From Individual Experience to Collective Evidence:
An Incident-Based Framework for Identifying Systemic Discrimination

Jessica Dai, Paula Gradu, Inioluwa Deborah Raji, Benjamin Recht
University of California, Berkeley (alphabetical order)

Does some system that individuals interact with cause disproportionate harm to a meaningful subgroup?

- Individuals are described by covariates $X \in \mathbb{R}^d$, and groups $\mathcal{G}$ are defined by a function of covariates.
  Individuals can be in multiple groups at once; we know the base rate μ_G^0 for every group.
- Reports arrive sequentially. We *stop the algorithm* when we identify a group that has been harmed.

Given a reporting system, does any group report much more frequently than we would expect?

Example – vaccines (VAERS) and pharmaceuticals (FAERS)

- System: a particular vaccine or drug
- "Bad event": medical adverse event (in general, or for a particular symptom)
- Reporting: reflects true "bad event" (not necessarily causal)
- Main question: "does this treatment disproportionately cause adverse events for a particular subgroup?"

Example – (algorithmic) decision-making

- System: a decision-making system, e.g. for loan allocations
- "Bad event": an incorrect decision, e.g. denials to high-creditworthy individuals
- Reporting: correlated with (but not always equal to) "bad event"
- Main question: "does this decision-making system discriminate against a subgroup?"

Algorithm 1: Testing for overrepresentation in reports

Input: set of groups $\mathcal{G}$ with base preponderances $\{\mu_G^0\}_{G\in\mathcal{G}}$, relative strength β, test level α, group size c

for $t = 1, 2, \ldots$ **do**
  See an incident report X_t;
  for $G \in \mathcal{G}$ **do**
    Update (log-)likelihood of harm for G depending on whether $X_t \in G$;
    Using differential privacy, choose $\widehat{G}^\star$ to be the most impacted group based on reports seen thus far;
    Test $\widehat{G}^\star$ for harm at level α;
    if $\widehat{G}^\star$ *is likely to be harmed* **then**
      Return $\widehat{G}^\star$.

Algorithm 2: Formal statement of Algorithm 1

Input: set of groups $\mathcal{G}$ with base preponderances $\{\mu_G^0\}_{G\in\mathcal{G}}$, relative strength β, test level α, group size c

Initialize $\omega_0^G = 0$ and $\lambda^G = 1/2$ for all $G \in \mathcal{G}$;
Compute $\tau = \ln(1 + \frac{1}{1-c})$ and $\widetilde{\alpha} = \max_{\gamma\in(0,\alpha)}(\alpha - \gamma)\exp\left(-\frac{1}{16} - \frac{1}{4}\sqrt{\ln(2/\gamma)}\right)$;
for $t = 1, 2, \ldots$ **do**
  for $G \in \mathcal{G}$ **do**
    Update $\omega_t^G \leftarrow \omega_{t-1}^G + \ln\left(1 + \lambda_t^G(\mathbf{1}_{X_t\in G} - \beta\mu_G^0)\right)$;
    Sample $\xi_t^G \sim \mathrm{Lap}\left(4\tau\sqrt{2t}\right)$;
    Let $z_t = \frac{\mathbf{1}_{X_t\in G} - \beta\mu_G^0}{1+\lambda_t^G(\mathbf{1}_{X_t\in G} - \beta\mu_G^0)}$;
    Let $\lambda_{t+1}^G \leftarrow \mathrm{Proj}_{[0,1]}\left(\lambda_t^G + \frac{2}{2-\ln(3)}\frac{z_t}{1+\sum_{s\in[t]} z_s^2}\right)$;
  Let $\widehat{G}_t^\star \leftarrow \arg\max_{G\in\mathcal{G}}(\omega_t^G + \xi_t^G)$;
  if $\omega_t^{\widehat{G}_t^\star} \geq \ln(1/\widetilde{\alpha})$ **then**
    Return $\widehat{G}_t^\star$.

Main algorithmic idea: Sequential + multiple hypothesis testing for whether any group over-reports relative to their base rate.

Key questions:

1. **Correctness:** If we stop the test, is it true that our $\widehat{G}_t^\star$ is actually reporting disproportionately often?
2. **Stopping time:** If there do exist groups that are harmed, how quickly can we identify them?

Some technical ingredients:

- **"e-values"** (non-negative supermartingales) to ensure anytime-valid stopping (Alg. 2: L1, 5, 10)
- **Differentially-private selection** to avoid Bonferroni over groups (Alg. 2: L2, 6, 9, 10)
- **Online newton step** on λ_t^G to improve strength of the test (Alg. 2: L7, 8)

Detecting disproportionate reports of myocarditis in young men from COVID vaccines (VAERS data)

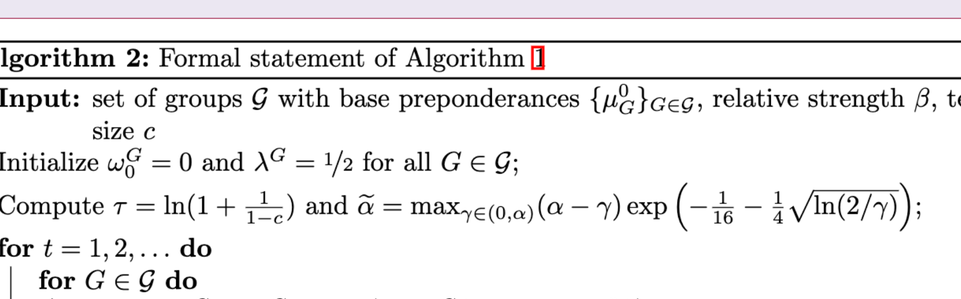

Figure 1: Results for incidents reporting Myocarditis or related sympotoms after taking the COVID-19 vaccine over 50 trials. *Left:* Comparison of stop time across algorithms; *Center:* The non-ONS algorithm baseline fail rate again decreases with an increase in α; *Right:* Comparison of validity across algorithms. $\alpha = [0.01 - 0.10]$, $\beta = 1.2$ for all.

Detecting disparity in loan denials to high-creditworthy individuals (HMDA data, simulated reports)

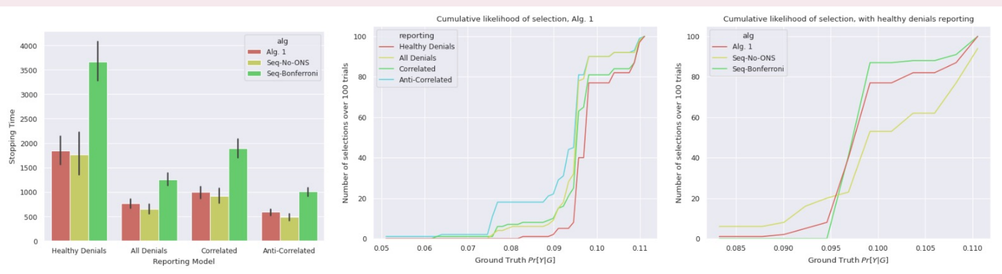

Figure 3: Impact of reporting models; 100 trials. *Left:* Stopping time by reporting model and algorithm. *Middle and right:* Ground-truth $\Pr[Y|G]$ vs cumulative selections. *Middle:* Alg. 2 across various reporting models; *Right:* comparison across algorithms for healthy denials reporting model. $\alpha = 0.01$, $\beta = 1.4$ for all.