Machine-Learning-Derived Enrichment Markers in Clinical Trials

David H. Millis, MD, MBA, PhD
Medical Officer, Division of Psychiatry
Center for Drug Evaluation and Research
U.S. Food and Drug Administration
February 20, 2020

Disclaimers

This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

Financial disclosures: none.

Morning Sessions: A Shared Theme

All four of the previous speakers discussed the use of machine learning methods for identifying clinically meaningful subpopulations of patients.

Dr. Ahmed	Improving clinical trial recruitment by identifying geographically-dispersed patients who share characteristics that make them likely to benefit from the drug
Dr. Geraci	Interactive process for identifying meaningful patient subpopulations by examining the variables that a ML algorithm uses to separate patients into subgroups
Dr. Tiller	• A ML algorithm to classify patients by likelihood of response to a treatment for depression • A separate ML algorithm to generate a set of rules to provide descriptions of the two subpopulations that would be meaningful to clinicians
Dr. Wall	Comparison of several machine learning algorithms that used a database of videos tagged by behavioral features to learn how to distinguish children with typical behavioral development from children with autism spectrum disorder

Previous FDA Approvals Related to ML

- Typically have involved devices and/or software that learn to assign patient data into known, previously-established categories that can be verified by a human expert
  - several examples of products that learn to distinguish normal from abnormal images (MRI, CT scans, mammograms)
  - system performance can be assessed by comparisons to a radiologist's interpretation of images in the training set
- No examples of an approval in which a ML algorithm identified a novel, previously unrecognized subpopulation of patients, with subsequent approval of a drug for use in that subpopulation

Motivation for This Presentation

- The previous talks raise interesting questions about how the FDA would provide oversight for a drug development program in which a machine learning algorithm has a key role in identifying the target population for the drug.
- Little experience in FDA with evaluating study protocols that use ML-based methods for enriching the study population in a clinical trial.
- My aim today: to point out some issues that would be part of our thinking in the event that a sponsor submits a proposal to incorporate machine learning methods into the inclusion criteria for a clinical trial.

A few caveats…

- This presentation should not be considered to represent FDA-approved industry guidance on the use of ML-based classifiers in drug development. Currently there are no FDA guidances that explicitly cover this topic.
- Sponsors considering the use of ML-based classifiers in drug development should seek consultation from the FDA during the earliest stage possible in the development program.

Outline

1. Overview of the concept of enrichment
2. Regulatory issues raised by enrichment strategies based on machine learning algorithms

[1] ENRICHMENT: OVERVIEW

[2] REGULATORY ISSUES FOR ENRICHMENT BASED ON MACHINE-LEARNING MODELS

Enrichment

- Definition*:
  - The prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is in fact present) is more likely than it would be in an unselected population
- Purpose:
  - To make it easier to demonstrate a drug effect
  - To facilitate better matching of patients to treatments once the drug enters clinical practice

** Source: FDA guidance, "Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products," March 2019.*

Enrichment Strategies

- Strategies to decrease heterogeneity
  - To reduce variability in the study population
- Prognostic enrichment strategies
  - Choosing patients with a greater likelihood of having a disease-related endpoint event or a substantial worsening in condition
- Predictive enrichment strategies
  - Choosing patients more likely to respond to the drug than other patients with the condition being treated

Strategies to Decrease Heterogeneity

- Defining entry criteria to ensure that patients in the study actually have the disease
- Making efforts to remove placebo responders prior to randomization
- Decreasing intra-patient variability by enrolling only patients who give consistent values on baseline assessments

Prognostic Enrichment Strategies

- Selecting patients with a greater likelihood of having a clinical event or a large change in a continuous measure
  - This allows a treatment effect to be more readily detected
  - Example: selecting patients with high risk of cancer recurrence may make it easier to detect the effect of a cancer treatment

Predictive Enrichment Strategies

- Identify patients more likely to respond to a particular intervention
  - measurement of a biomarker (genomic, proteomic) related to the study drug's mechanism
  - pathophysiological strategies:
    - measuring the patient's ability to metabolize a prodrug to its active metabolite
    - determining whether the patient's cells produce the molecular target of the drug
  - previous history of response to a drug in the same pharmacologic class
  - documented response in an open pre-randomization period in a randomized-withdrawal trial
  - factors identified in results from previous studies

Predictive Enrichment and Benefit-Risk Relationship

- Identification of a responder population can enhance the benefit-risk relationship of the drug by avoiding exposure and potential toxicity in people who are unlikely to benefit from the drug
- For drugs with significant toxicity and low overall response in a general population, identifying a responder population could make the risk more acceptable and facilitate continued drug development and approval

Studying the Marker-Negative Population

- Learning how the drug affects the marker-negative population can be useful:
  - Is treatment effect completely absent or just smaller in the marker-negative population?
  - Is safety profile the same or different in the marker-negative population?
  - If the marker-positive population is small compared to the marker-negative population, clinicians will more often have to decide whether to prescribe the drug for a marker-negative individual than for a marker-positive individual.

- Benefit-risk analysis for the marker-negative population:
  - Are treatment options equally constrained for both subpopulations?
  - Should use of the drug in the marker-negative population be permitted, discouraged, or contraindicated?
    - The answer depends on directly studying the marker-negative population.

Limiting study of the marker-negative population may be justified…

- There is a pathophysiological basis for concluding that the marker-negative population will not respond to the drug
  - for example: patients lack the molecular target for the drug
- Early studies show no treatment response in the marker-negative population

[1] ENRICHMENT: OVERVIEW

[2] REGULATORY ISSUES FOR ENRICHMENT BASED ON MACHINE-LEARNING MODELS

Which Enrichment Strategies Can be ML-Based?

- Machine learning algorithms could have a role in all three types of enrichment strategies: decreasing heterogeneity, prognostic, predictive
- Strategies for decreasing heterogeneity will likely have more relevance to clinical trial design than to clinical practice
- Prognostic and predictive strategies could have a role in both clinical trial design and in clinical practice, since they define characteristics of patients most likely to benefit from the drug

Performance of the Enrichment Classifier: Shape of the Classification Boundary

- How well does the shape of the boundary separating patient groups generalize from a small study population to the larger population encountered in clinical practice?
  - A model generated by analysis of a small research dataset may be subject to either underfitting or overfitting the data, and can suggest a boundary between subpopulations that may not be representative of the larger population of patients seen in clinical practice
    - High bias (underfitting): the model is overly simple
    - High variance (overfitting): the model matches the small population too closely; may be capturing noise in the data
  - Either of these can result in misclassification of patients in clinical practice

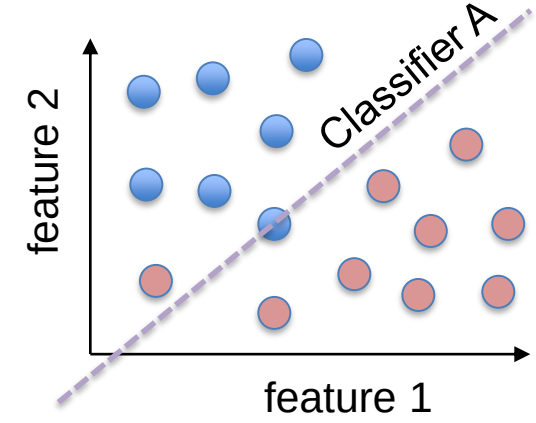
Performance of the Enrichment Classifier: Setting the Classification Threshold

- How far from the classifier boundary should a patient be in order for us to be certain that the patient belongs in one subpopulation or the other?
  - Adjusting this threshold will change the specificity and sensitivity of the classifier
  - If classifier lacks specificity: it is overly inclusive
    - The estimated difference between the effect in the enriched and non-enriched populations will be attenuated
    - Defeats the goal of the enrichment strategy
  - If the classifier lacks sensitivity: it is overly exclusive
    - Patients who could benefit from the drug will not be studied
    - Study subjects may be difficult to find
  - Setting the optimal cutoff to separate patient subpopulations may be difficult, depending on the anticipated tradeoffs between sensitivity and specificity

Performance of the Enrichment Classifier: How to Evaluate Over Time?

- Performance of the classifier may require reassessment over the course of the development program
  - For example: the model generated from a Phase 3 dataset may be different from the model generated earlier from a smaller Phase 2 dataset
- Question: if the model is complex and does not have an obvious clinical interpretation:
  - What metric do we use to assess how well the classifier assigns patients to the correct subpopulation?
    - Different task from assigning patients to the responder or non-responder group.
  - If the model changes over time, how can we tell whether the model is improving or being more distracted by noise in the data?

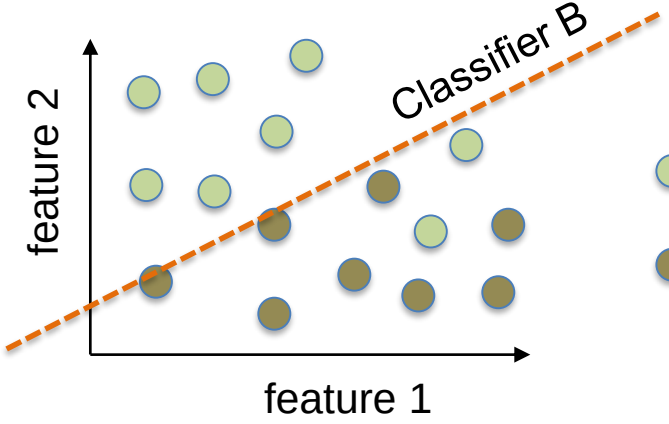
Model Interpretability



= responders

= non-responders

"Interesting… most of the patients above the line have just depression, while a lot of the patients below the line have both depression and anxiety. How would the classifier look if we label patients by these two categories?"



= depression + anxiety

= depression only

"The new model is very similar. In clinical practice, maybe assessing for both depression and anxiety can identify the patients most likely to respond to the drug."

Model Interpretability and Labeling

- Interpretability of the ML model may affect the regulatory path towards creating the product label.
- Question: Once the ML algorithm creates a model that identifies a potential target subpopulation, can we identify that subpopulation by some means other than using the ML-based model?
  - **If yes:**
    - the label might include minimal reference to the ML algorithm
    - The Indication section would describe the target population in terms that could be easily interpreted by a clinician
  - **If no:**
    - Running the new patient's data through the ML model may be a requirement for identifying patients for whom the drug will be effective
    - In this case, the ML model may take the role of a **companion diagnostic** for the drug
    - The regulatory pathway for the drug will include both demonstrating **efficacy of the drug** and demonstrating that the **companion diagnostic can identify patients** in the target subpopulation
    - Potential problem: If we don't have a clear understanding of the pattern in the patient data that the ML model is detecting, how can we measure the performance of the companion diagnostic? What will we measure its performance against?

ML-Based Enrichment Marker as a Companion Diagnostic

- Discussion of the possible use of an ML-based enrichment marker as a companion diagnostic for purposes of clinical trial enrichment should occur early in the drug development process
- No single current FDA guidance explicitly covers this situation, but several guidances discuss principles that would be relevant to such a development program:
  - "Qualification Process for Drug Development Tools," December 2019
  - "In Vitro Companion Diagnostic Devices," August 2014
  - "Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product," July 2016

Intended Population: All Patients, or Just Marker-Positive Patients?

- Whether the drug is approved only for patients matching the enriched population is not a simple decision
  - A study on an enriched population could support the approval of the drug for use in any patient with the disease – i.e., both marker-positive and marker-negative patients
  - If there is strong evidence that the drug would be ineffective or pose unacceptable risks in the non-enriched population, it may be necessary for the label to define an intended population that matches the characteristics of the enriched population studied in the clinical trials
  - However – describing the intended population in labeling may be difficult if the classifier defines a patient subpopulation that cannot be described in a way that can be easily interpreted by clinicians
- A clinically interpretable ML model may simplify labeling if it is important for clinicians to avoid giving the drug to patients outside the enriched population

Potential Bias in Model Development

- *(*Note: here we are talking not about bias vs variance, but bias vs neutrality.)*
- The ML model may have unanticipated biases based on:
  - Demographics and clinical characteristics of patients in the training set
  - The assessment tools selected by the investigators for inclusion in the classifier
- Tradeoff between explainability and bias:
  - using a small number of variables to build the model can reduce complexity and foster explainability
  - but the decision to leave some variables out of the model-building process may introduce investigator bias about which factors are likely to be important in defining the patient classifications
  - in the study protocol, justification for including or omitting certain variables in model training may help identify sources of bias
- Questions about how to help clinicians address possible bias in the model:
  - How might the composition of the training set be conveyed to clinicians?
  - Would the composition of the training set be included somewhere in the product labeling?
  - Would the label include guidance on how the clinician should proceed if a new patient does not match the characteristics of the training set?

Conclusions

- Use of enrichment designs can increase the likelihood of observing a treatment effect in clinical trials.
- The performance characteristics of ML-based enrichment markers must be understood both for use in clinical trials and for deployment in clinical practice.
- Clinicians may need guidance on how to use ML models to match individual patients to the population studied for the drug approval.
- Interpretability and complexity of ML-based models may have an impact on labeling and on whether the ML model itself will need to be evaluated as part of the drug approval process.
- Discussions with FDA on ML-based models to be used in clinical trial design or as companion diagnostics should occur early in the drug development process.


FDA
U.S. FOOD & DRUG
ADMINISTRATION