Large-Scale Study of Temporal Shift in Health Insurance Claims

Christina X Ji[1], Ahmed M Alaa[2], David Sontag[1]
[1]MIT CSAIL and IMES, [2]UC Berkeley and UCSF
Email: cji at mit dot edu @christina_x_ji

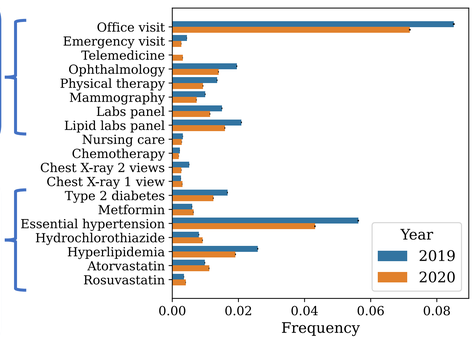
Paper:

Motivation

When clinical guidelines, patient characteristics, recording patterns, or data availability change over time, a machine learning **model trained on historical data may no longer be optimal**.

US guidelines in April 2020 recommended postponing non-urgent services.[1,2]

Condition claims decreased, but drug prescription rates did not, reflecting less care for chronic conditions.



Definition for Temporal Shift

Temporal shift affects an outcome at time t within a sub-population h_t if a metric ϕ such as AUC satisfies

$$\hat{\phi}_{\mathcal{D}_t}(\hat{f}_t, h_t) - \hat{\phi}_{\mathcal{D}_t}(\hat{f}_{t-1}, h_t) > 0$$

Current model $\hat{f}_t$ on current data $\mathcal{D}_t$ $>$ Previous model $\hat{f}_{t-1}$ on current data $\mathcal{D}_t$

Baseline definition from benchmark[3]:

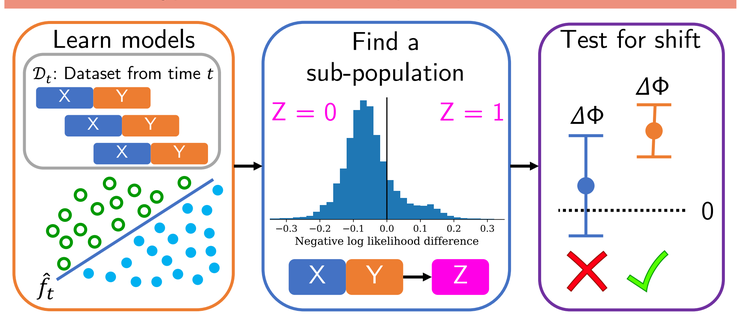
$$\hat{\phi}_{\mathcal{D}_{t-1}}(\hat{f}_{t-1}, h_t) - \hat{\phi}_{\mathcal{D}_t}(\hat{f}_{t-1}, h_t) > 0$$

Previous model $\hat{f}_{t-1}$ on previous data $\mathcal{D}_{t-1}$ $>$ Previous model $\hat{f}_{t-1}$ on current data $\mathcal{D}_t$

Algorithm to Test for Temporal Shift

To **test for temporal shift**, we perform 3 steps:

1. Learn outcome models
2. Find a sub-population affected by shift
3. Perform hypothesis test for temporal shift definition



Algorithm to Scan for Temporal Shift

We create an algorithm to scan for temporal shift in multiple outcomes across many time periods within many sub-populations

Repeat test for temporal shift for each outcome, time, and sub-population → Control false discovery rate with multiple hypothesis correction → Examine clinical significance

Set-up for Large-Scale Scan

Goal: **Assess yearly temporal shift in a large collection of outcomes** from 2015 to 2020

Dataset: Health insurance claims from Philadelphia area from 2014 to mid-2021

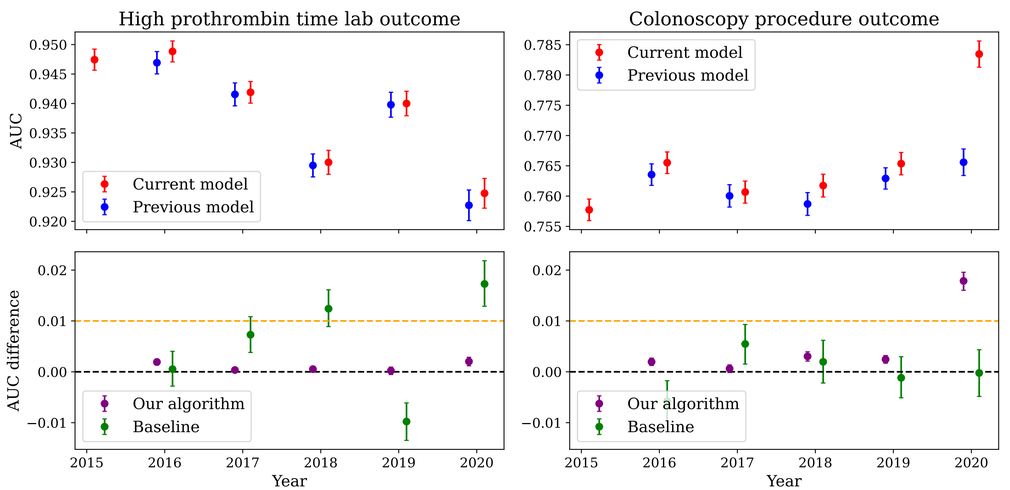
Outcomes: Defined as occur in the next 3 months:

- 100 initial condition diagnoses
- 100 abnormal lab measurements
- 42 procedure groups

- *Cohort:* Over 1.6 million patients with average of 34 samples per patient
- *Features:* Over 15,000
- *Number of scanned tasks:* 1010

Illustrative Tasks in Large-Scale Scan

Our algorithm can successfully detect temporal shift, while the baseline erroneously identifies performance changes due to other factors as temporal shift.



Baseline erroneously flags 2018 and 2020 as temporal shift

Our algorithm detects shift in 2020, while baseline does not.

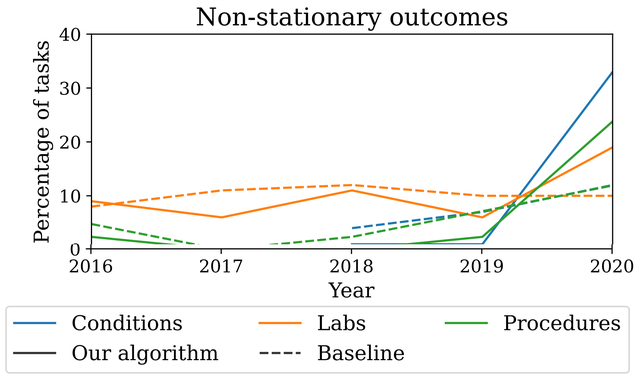
Top: AUC with bootstrap standard error Bottom: 90% bootstrap confidence intervals
Our algorithm: Difference between red at time t and blue at time t
Baseline: Difference between red at time $t-1$ and blue at time t
Dotted lines are thresholds: Statistical significance (of interval) Clinical significance (of point estimate)

Acknowledgments

We are grateful to Alejandro Buendia, Hunter Lang, Rebecca Boiarsky, Michael Oberst, Hussein Mozannar, Yury Polyanskiy, Leora Horwitz, and Saul Blecker for input on our algorithms, study design, and data extraction pipeline. We would also like to thank Aaron Smith-McLallen, Stephanie Gervasi, James Denyer, Eric Wilkinson, George Fenimore, Monique Wilkins, and the rest of the data science group at Independence Blue Cross for their expertise, data, and financial support.

Results from Large-Scale Scan

9.7% of tasks are affected by temporal shift at the population level



- 62 of 98 tasks with temporal shift are in 2020
- Lab outcomes account for 32 of 36 tasks before 2020
- 93% of tasks have some sub-population affected

Case Studies

Temporal shifts detected by our algorithm are driven by changes in clinical guidelines

- COVID-19 pandemic: 781 features affected by domain shift
- 4 low eGFR lab outcomes have label shift in 2018 driven by switch from MDRD to CKD-EPI formula
- Inpatient consultations dropped more swiftly among Medicare patients in 2019 due to reimbursement policy changes

Conclusion

Temporal shift is prevalent in healthcare. To combat the negative effects on model performance, **detecting when to train a new model is important**.

We propose algorithms to test for temporal shift and scan for shift in a large collection of outcomes, time points, and sub-populations.

We demonstrate our methods in a large-scale study and find 9.7% of tasks are affected by temporal shift at the population level.

References

[1]Berkenstock et al. Changes in patient visits and diagnoses in a large academic center during the COVID-19 pandemic. BMC ophthalmology, 2021.
[2]Hartnett et al. Impact of the COVID-19 pandemic on emergency department visits–United States, January 1, 2019–May 30, 2020. Morbidity and Mortality Weekly Report, 2020.
[3]Yao et al. Wild-time: A benchmark of in-the-wild distribution shift over time. NeurIPS, 2022.