Towards a Post-Market Monitoring Framework for Machine Learning-based Medical Devices: A case study

Jean Feng, Adarsh Subbaswamy, Alexej Gossmann, Harvineet Singh, Berkman Sahiner, Mi-Ok Kim, Gene Pennello, Nicholas Petrick, Romain Pirracchio, Fan Xia

Workshop on Regulatable Machine Learning, NeurIPS 2023

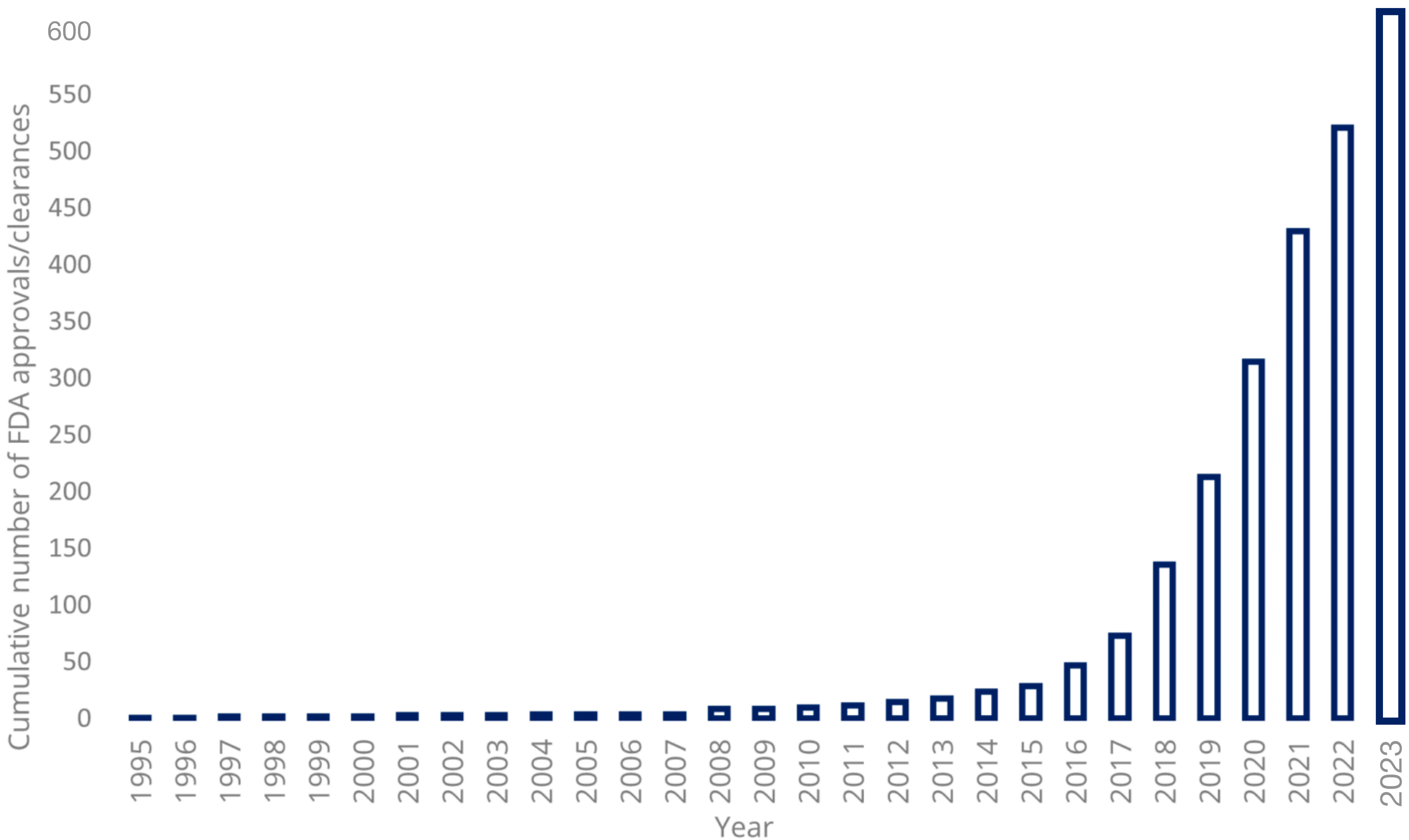
University of California San Francisco

FDA

CERSI UCSF-Stanford

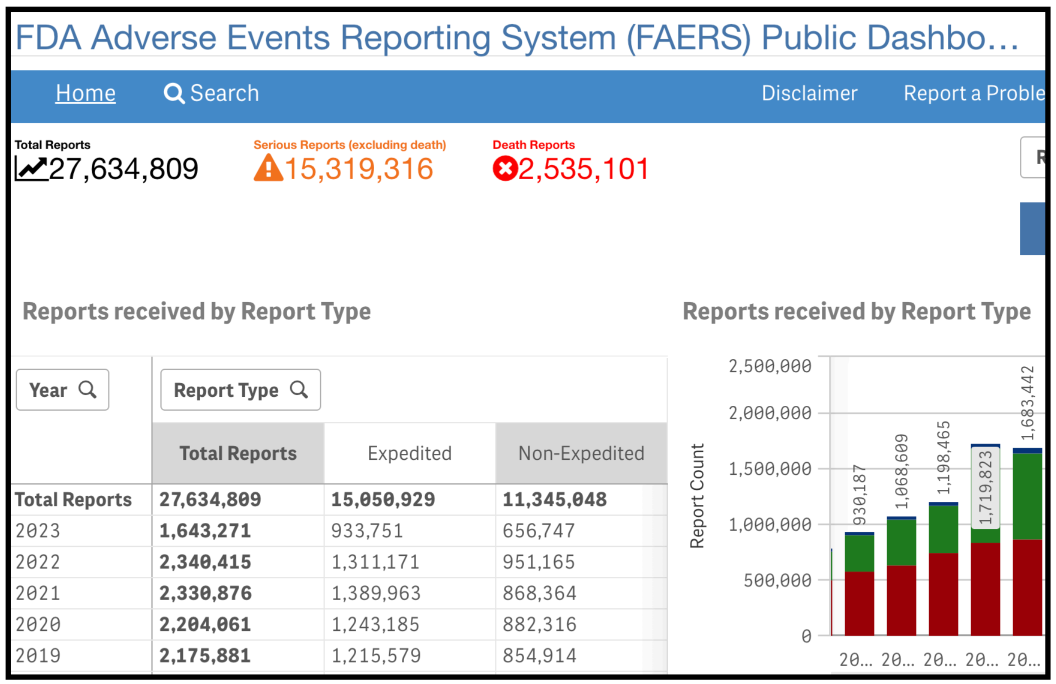
Disclaimer: The views presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of the FDA/HHS, PCORI, or the U.S. Government.

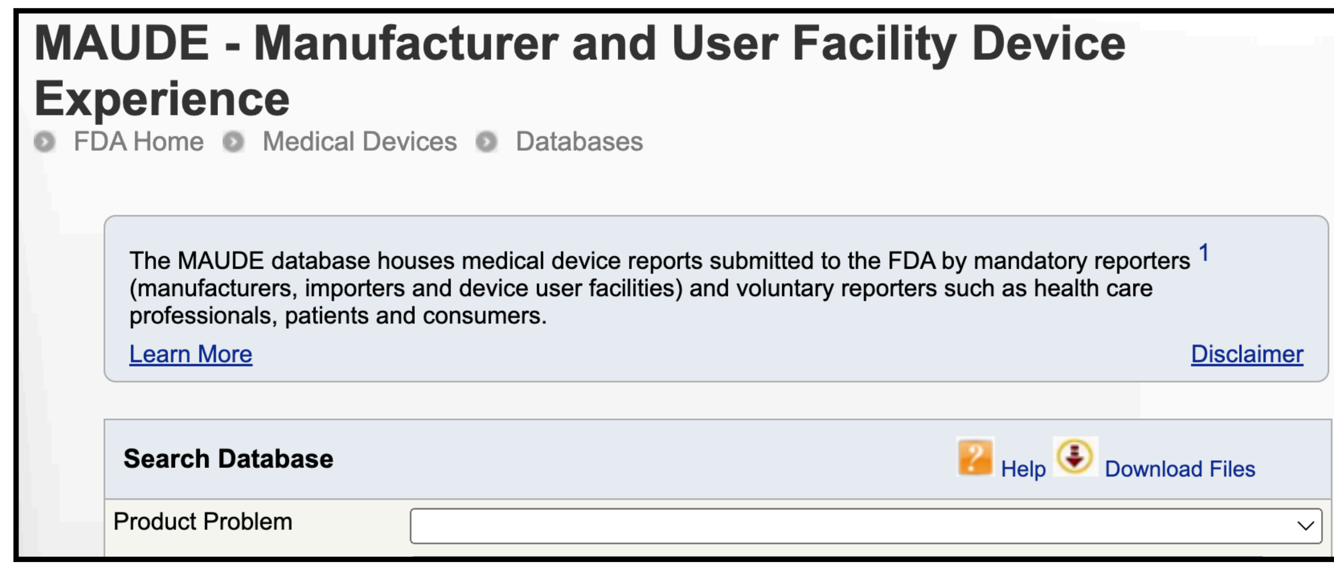
FDA approvals of AI/ML-Enabled Medical Devices



Source: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

Post-market surveillance/ reporting systems





Sentinel

MedWatch
www.FDA.gov/MedWatch

The regulatory landscape

U.S. FOOD & DRUG ADMINISTRATION | Health Canada Santé Canada | Medicines & Healthcare products Regulatory Agency

Guiding Principles

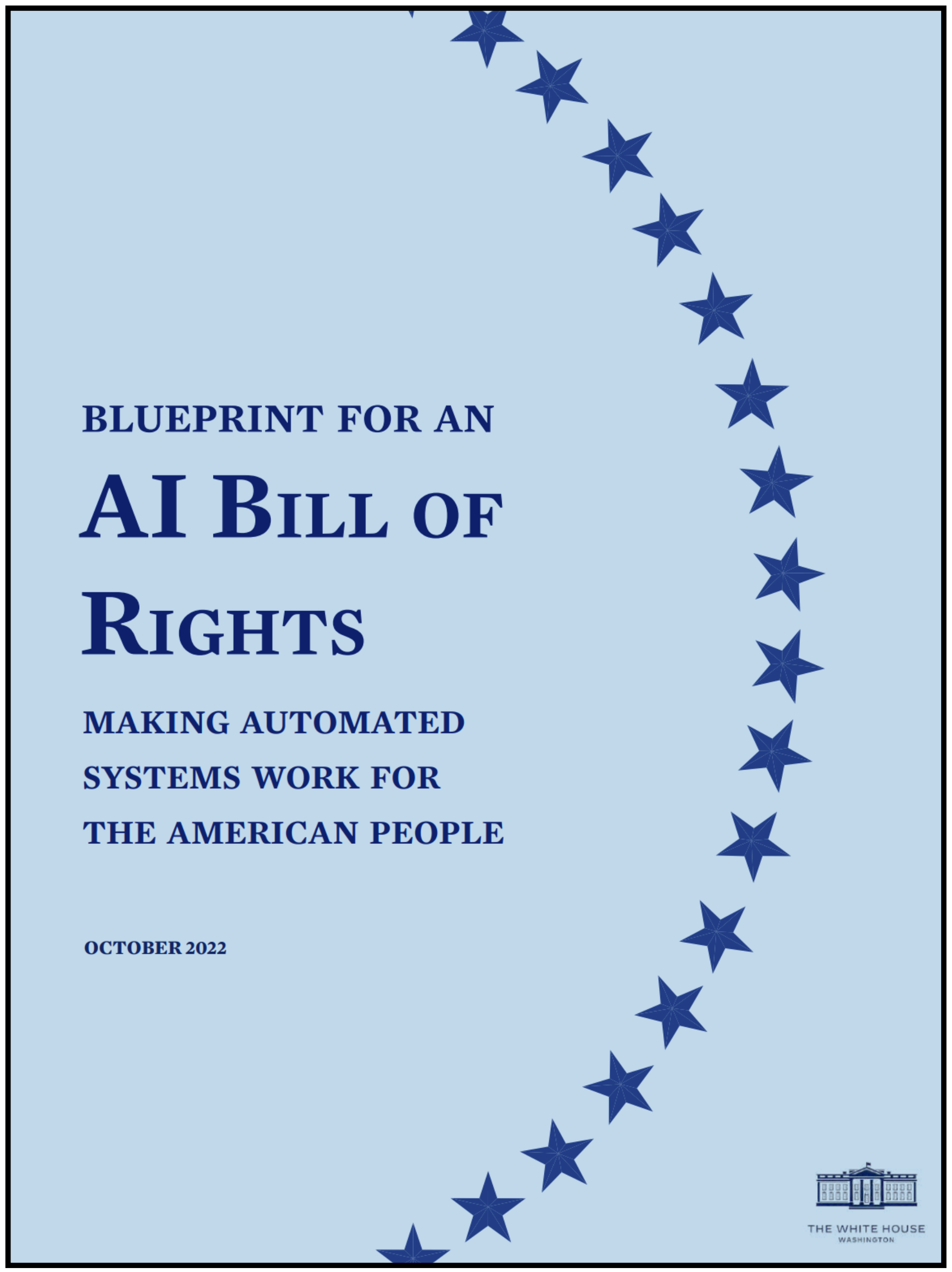
1. **Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle:** In-depth understanding of a model's intended integration into clinical workflow, and the desired benefits and associated patient risks, can help ensure that ML-enabled medical devices are safe and effective and address clinically meaningful needs over the lifecycle of the device.
2. **Good Software Engineering and Security Practices Are Implemented:** Model design is implemented with attention to the "fundamentals": good software engineering practices, data quality assurance, data management, and robust cybersecurity practices. These practices include methodical risk management and design process that can appropriately capture and communicate design, implementation, and risk management decisions and rationale, as well as ensure data authenticity and integrity.
3. **Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population:** Data collection protocols should ensure that the relevant characteristics of the intended patient population (for example, in terms of age, gender, sex, race, and ethnicity), use, and measurement inputs are sufficiently represented in a sample of adequate size in the clinical study and training and test datasets, so that results can be reasonably generalized to the population of interest. This is important to manage any bias, promote appropriate and generalizable performance across the intended patient population, assess usability, and identify circumstances where the model may underperform.
4. **Training Data Sets Are Independent of Test Sets:** Training and test datasets are selected and maintained to be appropriately independent of one another. All potential sources of dependence, including patient, data acquisition, and site factors, are considered and addressed to assure independence.
5. **Selected Reference Datasets Are Based Upon Best Available Methods:** Accepted, best available methods for developing a reference dataset (that is, a reference standard) ensure that clinically relevant and well characterized data are collected and the limitations of the reference are understood. If available, accepted reference datasets in model development and testing that promote and demonstrate model robustness and generalizability across the intended patient population are used.
6. **Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device:** Model design is suited to the available data and supports the active mitigation of known risks, like overfitting, performance degradation, and security risks. The clinical benefits and risks related to the product are well understood, used to derive clinically meaningful performance goals for testing, and support that the product can safely and effectively achieve its intended use. Considerations include the impact of both global and local performance and uncertainty/variability in the device inputs, outputs, intended patient populations, and clinical use conditions.
7. **Focus Is Placed on the Performance of the Human-AI Team:** Where the model has a "human in the loop," human factors considerations and the human interpretability of the model outputs are addressed with emphasis on the performance of the Human-AI team, rather than just the performance of the model in isolation.
8. **Testing Demonstrates Device Performance During Clinically Relevant Conditions:** Statistically sound test plans are developed and executed to generate clinically relevant device performance information independently of the training data set. Considerations include the intended patient population, important subgroups, clinical environment and use by the Human-AI team, measurement inputs, and potential confounding factors.
9. **Users Are Provided Clear, Essential Information:** Users are provided ready access to clear, contextually relevant information that is appropriate for the intended audience (such as health care providers or patients) including: the product's intended use and indications for use, performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model. Users are also made aware of device modifications and updates from real-world performance monitoring, the basis for decision-making when available, and a means to communicate product concerns to the developer.
10. **Deployed Models Are Monitored for Performance and Re-training Risks Are Managed:** Deployed models have the capability to be monitored in "real world" use with a focus on maintained or improved safety and performance. Additionally, when models are periodically or continually trained after deployment, there are appropriate controls in place to manage risks of overfitting, unintended bias, or degradation of the model (for example, dataset drift) that may impact the safety and performance of the model as it is used by the Human-AI team.

Good Machine Learning Practice for Medical Device Development: Guiding Principles

(FDA 2021)

"Deployed Models Are Monitored for Performance and Re-training Risks are Managed"

The regulatory landscape

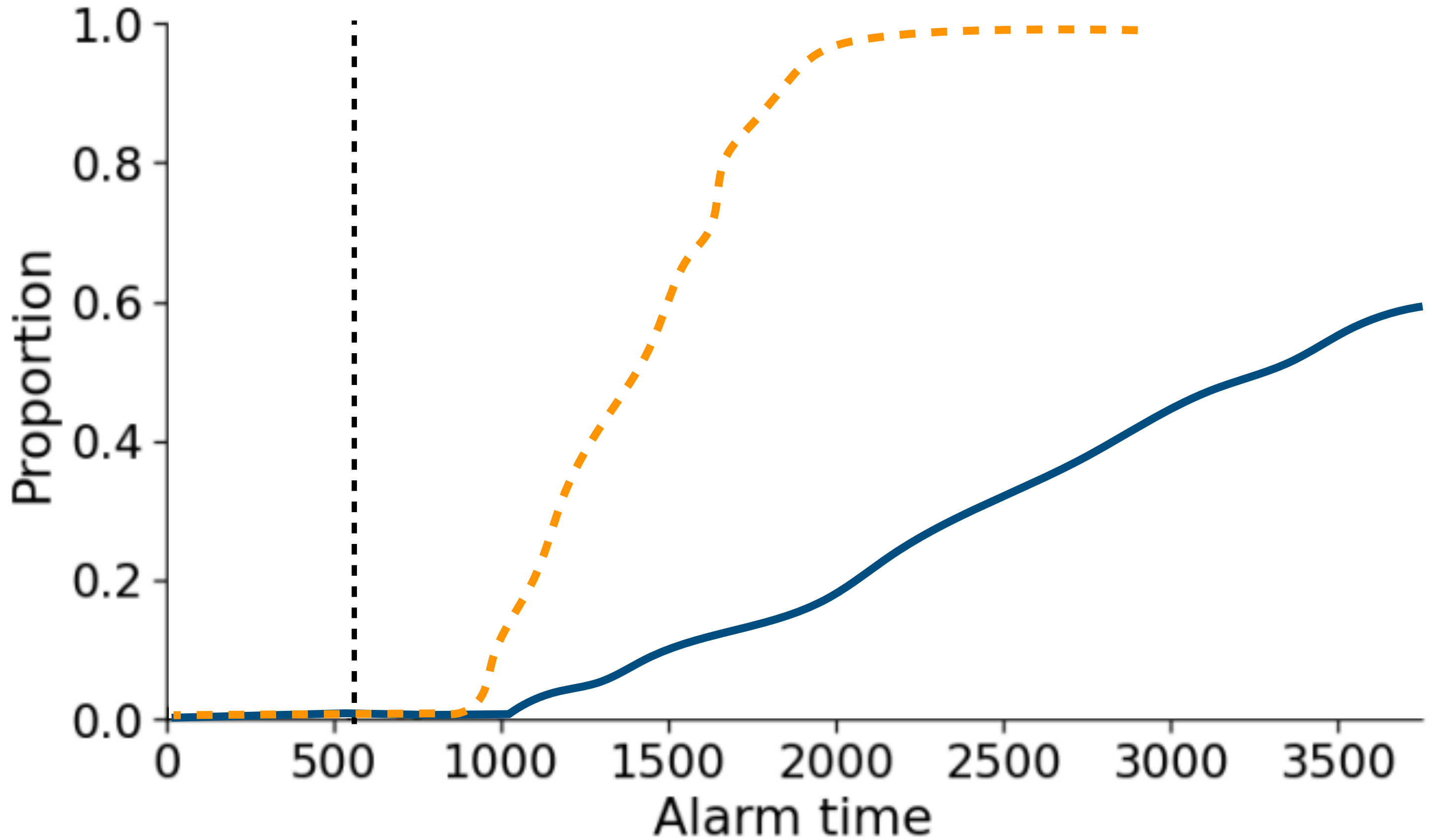


"Automated systems should have **ongoing monitoring procedures**… in place to ensure that their performance does not fall below an acceptable level over time, based on changing real-world conditions or deployment contexts, post-deployment modification, or unexpected conditions."

What's so hard about monitoring?

- A common proposal is to monitor the same metrics used for initial model approval. However, model monitoring is ***not simply*** model evaluation.
- Consider a model that was initially approved based on its negative and positive predictive values (NPV/PPV). We could try to monitor based on:
  - Option —: the same metrics of NPV/PPV
  - Option --: strong calibration

} *Q1: What is the monitoring criterion?*



*The goal of model monitoring is **detect performance decay as quickly as possible,** so to minimize the number of individuals exposed to a defective product.*

What’s so hard about monitoring?

Observational data: Easy to collect, but exhibits many potential sources of bias. ML algorithm itself may be a major source of bias.

Interventional data: Harder to collect, but can explicitly eliminate biases.

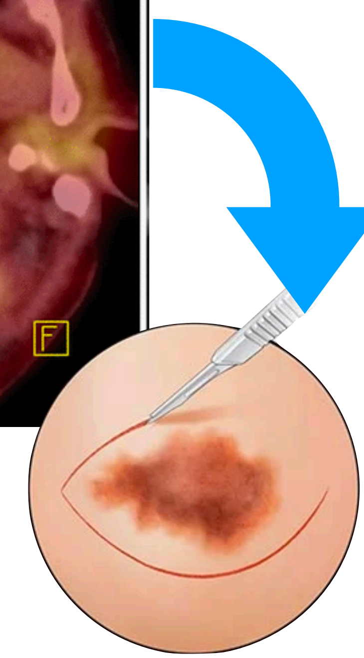
Q2: What data should we analyze/collect?

Q3: What assumptions are required?

Clinical risk prediction algorithm

Interfering medical interventions (IMI)

Diagnostic algorithm



Verification bias

A systematic framework is needed

How can we answer the many design questions, e.g.

Q1: What is the monitoring criterion?

Q2: What data should we analyze/collect?

Q3: What assumptions are required?

Our workshop paper takes the first steps towards building out ***a post-market monitoring framework*** *that brings together tools from* ***causal inference*** *and* ***statistical process control****.*

***Case study:** Risk prediction algorithm*

A postmarket monitoring framework

1. Define potential monitoring criteria

2. Enumerate sources of bias and define the causal model

3. Describe candidate monitoring strategies

4. Compare the pros and cons of candidate strategies

***Case study:** Risk prediction algorithm*

A postmarket monitoring framework

1. Define potential monitoring criteria

Criterion 1: NPV/PPV levels are maintained

$$H_0: \begin{cases} \Pr(Y_t(a) = 0 \mid \hat{y}_t(X_t) = 0) \geq c_{a0} \\ \Pr(Y_t(a) = 1 \mid \hat{y}_t(X_t) = 1) \geq c_{a1} \end{cases}$$

Criterion 2: NPV/PPV levels within subgroups are maintained

Criterion 3: Strong calibration is maintained

2. Enumerate sources of bias and define the causal model

3. Describe candidate monitoring strategies

4. Compare the pros and cons of candidate strategies

***Case study:** Risk prediction algorithm*

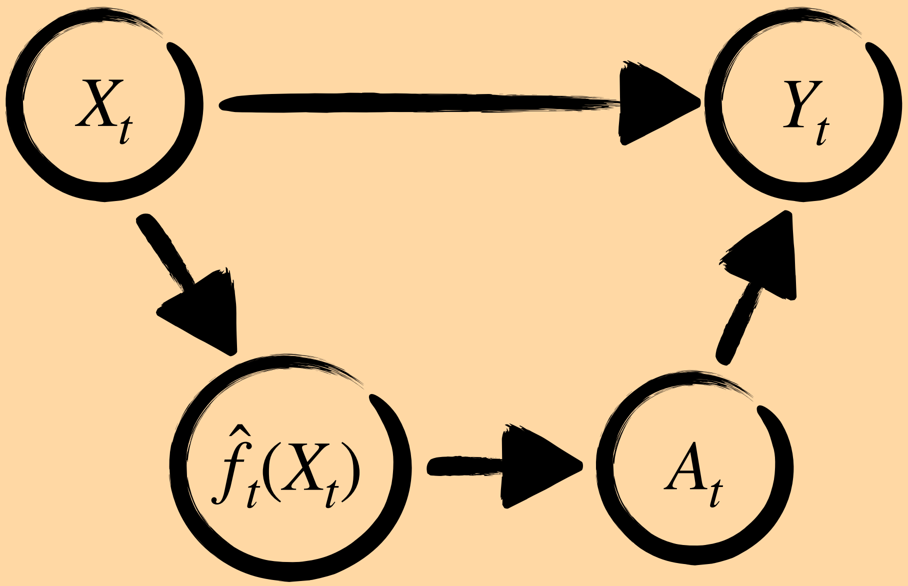
A postmarket monitoring framework

1. Define potential monitoring criteria

2. Enumerate sources of bias and define the causal model

<u>Potential Biases in Observational Data</u>

- Spectrum bias
- Off-label Use
- Interfering Medical Interventions (IMI)
- Circular Definitions
- …



3. Describe candidate monitoring strategies

4. Compare the pros and cons of candidate strategies

Case study: Risk prediction algorithm

A postmarket monitoring framework

1. Define potential monitoring criteria

2. Enumerate sources of bias and define the causal model

3. Describe candidate monitoring strategies

Criterion ***Data Source***

{1, 2, 3} x {Observational, Interventional}

4. Compare the pros and cons of candidate strategies

***Case study:** Risk prediction algorithm*

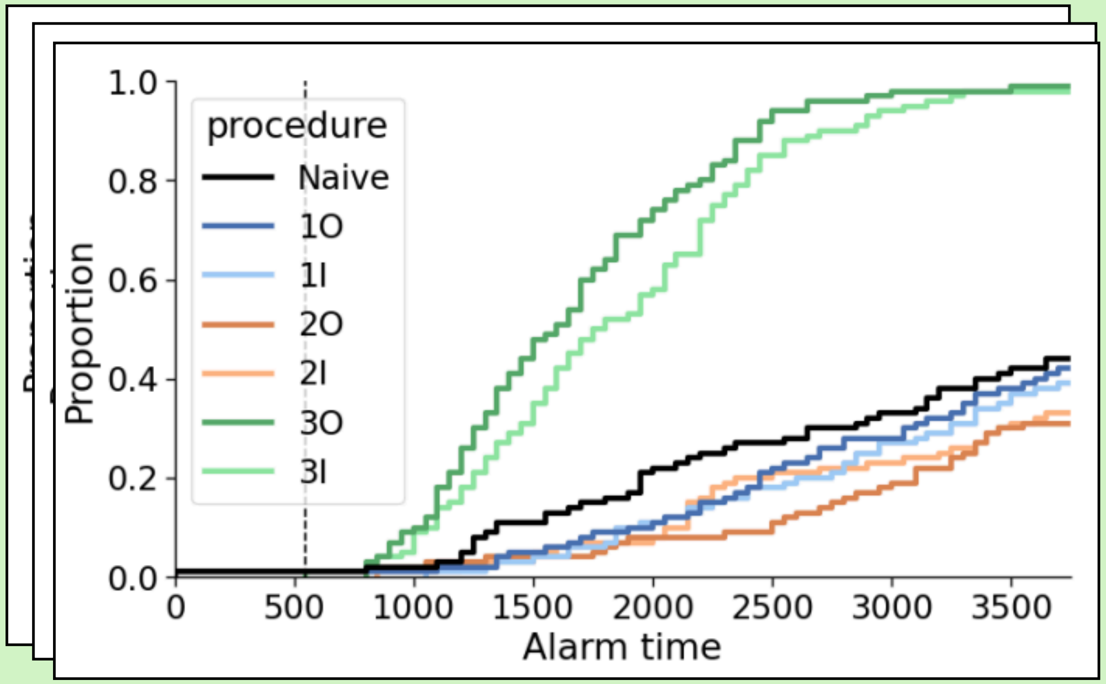
A postmarket monitoring framework

1. Define potential monitoring criteria

2. Enumerate sources of bias and define the causal model

3. Describe candidate monitoring strategies

4. Compare the pros and cons of candidate strategies



Procedure	Interpretability	Fairness	Assumptions
1I	High	None	Positivity
1O	High	None	Positivity, Conditional Exchangeability
2I	High	Moderate	Positivity
2O	High	Moderate	Positivity, Conditional Exchangeability
3I	Medium	Strong	None
3O	Medium	Strong	Conditional Exchangeability

A postmarket monitoring framework

1. Define potential monitoring criteria

2. Enumerate sources of bias and define the causal model

3. Describe candidate monitoring strategies

4. Compare the pros and cons of candidate strategies

Select final strategy after discussion with team members and stakeholders

Thank you!

https://arxiv.org/abs/2311.11463

Funding: This work was funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2022C1-25619).