ML components have increased rapidly in the past few years. In 2016 and 2017, we identified only one such submission each year. From 2017 to 2020, the numbers of submissions increased by approximately twofold to threefold yearly. Then in 2021, the number of submissions increased sharply to 132 (approximately 10-fold as compared with that in 2020). This trend of increasing submissions with AI/ML components is consistent with our expectation based on the observed increasing collaborations between the pharmaceutical and technology industries.

Figure 1b illustrates the distributions of these submissions by therapeutic area. Oncology, psychiatry, gastroenterology, and neurology were

[1]Office of Clinical Pharmacology, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [2]Office of Medical Policy, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [3]Office of New Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [4]Digital Health Center of Excellence, Center for Devices and Radiological Health (CDRH), US Food and Drug Administration, Silver Spring, Maryland, USA.
*Correspondence: Hao Zhu (hao.zhu@fda.hhs.gov)
†These authors contributed equally.

Received March 16, 2022; accepted May 19, 2022. doi:10.1002/cpt.2668

Updated analysis with submissions in 2022-2025

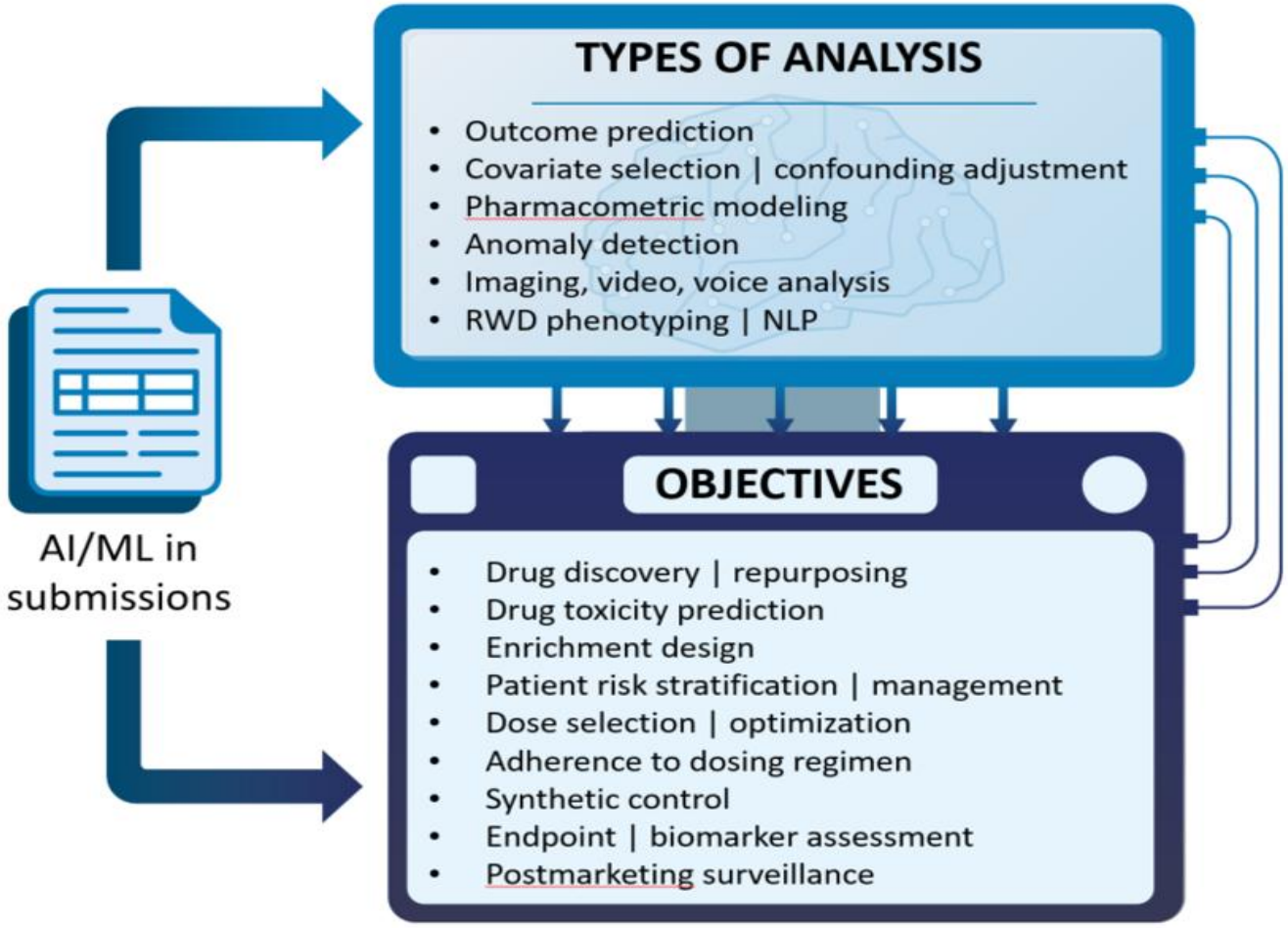


(Figure credit: Dr. Menglun Wang)

Proprietary

Types of AI/ML Related Analyses & Objectives

- A wide range of AI/ML algorithms
  - Neural network, LASSO, Elastic network, XGBoost, Random forest



Proprietary

Regulatory Decision-Making Example: ML-based Patient Population Selection

ARTICLE

Using Machine Learning to Determine a Suitable Patient Population for Anakinra for the Treatment of COVID-19 Under the Emergency Use Authorization

Qi Liu[1,†], Raj Nair[2,*,†], Ruihao Huang[1,†], Hao Zhu[1,†], Austin Anderson[2], Ozlem Belen[2], Van Tran[3], Rebecca Chiu[3], Karen Higgins[3], Jianmeng Chen[1], Lei He[1], Suresh Doddapaneni[1], Shiew-Mei Huang[1], Nikolay P. Nikolov[2] and Issam Zineh[1]

Liu et al. CPT doi:10.1002/cpt.3191

This was the first time that CDER used AI/ML for a regulatory decision, in this case to identify a population for a drug therapy.

1.1 Patient Population Identification

KINERET is authorized for emergency use for the treatment of COVID-19 in hospitalized adults with positive results of direct SARS-CoV-2 viral testing with pneumonia requiring supplemental oxygen (low- or high-flow oxygen) who are at risk of progressing to severe respiratory failure and likely to have an elevated plasma suPAR.

In the SAVE-MORE trial used to support the efficacy and safety of KINERET in COVID-19, key exclusion criteria were: pO2/FiO2 ratio < 150 mmHg, requirement for non-invasive ventilation (NIV), requirement for mechanical ventilation (MV), requirement for extra-corporeal membrane oxygenation (ECMO), and < 1500 neutrophils/mm^3.

All enrolled patients were required to have a plasma soluble urokinase plasminogen activator receptor (suPAR) level ≥ 6 ng/mL *[see Clinical Studies (14.1)]*. The suPAR assay is not commercially available in the United States. In order to identify a comparable population as was studied in the SAVE-MORE trial, an alternative patient identification method was developed to select patients most likely to have suPAR ≥ 6 ng/mL based on commonly measured patient characteristics. Patients meeting at least three of the following eight criteria are considered likely to have suPAR ≥ 6 ng/mL at baseline:

1. Age ≥ 75 years
2. Severe pneumonia by WHO criteria[1]
3. Current/previous smoking status
4. Sequential Organ Failure Assessment (SOFA)[2] score ≥ 3
5. Neutrophil-to-lymphocyte ratio (NLR) ≥ 7
6. Hemoglobin ≤ 10.5 g/dL
7. Medical history of ischemic stroke
8. Blood urea > 50 mg/dL and/or medical history of renal disease

Fact Sheet of Kineret ®

https://www.fda.gov/media/163546/download

https://www.fda.gov/drugs/spotlight-cder-science/using-machine-learning-identify-suitable-patient-population-anakinra-treatment-covid-19-under

Proprietary

FDA Qualified AI-Based Drug Development Tool (DDT)

FDA Qualifies First AI Drug Development Tool, Will Be Used in 'MASH' Clinical Trials

[12/8/2025] The U.S. Food and Drug Administration (FDA) has qualified the first AI drug development tool, the AI-Based Histologic Measurement of NASH (AIM-NASH), to help pathologists assess metabolic dysfunction-associated steatohepatitis (MASH) disease activity in clinical trials. This cloud-based tool helps pathologists score liver biopsy components, including fat infiltration (steatosis), inflammation (hepatocellular ballooning and lobular inflammation), and scarring (fibrosis) stage.

MASH, a significant public health challenge affecting millions of Americans, is a severe form of metabolic-associated fatty liver disease that develops when fat buildup in the liver causes inflammation and scarring. MASH can lead to cirrhosis (severe liver scarring), hepatic decompensation (worsening of liver function), liver cancer, liver transplantation, or death.

Currently in MASH clinical trials, multiple experts independently assess liver histology, a time-consuming process made complicated by variable scoring. AIM-NASH could help standardize histologic assessment and reduce the time and resources needed for MASH drug development.

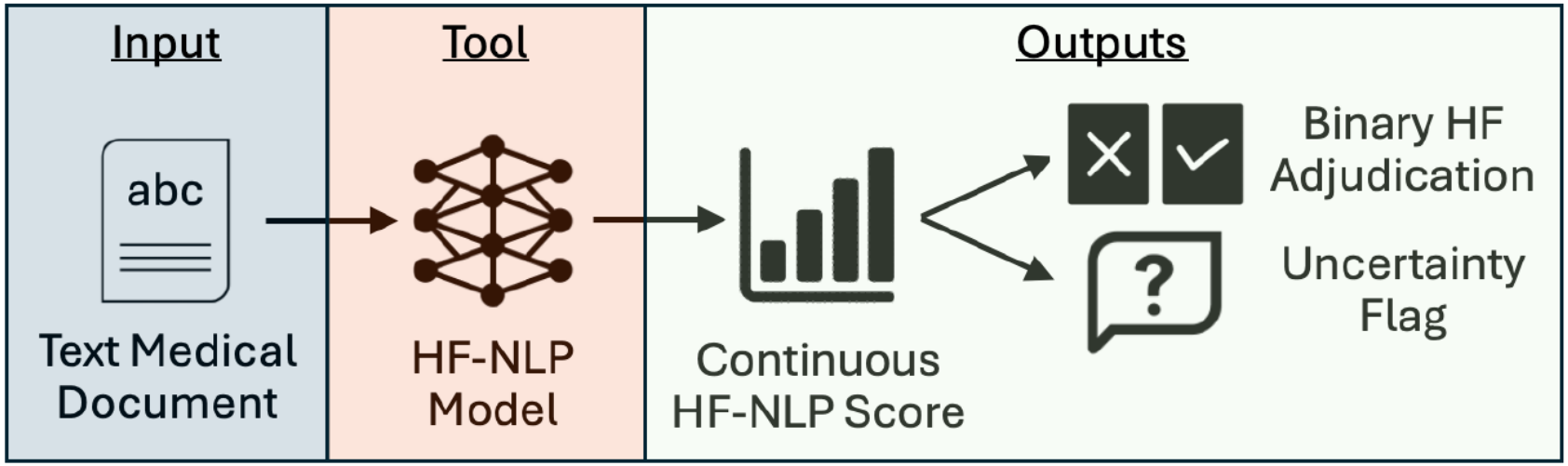
Proprietary

FDA Accepted Letter of Intent for AI/Natural Language Processing-Based DDT

FDA has completed its review of the Letter of Intent (LOI) for Drug Development Tool (DDT), DDT-IST-000027, received on April 10, 2024, by the CDER Innovative Science and Technology Approaches for New Drugs (ISTAND) Pilot Program, submitted under section 507 of the Federal Food, Drug, and Cosmetic (FD&C) Act.[1]

The LOI is for the Heart Failure Natural Language Processing (HF-NLP) model.

FDA has made a determination to **accept** the LOI into the CDER ISTAND Pilot Qualification Program. Note that projects in the ISTAND Pilot Qualification Program may be eligible to apply for DDT grant programs which are offered for further DDT tool development as funding permits.[2]

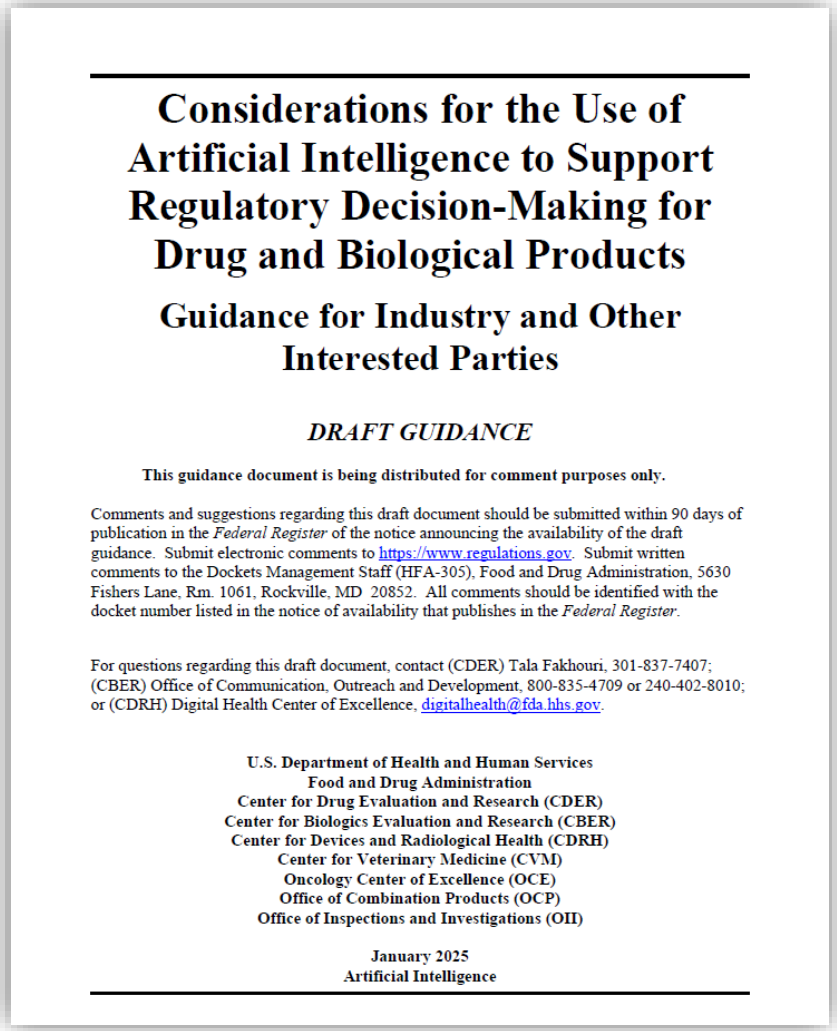


https://force-dsc.my.site.com/ddt/s/ddt-project?ddtprojectid=205

Proprietary

Considerations for AI Use in Regulatory Decision Making

- Draft Guidance published in Jan 2025
- AI used to produce data or information to support regulatory decision making
- Informed by:
  - FDA's experience with reviewing over 300 submissions with AI
  - Over 800 comments received on the 2023 and 2024 discussion papers
  - Current regulatory science research
- Provides a risk-based framework for establishing and evaluating the credibility of AI use in regulatory decision making
- Help ensure that AI models used to answer regulatory questions are credible for a particular context of use and are supported with the appropriate level of evidence



Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry and Other Interested Parties

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Tala Fakhouri, 301-837-7407; (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010; or (CDRH) Digital Health Center of Excellence, digitalhealth@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)
Center for Veterinary Medicine (CVM)
Oncology Center of Excellence (OCE)
Office of Combination Products (OCP)
Office of Inspections and Investigations (OII)

January 2025
Artificial Intelligence

Source: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

Proprietary

Considerations for AI Use in Regulatory Decision Making

- **Step 1:** State the question of interest that will be addressed by the AI model
- **Step 2:** Define the context of use for the AI model
- **Step 3:** Assess the AI model risk
  - Model influence:
    - The contribution of the evidence derived from the AI model relative to other contributing evidence used to inform the question of interest
  - Decision Consequence:
    - Describes the significance of an adverse outcome resulting from an incorrect decision concerning the question of interest
- **Step 4:** Develop a plan to establish AI model credibility within the context of use
- **Step 5:** Execute the plan
- **Step 6:** Document the results of the credibility assessment and discuss deviations from the plan
- **Step 7:** Determine the adequacy of the credibility assessment

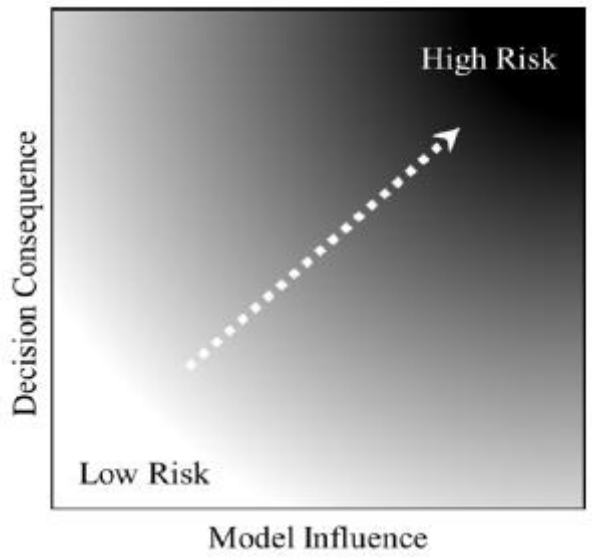


Figure 1. Model risk matrix. The model risk moves from low to high as decision consequence or model influence increases. The ratings for decision consequence and model influence are independently determined.

The credibility assessment activities used to establish the credibility of AI model outputs should be commensurate with the AI model risk.

Source: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

Proprietary

Determine the adequacy of the credibility assessment

- If either the sponsor or FDA determine that *model credibility is not sufficiently established for the model risk*, several outcomes are possible:
  - The sponsor *may downgrade the model influence by incorporating additional types of evidence* in conjunction with the evidence from the AI model to answer the question of interest;
  - The sponsor *may increase the rigor of the credibility assessment activities* or augment the model's output by adding additional development data;
  - The sponsor *may establish appropriate controls to mitigate risk*;
  - The sponsor may consider the credibility of the AI model's output inadequate for the COU; therefore, *the model's COU would be rejected or revised in an iterative fashion*.

Source: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

Proprietary

Hypothetical Case Study: LLM-Based Clinical Endpoint Assessments

- **Motivation:** Clinical endpoints requiring human evaluation (e.g., clinical interviews, medical note reviews) can be affected by variability in how raters interpret, extract, and score patient information. Differences across raters, sites, and assessment sessions may introduce inconsistency and reduce the reliability of endpoint measurement and adjudication.

- **COU**: A sponsor proposes to use an LLM-based evaluation tool to process clinical source data (e.g., patient interview transcripts, medical notes) and generate scores or adjudications for a clinical outcome assessment. The model output may be used in different ways, for example:

  - As an **assistive tool** for trained human raters or adjudicators
  - as the primary source of the endpoint score or adjudication without human oversight

*This study is intended to illustrate general considerations for evaluating whether an AI/LLM model credibility assessment is commensurate with the model risk for a proposed context of use. It does not describe any specific submission, sponsor, product, indication, clinical program, or regulatory interaction.

Proprietary

Model Risk Analysis

- Model influence:
  - Low to Moderate. Human oversight is required for the LLM-Based Clinical Endpoint Assessments.
  - **High**. No human oversight is proposed in using the LLM-Based Clinical Endpoint Assessments. It is the ***sole*** factor to inform the question of interest.

- Decision consequence (If LLM makes inaccurate decision):
  - **Low to Moderate in Phase II,** as it is not intended to support regulatory approval.
  - **High** in Phase III, as it is intended to support regulatory approval.

Proprietary

Recommendations on establishment of the credibility assessment

- What if credibility assessment activities are not sufficient to be commensurate with the model risk of using LLM scorer as a primary endpoint assessment?
  - *In general, when model credibility assessment is not sufficiently established for the model risk[1], several outcomes are possible:*
    - *Downgrade the model influence by incorporating additional types of evidence in conjunction with the evidence from the AI model to answer the question of interest (i.e., a human-in-the loop approach).*

    For example, the LLM might not serve as the sole factor supporting the primary endpoint, with human raters continuing to play a role in conducting the primary endpoint assessment. By downgrading the model influence, the overall model risk may be reduced.
    - *Increase the rigor of the credibility assessment activities (i.e., additional validation dataset(s))*
    - *Establish appropriate controls to mitigate risk (i.e., the quality control measures)*

1. Page 15: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

Proprietary

Review Example 2: ML for Patients Risk Stratification

- **COU:** In previous trials of Drug A, all patients went through inpatient monitoring after dosing due to concern of a potentially life-threatening AE. The sponsor proposed to use a ML model to predict a patient's risk for this AE based on baseline characteristics and lab values. If predicted to be low risk, the patient will have outpatient monitoring instead of inpatient monitoring.

  Initially the sponsor was not clear on some components of COU (e.g., in clinical trials only or in both trials and real-world application?). FDA sent them relevant comments.

  The sponsor also mentioned that this model could also be used for dosing regimen optimization. If so, it will need a different model risk analysis.

Proprietary

Review Example 2: ML for Patients Risk Stratification (continued)

FDA's model risk analysis:

- **Model influence: High.** The model prediction will be the sole factor to determine whether a patient will go through inpatient or outpatient monitoring after dosing.
- **Decision consequence: High.** The model prediction will determine whether a patient will go through inpatient or outpatient monitoring after dosing
  - If the model erroneously predicts a low risk patient to be high risk, it may not be of major concern, as this patient will go through inpatient monitoring even though it might be unnecessary.
  - If the model erroneously predicts a high risk patient to be low risk, it will be a major concern, as this patient will not have the necessary inpatient monitoring for the potentially life-threatening AE.

Negative predictive value, i.e. probability (No AE/Negative Model prediction) and sensitivity, will be the major performance metrics to determine the acceptance of the model from a regulatory perspective.

Proprietary

Review Example 3 : ML-based Enrichment Trial Design

COU:

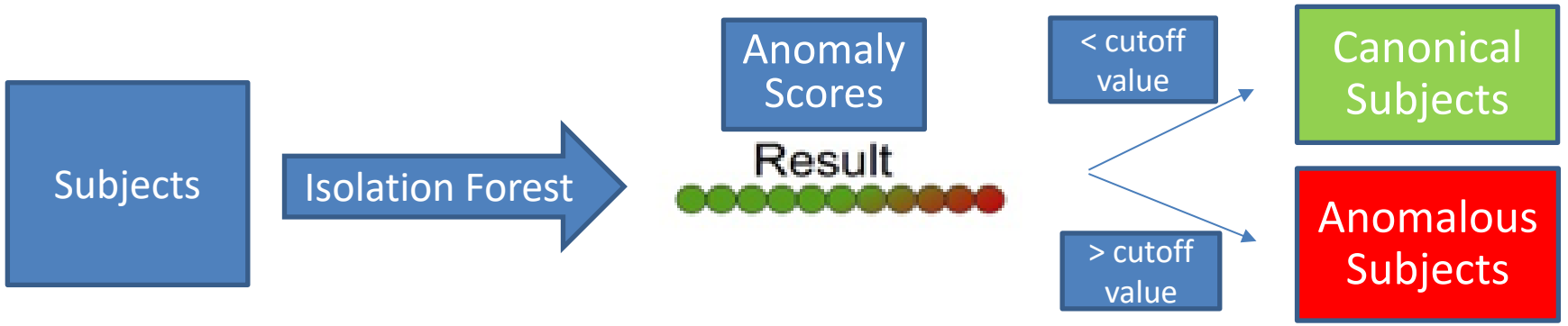
- Drug C failed to reach statistical significance in prespecified primary efficacy analysis for the phase 2 trial.
  - Unexpectedly large improvement in the primary efficacy endpoint from baseline for the placebo-treated cohort.
    - This primary efficacy endpoint was a composite endpoint with many items.
  - Motivation to explore ML method to identify patients with abnormal patterns.

- The sponsor is proposing an ML-based enrichment strategy for two planned Phase 3 studies
  - Decrease inter-subject variability prior to randomization:
    - Select canonical subjects whose symptoms are characterized by greater similarity to the typical patient in the target population.

Proprietary

Review Example 3 : ML-based Enrichment Trial Design (continued)

- The dataset for ML development includes:
  - The phase II trial of Drug C
  - Multiple trials of Drug D for the same indication
- The anomaly score is calculated by applying a trained isolation forest model on the components of the efficacy endpoint at screening and baseline.
  - The anomaly score for a given subject indicates how easily this subject can be isolated from the rest of the subjects.



Proprietary

Example 3: ML-based Enrichment Trial Design (continued)

- Encouraging trend was observed when ML-based inclusion criterion was applied retrospectively to the phase II data of drug C
  - Larger effect size

- In general, similar trends were observed when it was applied to trials of drug D

Proprietary

Example 3: ML-based Enrichment Trial Design (continued)

- FDA's major comment to sponsor

  The sponsor proposed to use the ML-based inclusion criterion in two phase III adequate and well-controlled studies.

  FDA:

  Suggest using the ML-based inclusion criterion in only one of the two planned adequate and well-controlled studies.

  Due to the potential challenges related to the **explainability** of the model and the **generalizability** of the results, it would be helpful to include the "anomalous" patients in at least one of the adequate and well-controlled studies.

Proprietary

Thank You!