Statistically Valid Post-Deployment Monitoring Should Be Standard for AI-Based Digital Health

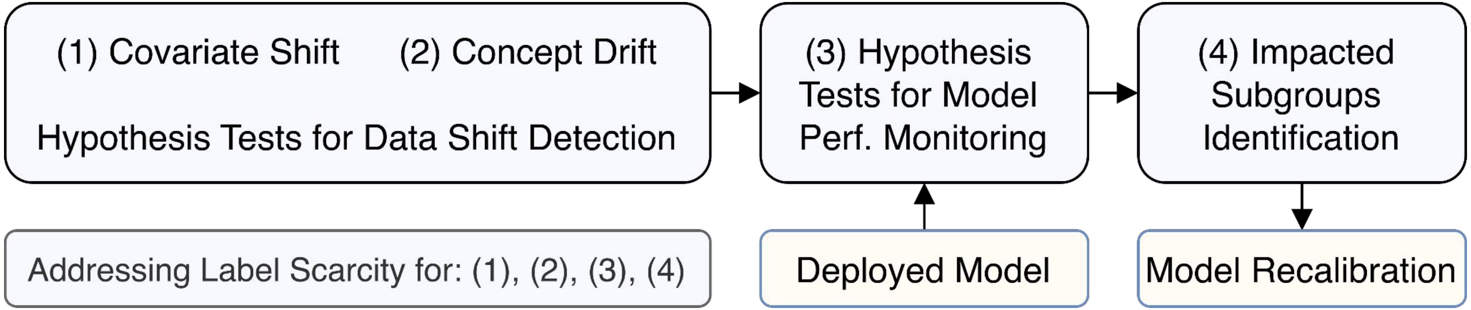
Pavel Dolin, Weizhi Li, Gautam Dasarathy, Visar Berisha

Arizona State University

Motivation and Problem

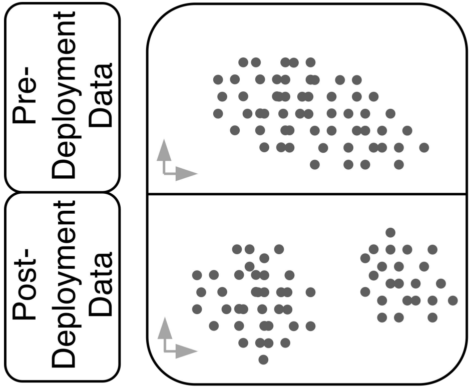
- A recent review found that only 9% of FDA-registered AI-based healthcare tools include a post-deployment surveillance plan
- Post-deployment monitoring is needed
  - Once the model is trained it faces changing distributions, unanticipated use cases, and data shift
  - Hard to monitor due to limited access to labels
- FDA expects statistically valid post-deployment testing
- Gap: How do we do label-efficient statistically valid testing in post-deployment settings

Proposal



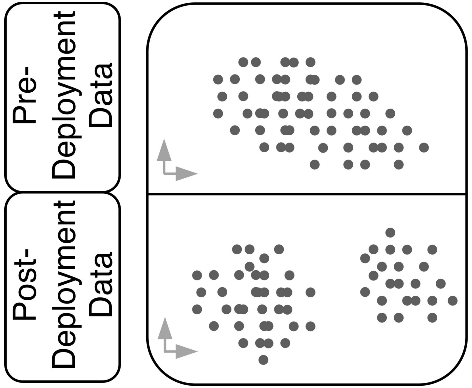
Framing Post-Deployment Monitoring as Hypothesis Testing. (1) Hypothesis tests for covariate shift, (2) concept drift. If a statistically significant change is observed, (3) a hypothesis test for model performance degradation is performed. If a model is affected by the change, (4) impacted subgroup identification is performed and used for target label collection and model recalibration. One of the open problems is addressing label scarcity for each of the described stages.

1. Covariate Shift



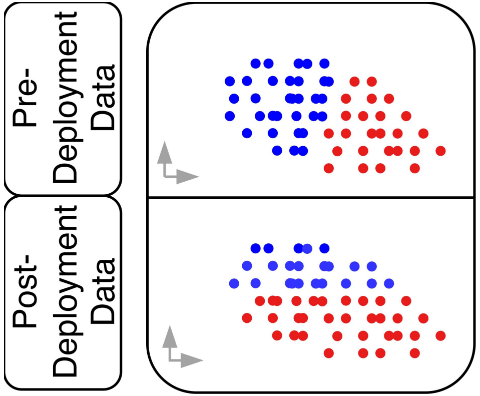
$$
\begin{aligned}
H_0 : \quad & p_{t_0}(\mathbf{s}, \mathbf{c}) = p_{t_1}(\mathbf{s}, \mathbf{c}), \\
H_1 : \quad & p_{t_0}(\mathbf{s}, \mathbf{c}) \neq p_{t_1}(\mathbf{s}, \mathbf{c}).
\end{aligned}
$$

1. Covariate Shift



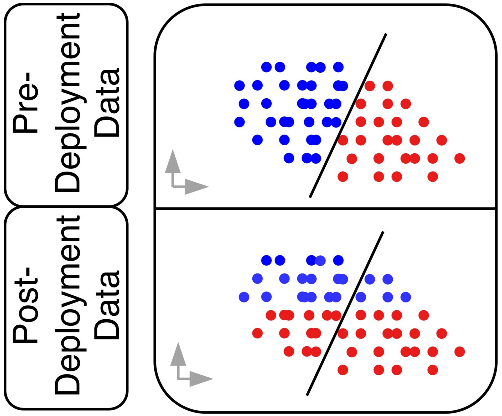
$$\begin{aligned} H_0 &: \quad p_{t_0}(\mathbf{s}, \mathbf{c}) = p_{t_1}(\mathbf{s}, \mathbf{c}), \\ H_1 &: \quad p_{t_0}(\mathbf{s}, \mathbf{c}) \neq p_{t_1}(\mathbf{s}, \mathbf{c}). \end{aligned}$$

2. Concept Drift



$$\begin{aligned} H_0 &: \quad p_{t_0}(\mathbf{s}, \mathbf{c}, y) = p_{t_1}(\mathbf{s}, \mathbf{c}, y) \\ H_1 &: \quad p_{t_0}(\mathbf{s}, \mathbf{c}, y) \neq p_{t_1}(\mathbf{s}, \mathbf{c}, y). \end{aligned}$$

3. Model Performance Monitoring



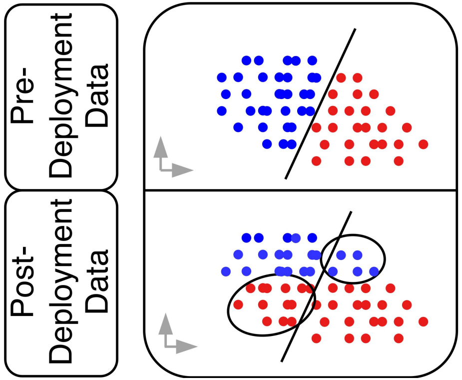
3.1 Performance Deviation Testing

$$
\begin{aligned}
H_0 &: \quad M_{t_0} - M_{t_1} \leq \tau_{\text{deg}}, \\
H_1 &: \quad M_{t_0} - M_{t_1} > \tau_{\text{deg}}
\end{aligned}
$$

3.2 Specification Threshold Testing

$$
\begin{aligned}
H_0 &: \quad M_{t_1} \geq \tau_{\text{spec}}, \\
H_1 &: \quad M_{t_1} < \tau_{\text{spec}}.
\end{aligned}
$$

4. Impacted Subgroups identification

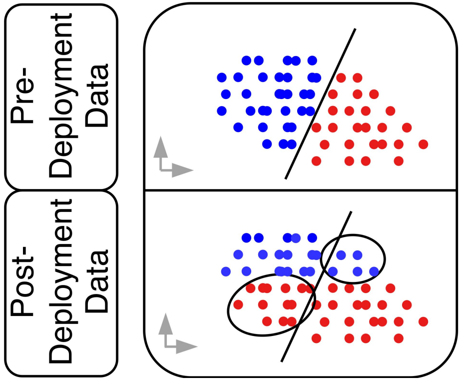


$$M_t^{\mathcal{G}} \;=\; g\left(f, p_t\left(\mathbf{s}, \mathbf{c}, y \mid \mathcal{G}\right)\right)$$

$$\max_{\mathcal{G} \subseteq \mathcal{S} \times \mathcal{C}} M_{t_0}^{\mathcal{G}} - M_{t_1}^{\mathcal{G}}$$

4. Impacted Subgroups identification

Open Challenges



$$M_t^{\mathcal{G}} = g\left(f, p_t\left(\mathbf{s}, \mathbf{c}, y \mid \mathcal{G}\right)\right)$$

$$\max_{\mathcal{G} \subseteq \mathcal{S} \times \mathcal{C}} M_{t_0}^{\mathcal{G}} - M_{t_1}^{\mathcal{G}}$$

- Identifying features where the model's discriminative power has shifted
- Detecting subgroups that experience disproportionate performance degradation
- Uncovering complex interaction patterns

Call to Action & Exciting Open Challenges

- Choosing the right test statistic for high dimensional settings and in the presence of mixed data types
- Addressing label scarcity
  - Data Shifts Detection
  - Model Performance Degradation
- Impacted Subgroup Identification

References

[1] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. Pattern Recognition, 45(1):521–530, January 2012.

[2] Dong Liu, William Baskett, David Q Beversdorf, and Chi-Ren Shyu. Exploratory data mining for subgroup cohort discoveries and prioritization. IEEE Journal of Biomedical and Health Informatics, 24(5):1456–1468, 2020

[3] Daniel Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In Machine Learning and Knowledge Discovery in Databases, pages 1–16. Springer, 2008.

Thank you!