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

This manuscript is a working draft as of December 2024; please do not distribute.

Abstract

When an individual reports a negative interaction with some system, how can their personal experience be contextualized within broader patterns of system behavior? We study the *incident database* problem, where individual reports of adverse events arrive sequentially, and are aggregated over time. In this work, our goal is to identify whether there are subgroups—defined by any combination of relevant features—that are disproportionately likely to experience harmful interactions with the system. We formalize this problem as a sequential hypothesis test, and identify conditions on reporting behavior that are sufficient for making inferences about disparities in true rates of harm across subgroups. We show that algorithms for sequential hypothesis tests can be applied to this problem with a standard multiple testing correction. We then demonstrate our method on real-world datasets, including mortgage decisions and vaccine side effects; on each, our method (re-)identifies subgroups known to experience disproportionate harm using only a fraction of the data that was initially used to discover them.

1 Introduction

The impact of injustice is most acutely felt by the individual. But if an individual experiences harm, how can they know whether their experience is an isolated incident or part of a larger pattern of discrimination?

Fairness work has historically focused on model developers and third-party auditors as the main actors involved in creating fair mechanisms, motivating methods to construct models that are fair with respect to pre-defined subgroups at development time (e.g. as surveyed in Pessach and Shmueli [2022])—or in identifying unfair ones, motivating post-hoc audits that occur after the entire decision-making process has completed (e.g., [Byun et al., 2024, Martinez and Kirchner, 2021]). However, in most applications where fairness is a concern, problems with the system may only emerge over time, and it is not necessarily obvious which subgroups might be important. Moreover, such approaches to fairness provide no mechanism for individuals to raise concerns.

It is exactly this question of individual agency that drives our work. In addition to normative concerns, which suggest that individuals ought to have a voice in expressing concerns with their treatment (see, for example, the literature on contestability of algorithmic decisions, e.g. Vaccaro et al. [2019]), recent legislation has also highlighted individual reporting as a policy mandate for the governance of AI systems (e.g., the EU AI act [European Parliament, 2023]). While such legislation has yet to see full implementation, mechanisms for individual incident reporting already exist in a variety of application domains, including consumer finance, medical devices, and vaccines and pharmaceuticals. A key component of reporting databases in the latter settings is that information from individual reports are aggregated to build collective knowledge about specific vaccines or pharmaceuticals—and, when applicable, this aggregated information can drive downstream decisionmaking, such as updating vaccine guidelines or drug treatment protocols (e.g. Oster et al. [2022]).

Fairness is an especially salient application for incident reporting systems: while individuals bear the harm, commonly-accepted (and legally-legible) notions of fairness are understood at an aggregate level. In fact, existing examples of (algorithmic) discrimination lawsuits (e.g., Gilbert [2023] in hiring, or in housing) are often structured as class actions, even as they are initiated by individuals based on their personal

experiences. Crucially, individuals themselves may not know whether their experience with the system was inherently problematic, and deserving of redress, until it is placed in context with the experiences of others. On the other hand, while existing incident databases do not typically analyze reporting behavior, it may be necessary to consider reporting more carefully in order for incident databases to be useful for fairness auditing in more general settings, such as for algorithms that make allocation decisions.

In this paper, we consider what a realistic approach to assessing fairness claims from an incident database might look like in practice. We are primarily interested in designing a framework for the general public to report and contest large-scale harms by leveraging reports of *individual experience* to inform *collective evidence* of discrimination. Our contributions are as follows.

1. **Model.** We propose *incident databases*, which allow individuals to submit reports of negative interactions, as a new mechanism for post-deployment fairness auditing. We show how to find evidence of disparately impacted subgroups without requiring knowledge of expected incidence rates. In particular, we identify conditions on reporting behavior and show how they can be used to to make inferences about rates of true harm (Section 3).

2. **Algorithms.** Our formalization of the problem allows us to leverage known approaches to sequential hypothesis testing, with theoretical guarantees that inherit from properties of those algorithms. We show how to instantiate two reasonable algorithms for our proposed test and provide theoretical guarantees for each (Section 4).

3. **Real-world validation.** We illustrate the usefulness of our approach using real-world datasets, for applications with known disparity in per-subgroup rates of harm. On both real vaccine incident reports and on mortgage allocation decisions, our algorithm correctly identifies groups that disproportionately experience harm—and does so using a comparatively small number of reports (Section 5).

1.1 Related work

The incident database problem we study is at the intersection of various challenges addressed in fairness and statistics. We give an overview here and provide more detailed technical discussion of the most closely-related works in Appendix A.

Algorithmic accountability via (individual) reports. Some recent work considers methods for learning about fairness problems via individual reports from both theoretical [Globus-Harris et al., 2022] and practical [Agostini et al., 2024] perspectives. However, most discussion of individual experiences in machine learning fairness literature is limited to contexts where the objective is to assess, appeal, contest or seek recourse for that individual to change their *individual* outcomes, rather than forming a *collective* judgment about the system as a whole [Sharifi-Malvajerdi et al., 2019, Ustun et al., 2019, Karimi et al., 2022]. Other work on identifying fairness-related issues via reporting data has typically focused on learning in batch and/or post-hoc contexts. Positive-unlabeled (PU) learning has been suggested as a mechanism for learning from reporting data, especially in the context of modeling disparate reporting rates across subgroups (e.g., Shanmugam et al. [2024], Wu and He [2022]). In other works, identifying disparate reporting rates is itself is the central challenge (e.g., Liu and Garg [2022], Liu et al. [2024]).

On the other hand, an emerging body of literature from the human-computer interaction community develops the concept of *contestability* (e.g., Almada [2019], Vaccaro et al. [2019], Landau et al. [2024], Karusala et al. [2024]); though contestability is still typically understood in terms of individual outcomes, we see our work as one possible path to implementing this ideal, with an eye towards empowering contestability at larger scale.

Fairness auditing as hypothesis testing. Cen and Alur makes a direct connection between legal AI fairness audit requirements and hypothesis testing, although mainly considers a post-hoc setting. Cherian and Candès [2023] take a multiple testing approach for handling a large number of groups, but this test is again post-hoc (or entirely pre-deployment). Perhaps the most closely related works are that of Chugg et al. [2024] and Feng et al., who propose applying a sequential hypothesis test with the explicit goal of quickly identifying bias in deployed systems in real time.

Identifying and defining subgroups. One approach to subgroup definition, following the line of work in multicalibration Hébert-Johnson et al. [2018], is to simply enumerate over all possible combinations of covariates. For sequential problems, per-group guarantees can be provided for subgroups that are learned online [Dai et al., 2024]; in the context of sequential experiments, Adam et al. [2024] propose an approach to early stopping that does not require the experimenter to pre-specify the group experiencing harm.

Sequential and multiple hypothesis testing. We leverage the recent literature on e-values (e.g. Waudby-Smith and Ramdas [2024], Vovk and Wang [2021]), which can be used to construct sequential tests that have validity guarantees in finite samples. While existing literature suggests methods for global null testing that can aggregate e-processes (e.g., Cho et al. [2024] or Chi et al. [2022]), such approaches are unable to provide per-hypothesis guarantees. More classical approaches include Wald's Sequential Probability Ratio Test (SPRT) and its extensions, such as Max-SPRT [Kulldorff et al., 2011], or a sequential generalization of the Holm procedure Bartroff and Song [2014].

Incident database analysis Sequential hypothesis tests have been used for real-world monitoring of adverse incidents in vaccines and medical devices (see, e.g., Shimabukuro et al. [2015]). Descriptive studies have identified disparate adverse impacts in pharmaceutical [Lee et al., 2023, Whitley and Lindsey, 2009] and vaccine settings [Oster et al., 2022].

Finally, we note that for AI systems, the term "incident database" been used to describe systems for monitoring the adverse impact of algorithmic deployments (e.g., Turri and Dzombak [2023], Feffer et al. [2023], Raji et al. [2022]). However, in the context of our work, we are actively excluding accident catalog databases, which include the colloquially named "AI incident" databases that draw direct inspiration from them (e.g., McGregor [2021], Ojewale et al. [2024]). Instead, we focus on reporting databases that provide records of individual experiences of adverse events that are tied to specific systems.

2 Model, Notation, and Preliminaries

The goal of constructing an incident database is to determine whether some system that individuals interact with—for example, an (algorithmic) loan decision system, or a medical treatment—results in disproportionate harm to some meaningful subgroups. For the incident database associated with a particular system, we will use $Y \in \{0, 1\}$ as an indicator variable that denotes the undesirable event corresponding to that system. For example, in loan decisions, this could correspond to the event that a highly-qualified individual was denied a loan; in the medical setting, this may be an adverse physical side effect due to the treatment.

Subgroup definitions. Individuals are characterized with feature vectors $X \in \mathcal{X}$, and we index individuals as X_i ("features of individual i") or X_t ("features of the individual who reports at time t"). Every individual X_i "belongs to" at least one group G, and we will denote the event that X_i belongs to G as $\{X_i \in G\}$; we will use $\mathcal{G}$ to denote the set of all possible groups. This set of possible groups $\mathcal{G}$ can be defined arbitrarily as long as all groups can be determined as a function of covariates $\mathcal{X}$. We allow for groups to be overlapping—that is, we allow each individual X_i to be in multiple groups so that $|\{G' \in \mathcal{G} : X_i \in G'\}| \geq 1$. For example, it is possible to set $\mathcal{G} := 2^{\mathcal{X}}$ as in Hébert-Johnson et al. [2018].

Reference population. The system for which the database is constructed naturally has a corresponding reference population of eligible individuals. For example, this could be everyone who has applied for a loan, or everyone who has been prescribed a certain medication. Thus, given a set of groups $\mathcal{G}$, we assume that it is possible to compute the composition of the reference population.

Assumption 2.1 (Reference population)**.** *For every $G \in \mathcal{G}$, the quantity $\mu_G^0 := \Pr[X \in G]$ is known. Throughout this work, we refer to the set $\{\mu_G^0\}_{G \in \mathcal{G}}$ as* base preponderances.

Reporting. As the database administrator, the high-level goal is to determine whether there exists some subgroup $G \in \mathcal{G}$ where $\Pr[Y \mid X \in G]$ is abnormally high. Crucially, the database does not have access to information about every individual who has interacted with the system; instead, individuals *may* report to

the database if they believe that they experienced bad event Y. We let R_i be a random variable representing whether individual i decides to report (with $R_i = 0$ indicating no report). A key quantity for each group is $\mu_G := \Pr[X_i \in G \mid R_i = 1]$, that is, the proportion of *reports* that each G comprises; we sometimes refer to $\{\mu_G\}_{G\in\mathcal{G}}$ as (reporting) preponderances. A central claim of this paper is that comparing μ_G to μ_G^0—that is, the extent to which group G is (over)represented within the reporting database—can be a useful signal for $\Pr[Y \mid G]$ in a wide class of applications.

In our model, reports are received sequentially, and each individual X_t belongs to each group G with probability μ_G—that is, $\mathbf{1}[X_t \in G] \sim_{\text{i.i.d.}} \text{Bern}(\mu_G)$.[1] The i.i.d. model of course simplifies the analysis and exposition, but itself is not intrinsic to modeling the incident reporting problem as a sequential hypothesis test. As we will show in Section 4.2, the explicit i.i.d. assumption can be relaxed; more generally, any probabilistic model for sequential testing can be adapted to incident reporting.

3 Identifying Discrimination by Modeling Preponderance

A major challenge of assessing potentially-differential rates of harm across subgroups using only reporting data is to relate the event that someone submits a report to the event that they experienced harm. That is, if someone did experience a negative outcome, how likely is it for them to have reported it, and conversely, how if someone submitted a report, how likely is it to reflect "true" harm? Moreover, as is known from prior work, reporting rates themselves can vary across subgroups.

Our central proposal is to conduct a hypothesis test for each group to determine whether it is overrepresented by a factor of β among reports. That is, for each $G \in \mathcal{G}$, we test the following hypotheses:

$$\mathcal{H}_0^G : \mu_G < \beta\mu_G^0 \qquad\qquad \mathcal{H}_1^G : \mu_G > \beta\mu_G^0. \tag{1}$$

In Section 4, we will discuss concrete algorithms for conducting this test sequentially and their corresponding theoretical guarantees. Before doing so, we first argue that testing for preponderance among reports, i.e., tracking μ_G in this way, can be a meaningful way to identify discrimination. In Sections 3.1 and 3.2, we describe two distinct ways that this particular test can be interpreted; see Appendix B for a discussion of some practical considerations for the modeling task.

3.1 Preponderance as relative risk

The first interpretation of our test allows us to make inferences about relative risk, the ratio between the rate of harm experienced by group G and on average over the population. In this interpretation, the key quantity is the *report-to-incidence ratio*.

Definition 3.1 (Report-to-incidence ratio)**.** *We define the* report-to-incidence-ratio (RIR) *as* $\rho := \frac{\Pr[R=1]}{\Pr[Y=1]}$*, and the group-conditional analogue as* $\rho_G := \frac{\Pr[R=1|G]}{\Pr[Y=1|G]}$.

In Proposition 3.2, we show that if the group-conditional RIR of some group G is at most some constant multiple of the population-wide RIR, then we can easily convert a lower bound on report preponderence into a lower bound on true relative risk.

Proposition 3.2. *Define the relative risk of group G to be* $\text{RR}_G := \frac{\Pr[Y=1|G]}{\Pr[Y=1]}$*. Suppose that for some group G we have $\rho_G \le b \cdot \rho$. Suppose that we determine that $\mu_G \ge \beta\mu_G^0$ for some $\beta > 1$. Then the true relative risk experienced by G is at least* $\text{RR}_G \ge \beta/b$.

Proof. First, note that by definition of ρ, ρ_G, and RR_G, we have

$$\rho_G \le b \cdot \rho \iff \frac{\Pr[R=1 \mid G]}{\Pr[Y=1 \mid G]} \le b \cdot \frac{\Pr[R=1]}{\Pr[Y=1]} \iff \text{RR}_G \ge \frac{\Pr[R=1 \mid G]}{\Pr[R=1]} \cdot \frac{1}{b}.$$

By Bayes' rule, $\frac{\Pr[R=1|G]}{\Pr[R=1]} = \frac{\Pr[G|R=1]}{\Pr[G]} = \frac{\mu_G}{\mu_G^0}$; furthermore, by assumption, we have $\frac{\mu_G}{\mu_G^0} \ge \beta$. The result follows from combining with the previous display. □

[1] Note that, because we allow groups to overlap, we cannot enforce $\sum_G \mu_G^0 = 1$ or $\sum_G \mu_G = 1$, and moreover the events $\{X_t \in G\}$ and $\{X_t \in G'\}$ are correlated for any $G, G' \in \mathcal{G}$, i.e. the independence does not hold across groups. The key point in our case is independence across time.

For example, suppose we take $\max_G \rho_G/\rho \leq b = 1.25$, i.e., no group over-reports 25% more frequently than the population average. Then, if a test identifies a group G for which $\mu_G \geq 1.75 \cdot \mu_G^0$, this implies that the *true* relative risk of harm for group G is at least $\mathrm{RR}_G \geq 1.4$—that is, G experiences harm 40% more frequently relative to the population average.

3.2 Preponderance as true incidence rate

We now discuss an alternate way to convert a lower bound on preponderance into a guarantee on real-world harm. In this case, we can infer the true incidence rate of harm (that is, no longer relative to the average) if we are able to estimate—or willing to make assumptions on—true and false reporting behavior in groups. Moreover, assumptions (or estimations) of these reporting rates need only be made in relation to the population average reporting rate $\Pr[R]$.

Definition 3.3 (Reporting rates)**.** *Let $r := \Pr[R]$ be the average reporting rate over the full population. Let $\gamma_G^{TR} := \frac{1}{r}\Pr[R_i = 1 \mid Y_i = 1, X_i \in G]$, $\gamma_G^{FR} := \frac{1}{r}\Pr[R_i = 1 \mid Y_i = 0, X_i \in G]$, Finally, let $\mathrm{IR}_G := \Pr[Y \mid G]$ represent the true incidence rate, i.e. the likelihood that an individual in G experiences Y.*

Note that $r \cdot \gamma_G^{\mathrm{TR}}$ represents the (possibly group-conditional) rate at which an individual $X_i \in G$ who experiences Y actually reports, while $r \cdot \gamma_G^{\mathrm{FR}}$ represents the rate that an individual $X_i \in G$ who does not experience Y reports. Thus, γ_G^{TR} and γ_G^{FR} represent how much more (or less) a particular group G makes true or false reports *relative to the population average* rate. We make the relationships between γ_G^{TR}, γ_G^{FR}, and our quantity of interest IR_G, more precise in the following.

Proposition 3.4. *Suppose that, for some G, it is determined that $\mu_G \geq \beta\mu_G^0$ for some $\beta > 1$. As long as $\gamma_G^{TR} > \gamma_G^{FR}$ for every $G \in \mathcal{G}$, $\mathrm{IR}_G \geq \frac{\beta \Pr[R] - \gamma_G^{FR}}{\gamma_G^{TR} - \gamma_G^{FR}}$.*

Proof of Proposition 3.4. Recall that we have defined $\mu_G = \Pr[G \mid R]$, and $\mu_G^0 = \Pr[G]$ is known by Assumption 2.1. By Bayes' rule, we have $\mu_G = \Pr[G \mid R] = \frac{\Pr[G]\Pr[R|G]}{\Pr[R]} = \mu_G^0 \frac{\Pr[R|G]}{r}$, where randomness is due to reporting. Now, let us decompose $\Pr[R \mid G]$ by "true" reports ($Y = 1$) and "false" reports ($Y = 0$). By the law of total probability, $\Pr[R \mid G] = r \cdot \big(\gamma_G^{\mathrm{TR}}\mathrm{IR}_G + \gamma_G^{\mathrm{FR}}(1 - \mathrm{IR}_G)\big)$; more precisely,

$$\begin{aligned}\frac{1}{r}\Pr[R \mid G] &= \Pr[R \mid G, Y = 1]\Pr[Y \mid G] + \Pr[R \mid G, Y = 0](1 - \Pr[Y \mid G]) \\ &= \gamma_G^{\mathrm{TR}}\mathrm{IR}_G + \gamma_G^{\mathrm{FR}}(1 - \mathrm{IR}_G) \\ &= \gamma_G^{\mathrm{FR}} + \mathrm{IR}_G(\gamma_G^{\mathrm{TR}} - \gamma_G^{\mathrm{FR}});\end{aligned}$$

combining this with the Bayes' rule computation, cancelling the $\frac{1}{r}$ factor, gives us $\mathrm{IR}_G = \frac{\frac{\mu_G}{\mu_G^0} - \gamma_G^{\mathrm{FR}}}{\gamma_G^{\mathrm{TR}} - \gamma_G^{\mathrm{FR}}}$. The result follows from the assumption that $\mu_G/\mu_G^0 \geq \beta$. □

Proposition 3.4 shows that the exact computation of IR_G depends on reporting rates γ_G^{TR} and γ_G^{FR}. While these quantities are not directly estimable from reporting data—in fact, estimating reporting rates is itself a distinct research challenge (e.g., Liu et al. [2024])—these results can nevertheless guide qualitative interpretation of how severe IR_G is.

For example, suppose a test is run for $\beta = 1.5$. Suppose that G overreports relative to the population average, with $\gamma_G^{\mathrm{FR}} = \Pr[R]$ (that is, G *falsely reports* at the same rate at the population average, which includes both true and false reports) and $\gamma_G^{\mathrm{TR}} = 2\Pr[R]$. Under these (generous) assumptions, we have that $\mathrm{IR}_G = 0.5$, an extremely high incidence rate for any application—regardless of incidence rates for other groups. Alternatively, suppose reporting rates did not vary by group. Then, if G is flagged at $\beta > 1$, there must be some G' with $\mathrm{IR}_G - \mathrm{IR}_{G'} \geq (\beta - 1)\frac{\Pr[R]}{\gamma_G^{\mathrm{TR}} - \gamma_G^{\mathrm{FR}}}$. If it is further assumed that $\gamma_G^{\mathrm{FR}} = 0$, then $\mathrm{IR}_G - \mathrm{IR}_{G'} \geq \beta - 1$.

4 Identifying Subgroups with High Reporting Overrepresentation

How might the test proposed in Equation (1) be carried out in practice, with reports arriving over time? At a high level, our algorithms follow the protocol outlined in Algorithm 1. For each group G, we maintain a test statistic ω_t^G that is updated as reports X_t are received over time. At each time t, each of these test statistics are compared to a threshold $\theta_t(\alpha)$, which depends on the test level α; the null hypothesis $\mathcal{H}_0^G$ for group G is rejected if $\omega_t^G > \theta_t(\alpha)$. For ease of exposition, Algorithm 1 is written so that groups corresponding to rejected nulls are collected in a set $\mathcal{G}^{\text{Flag}}$; in practice, a database administrator may choose to stop the test entirely as soon as one harmed group has been found.

In this section, we provide two algorithms that instantiate this sequential hypothesis test. In Section 4.1, we give a simple sequential Z-test-inspired approach which leverages a finite-time Law of the Iterated Logarithm. Section 4.2 presents a more complicated algorithm that uses recent developments in anytime-valid inference. The main differences in each algorithm lie in how they implement Lines 1 and 6 of Algorithm 1—that is, how test statistics and thresholds are computed. For both, handling for multiple hypothesis testing across groups is handled by a simple Bonferroni correction.

Algorithm 1: General protocol for testing overrepresentation

Input: Set of groups $\mathcal{G}$; base preponderances $\{\mu_G^0\}_{G\in\mathcal{G}}$; test level α; relative strength β

1 Initialize test statistic ω_0^G for every $G \in \mathcal{G}$ and compute threshold $\theta_0(\alpha)$;
2 Initialize set of rejected nulls (flagged groups) $\mathcal{G}^{\text{Flag}} := \emptyset$;
3 **for** $t = 1, 2, \ldots$ **do**
4 See X_t;
5 **for** $G \in \mathcal{G}$ **do**
6 Update test statistic ω_t^G and compute threshold $\theta_t(\alpha)$;
7 **if** $\omega_t^G \geq \theta_t(\alpha)$ **then**
8 Add G to $\mathcal{G}^{\text{Flag}}$ and take requisite action for G, if applicable.

4.1 Sequential Z-test

One simple observation that arises from the model presented in Section 2 is that if reports are arriving according to some underlying distribution—that is, for every group G, there is an underlying μ_G from which the sequence of reports X_t is drawn—then one might expect to be able to use concentration as a tool to conduct this test, since as time passes, the fraction of reports within the database from group G should converge to the true mean μ_G. We refer to this style of approach as a sequential Z-test, as it relies on measuring deviation from the mean.

Updating the test statistic ω_t^G. Given this intuition, the test statistic itself is a simple count of the number of times a report from each group has been seen, i.e. (with ω_0^G initialized at 0),

$$\omega_t^G \leftarrow \omega_{t-1}^G + \mathbf{1}[X_t \in G]. \tag{2}$$

Setting the threshold $\theta_t(\alpha)$. Given the way that ω_t^G accumulates evidence, one natural way to construct the threshold involves plus a correction term for both sample complexity and repeated testing over time. With C set to either $\sqrt{\beta\mu_G^0(1-\beta\mu_G^0)}$ or $1/2$, the threshold (which includes a Bonferroni correction) is

$$\theta_t(\alpha) := t \cdot \beta\mu_G^0 + C\sqrt{\frac{3}{2}\sqrt{\frac{1}{2}\ln\left(|\mathcal{G}|\frac{(2+\log_2(t))^2}{\alpha}\right)}}. \tag{3}$$

Theoretical guarantees. Our first guarantee is a bound on the probability that any group is incorrectly flagged. Formally, we have the following theorem:

Theorem 4.1 (Validity). *Running Algorithm 1 with $\theta_t(\alpha)$ as in Equation (3), setting $C = 1/2$, and ω_t^G updated as in Equations (2), guarantees that the probability that $\mathcal{G}^{Flag}$ will ever contain a group G where its corresponding null $\mathcal{H}_0^G$ is true is at most α, i.e.*

$$\Pr\left[\exists t : \exists G \in \mathcal{G}^{Flag} \text{ s.t. } \mathcal{H}_0^G \text{ holds}\right] \leq \alpha.$$

Detailed proofs of this result can be found in Lemma 1 of Jamieson et al. [2014], and Theorems 2 and 6 of Balsubramani [2014]. The choice of C affects the nature of the guarantee: the true, finite-sample anytime-validity guarantee requires $C = 1/2$. If instead $C = \sqrt{\beta\mu_G^0(1-\beta\mu_G^0)}$, then, strictly speaking, the guarantee holds only asymptotically. However, a higher value of C affects stopping time unfavorably, so the asymptotic approximation can be useful practically.

Notably, however, due to Theorem 3 of Balsubramani [2014], this is *not* a test of power one. In fact, there is an unavoidable anticoncentration due to the second term of Equation (3).[2]

Theorem 4.2 (Power). *Let T be the stopping time of Algorithm 1 with ω_t^G updated as in Equations (2) and $\theta_t(\alpha)$ as in Equation (3), with any setting of C. Let $\Delta_G := \mu_G - \beta\mu_G^0$, $\Delta_{\max} = \max_{G\in\mathcal{G}} \Delta_G$. If $\Delta_{\max} > 0$, then, with strict inequality, $\Pr[T < \infty] < 1$.*

4.2 Betting-style approach

We refer to our second algorithm as a *betting-style* approach, due to the way we construct our test statistics [Shafer, 2021, Waudby-Smith and Ramdas, 2024, Chugg et al., 2024, Vovk and Wang, 2021]. We direct the reader to these references for more detailed technical exposition. The betting-style approach rests on the key idea that we can pose the problem of rejecting some null hypothesis $\mathcal{H}_0^G$ as that of 'making money' by 'betting' against it. We concretize this test as follows.

Updating the test statistic ω_t^G. We let ω_t^G be the logarithm of the 'wealth' accumulated up to some time t given a series of bets $\lambda_1, \ldots, \lambda_t \in [0, 1/\beta\mu_G^0]$,[3] (where ω_0 and λ_0 are taken to be equal to 0). This corresponds to the following update rule for the quantity ω_t^G, initializing $\omega_0^G = 0$:

$$\omega_t^G \leftarrow \omega_{t-1}^G + \ln\left(1 + \lambda_t^G(\mathbf{1}_{X_t\in G} - \beta\mu_G^0)\right) \tag{4}$$

To minimize stopping time, we would like to set λ_t in a way that maximizes the test statistic ω_t^G. Luckily, the problem of 'portfolio optimization' [Cover, 1991] already has been well studied in the online learning literature [Zinkevich, 2003, Hazan et al., 2016, Cutkosky and Orabona, 2018], and there exists a concrete rule that ensures the resulting ω_t^G is not too far from the best achievable in expectation. This strategy, called Online Newton Step [Hazan et al., 2007], amounts to the following update for $\{\lambda_t\}_{t\geq 1}$:

$$\lambda_{t+1}^G \leftarrow \underset{[0,1/(\beta\mu_G^0)]}{\text{Proj}}\left(\lambda_t^G + \frac{2}{2-\ln(3)}\frac{z_t}{1+\sum_{s\in[t]} z_s^2}\right). \tag{5}$$

where we denote $z_t = \dfrac{\mathbf{1}[X_t \in G] - \beta\mu_G^0}{1 + \lambda_t^G(\mathbf{1}[X_t \in G] - \beta\mu_G^0)}$, and set $\lambda_0 = 1/2$.

Setting the threshold $\theta_t(\alpha)$. The way we set ω_t^G in Equation (4) ensures that anytime-validity at level α is preserved if we reject $\mathcal{H}_0^G$ as soon as $\omega_t^G > \log(1/\alpha)$. That is, under $\mathcal{H}_0^G$, we have that $\Pr[\exists t : \omega_t^G > \log(1/\alpha)] \leq \alpha$.[4] Adding a Bonferroni correction, this gives $\theta_t(\alpha) := \log(|\mathcal{G}|/\alpha)$ for all t.

[2] That said, in practice, this does not appear to be a a problem (see Section 5).

[3] To interpret this quantity, note that taking $\lambda_t = 0$ means the wealth remains the same regardless of what happens in the next round. On the other hand, $\lambda_t = 1/\beta\mu_G^0$ means that if we receive evidence in accordance with the null we lose all the 'money', but, if we receive evidence *against* the null, i.e. $X_t \in G$, we maximally increase ω_t^G.

[4] This follows directly from the prior work referenced at the beginning of this section; at a high level, every sequence $\{\exp(\omega_t^G)\}_{t\geq 1}$ is a non-negative super-martingale; applying Ville's inequality provides a validity guarantee directly. $\exp(\omega_t^G)$ can also be referred to as an *e-value* [Vovk and Wang, 2021], a measure of evidence against a null hypothesis similar to a p-value.

Theoretical guarantees. The validity guarantees of this approach are very similar to those for the Z-test style approach; it is always valid in finite samples.

Theorem 4.3 (Validity)**.** *Running Algorithm 1 with* $\theta_t(\alpha) = \log(|\mathcal{G}|/\alpha)$ *and* ω_t^G *updated as per Equations* (4) *and* (5) *guarantees that the probability that* $\mathcal{G}^{Flag}$ *will ever contain a group* G *where its corresponding null* $\mathcal{H}_0^G$ *is true is at most* α*, i.e.*

$$\Pr\left[\exists t : \exists G \in \mathcal{G}^{Flag} \text{ s.t. } \mathcal{H}_0^G \text{ holds}\right] \leq \alpha.$$

Finally, a direct consequence of adapting the analysis of Chugg et al. [2024] is the following theorem bounding the expected stopping time of our procedure:

Theorem 4.4 (Power)**.** *Let* T *be the stopping time of Algorithm 1 with* $\theta_t(\alpha) = \log(|\mathcal{G}|/\alpha)$ *and* ω_t^G *updated as per Equations* (4) *and* (5)*. Let* $\Delta_G := \mu_G - \beta\mu_G^0$ *and* $\Delta_{\max} = \max_{G\in\mathcal{G}} \Delta_G$*. Then, if* $\Delta_{\max} > 0$*, this test is a test of power one, i.e.* $\Pr[T < \infty] = 1$*. Furthermore,*

$$\mathbb{E}[T] \leq \mathcal{O}\left(\frac{\log(|\mathcal{G}|/\alpha) + \log(1/\Delta_{\max}^2)}{\Delta_{\max}^2}\right).$$

Note that the above corresponds to the stopping time guarantee for running the procedure on only the group with the biggest gap *plus* a term equal to $\log(|\mathcal{G}|)$ in the numerator. This means that, in terms of worst-case guarantee on stopping time, the contribution of the Bonferroni correction is small relative to the contribution of the test level α and, especially, to the gap $\Delta_{\max}$.

5 Real-World Examples

To demonstrate the applicability of our approach, we apply our framework to two real-world datasets. We begin by showing that Algorithm our approach can correctly (and quickly) identifies that young men experience myocarditis after the COVID-19 vaccine; then, on mortgage allocation data, we show that we identify known instances of discrimination under many reasonable reporting models. See Appendix C for dataset details and further discussion.

5.1 Myocarditis from COVID-19 vaccines

It is by now well-known that COVID-19 vaccines appear to induce elevated risk of myocarditis among young men [Oster et al., 2022] . While initial suspicions of elevated myocarditis risk relied on case studies, a more systematic understanding—including the pattern of disproportionate impact on on young men—was made possible by post-hoc analysis of incident databases. If we had been able to run the hypothesis tests proposed in the preceding sections on the reports collected in VAERS, would we have correctly identified this problem—and if so, how quickly?

Concretely, we let Y_i be the event that individual i experiences myocarditis after receiving a COVID vaccine, and run the test with the end-goal of identifying elevated incidence rate $\Pr[Y_i \mid X_i \in G]$ for group(s) G corresponding to adolescent men (ages 12-17 and 18-29).

Setting β**.** For this application, absolute incidence rate (that is, $\Pr[Y = 1 \mid G]$) is the quantity of interest to use for determining β. As suggested by Proposition 3.4, setting β requires considering three quantities of interest: the threshold on an "unacceptable" incidence rate, the relative rates of true reporting γ_G^{TR}, and the relative rates of false reporting γ_G^{FR}. Then, we can set $\beta = \max_G\left((\gamma_G^{\text{TR}} - \gamma_G^{\text{FR}}) \cdot \text{IR} + \gamma_G^{\text{FR}}\right)$.

We will choose 0 as the threshold on an "unacceptable" incidence rate.[5] It is therefore sufficient to set $\beta = \max_G(\gamma_G^{\text{FR}})$. While this is quantity cannot be determined from report data alone, a conservative assumption could be that any group erroneously reports at most twice the average reporting rate over the

[5]One might follow existing practice and use the per-group expected rate of myocarditis to benchmark an unacceptable incidence rate (e.g. as provided in Table 2 of Oster et al. [2022], which suggests at most 2 cases per million doses). However, in addition to this expected incidence rate being very small (and, for any practical purposes, being vastly dominated by the other reporting terms), it also implicitly relies on reports so that the benchmark quantities are $r \cdot \text{IR}$, rather than just IR, and thus depend on the unknown reporting rate r.

Table 1: On real historical sequence of myocarditis reports, time to identification of harmed groups. In each cell, we report the number of total reports to the rejection of the hypothesis corresponding to (M, 18-29) and the number of total reports corresponding to (M, 12-17). In all tests, the (M, 18-29) group is identified first. **Note that for the asymptotic Z-test, a minimum stopping time of 100 was enforced for the purposes of asymptotic validity.*

	*Asymptotic Z-test**	*Finite-sample Z-test*	*Betting-style test*
$\beta = 2.0$	101 *(M, 18-29)*; 256 *(M, 12-17)*	69; 530	61; 241
$\beta = 2.5$	101; 302	74; 546	69; 259
$\beta = 3.0$	101; 324	111; 612	80; 302

population, with $\gamma_G^{\text{FR}} = 2.0$. If the algorithm is first $\beta = 2.0$, stopping and flagging a group very quickly, it may be natural to re-run the test with increasing values of β, as a higher β corresponds to a more severe true incidence rate, so we also test $\beta = 2.5$ and $\beta = 3$.[6]

Results. We begin by running our algorithms on the actual sequence of reports in chronological order, as received in VAERS. In particular, we consider Algorithm 1 instantiated with ω_t^G updated according to Equation (2) and $\theta_t(\alpha)$ as in (3) and $C = 1/2$ (*Asymptotic Z-test*); with ω_t^G updated according to Equation (2) and $\theta_t(\alpha)$ as in (3) and $C = \sqrt{\beta\mu_G^0(1-\beta\mu_G^0)}$ (*Finite-sample Z-test*), and with ω_t^G updated according to Equations (4) and (5), and $\theta_t(\alpha)\ln(|\mathcal{G}|/\alpha)$ (*Betting-style test*). We run all experiments for $\alpha = 0.1$. In Table 1, we report the stopping time—that is, the number of reports it takes for the first null to be rejected—of each algorithm for various values of β.

To explore the robustness of these results, we also run synthetic experiments, permuting the ordering of reports to get a sense of possible variance in the stopping time. We run 100 random permutations of the full set of reports. In Figure 1 1, we compare the performance of various algorithms on this set of reporting data. Each point on these plots reflects the number of trials (out of 100) in which a rejection has occurred by time t. Figure 1 tracks the number of reports it takes for each algorithm to reject the null hypothesis for any group—that is, a scenario when the test is stopped and an alarm is raised as soon as one harmed group is identified. To interpret the figure, the finite-sample z-test (`lilt`) stopped by time $t = 400$ in around 45 of 50 trials; in all 50 trials, it stopped by time $t = 700$. Overall, Figure 1 shows that the asymptotically-valid sequential z-test (dashed, red) identifies a harmed group the most quickly, but the e-value algorithm (solid, dark purple) performs very similarly.

Our experimental results suggest that our proposed tests would in fact have been effective in determining that young men were disproportionately affected by myocarditis. Moreover, though it is difficult to determine exact timelines and the nature of clinical practice during early phases of the vaccine rollout, it is possible that such a test could have identified problems using less data—that is, more quickly—than was actually used for this finding.

5.2 Mortgage Allocations

In 2021, Martinez and Kirchner [2021] found that, based on publicly-released data from the Home Mortgage Disclosure Act (HMDA), substantial racial disparities in 2019 loan approvals persisted even after controlling for financial characteristics of applicants—most notably, healthy debt-to-income ratios (DTI). If home loan applicants had been able to submit reports when they believed they had experienced unfavorable outcomes, could we have used those reports to discover the discrimination identified by Martinez and Kirchner [2021]—and if so, how accurately, and how quickly?

Unlike the COVID vaccine case study, we are interested primarily in disparity among applicants with healthy DTI, even though all loan applicants would have been eligible to submit reports. Concretely, we let $A_i = 0$ be the event that a loan is not made to applicant i, and $Z_i = 1$ be the event that applicant i has a healthy debt-to-income ratio. Then, we let $Y_i = \{A_i = 0, Z_i = 1\}$ be the event that individual i has a

[6]Note that re-using the data here is statistically valid due to the equivalence between one-sided hypothesis testing and confidence sequences.

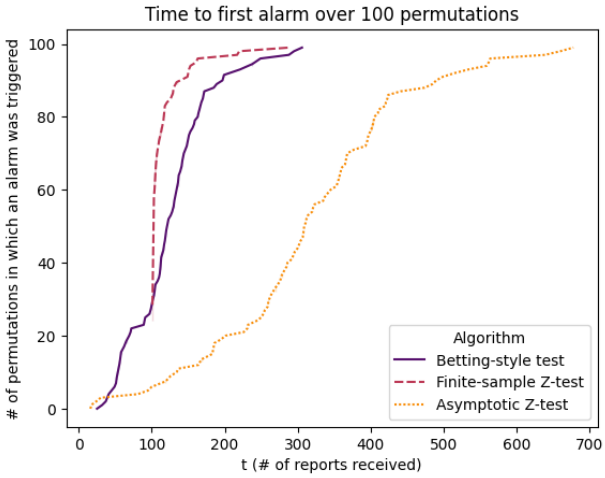

Figure 1: COVID experiment, with $\beta = 3$, over 100 random permutations of report database. Comparison across algorithms of the number of steps (t) it takes for each algorithm to reject the null hypothesis for any group—that is, stopping time, if the algorithm had been stopped as soon as one harmed group is identified. The asymptotically-valid sequential z-test (dashed, red) identifies a harmed group the most quickly, but the e-value algorithm (solid, dark purple) performs very similarly.

healthy DTI and did not receive a loan, and run the test with the end-goal of identifying groups that have relatively high rates of loan denials for applicants with healthy DTI, i.e. $\frac{\Pr[A_i=0, Z_i=1 | X_i \in G]}{\Pr[A_i=0, Z_i=1]}$.

Reporting models. The existence of verifiable disparities in this dataset allows us to evaluate the efficacy of our methods under varying models of reporting—that is, whether the groups returned by our algorithms do in fact reflect groups with high rates of healthy DTI denials, even if it is not the case that every report X_i corresponds to Y_i actually occurring. The dataset gives several levels of financial health with respect to DTI—in ascending order, "Struggling", "Unmanageable," "Manageable," and "Healthy." Modeling the idea that reporting behavior may be related to financial health, we use these categories to simulate four models for reporting.

(1) *Healthy DTI:* The (ideal) case where reports *only* come from individuals with "Healthy" DTI and were denied a loan.

(2) *Correlated:* A slightly more realistic case where "Healthy" denials report with probability 0.9, "Manageable" 0.5, "Unmanageable" 0.3, "Struggling" 0.1.

(3) *All Denials:* All denials submit reports regardless of financial health.

(4) The *Anti-Correlated:* The (unlikely) case where individuals with worse financial health are more likely to report, i.e. "Healthy" denials report with probability 0.1, "Manageable" 0.5, "Unmanageable" 0.7, "Struggling" 0.9.

Setting β. In this application, the quantity of interest is relative risk, so we draw on Proposition 3.2 to inform our setting of β. We will set our relative risk threshold to be 1.2—that is, we want our algorithm to raise an alarm when any group experiences event Y 20% more frequently than average over the population. We test at $\beta = 1.4$, reflecting a maximum group-conditional RIR (recall Definition 3.1) of $7/6$; $\beta = 1.6$, which corresponds to $\max_G \text{RIR}_G \leq 8/7$; and $\beta = 1.8$, which corresponds to $\max_G \text{RIR}_G \leq 9/7$.

Results. As with the COVID case study, we run all three algorithms discussed in Section 4; we also run them for all four reporting models discussed above, and for $\beta = \{1.4, 1.6, 1.8\}$. For each combination of algorithm, reporting model, and β, we again run 100 random permutations, and terminate the algorithm after 10,000 steps. [7]

One important question for this application is the extent to which our tests identify the type of harm we are interested in, depending on various reporting models: while the algorithms guarantee statistical validity in terms of overrepresentation (i.e., in terms of whether $\mu_G \geq \beta \mu_G^0$), they cannot intrinsically guarantee

[7]Note that since we are simulating reporting, we cannot run any experiment for the true historical sequence of reports, as we did in Table 1.

Table 2: Percentage of groups in $\mathcal{G}^{\text{Flag}}$ where $\text{RR}_G \geq 1.5$, on average over 100 trials, as well as median size of $\mathcal{G}^{\text{Flag}}$. Results are shown for the *Asymptotic Z-test*, which was allowed to run for 10,000 steps.

	Healthy DTI	*Correlated*	*All Denials*	*Anti-Correlated*
$\beta = 1.4$	93.7% ($\lvert\mathcal{G}^{\text{Flag}}\rvert \approx 8$)	92.3% (13)	76.1% (20)	59.0% (28)
$\beta = 1.6$	88.3% (3)	94.7% (6)	91.1% (10)	78.2% (16)
$\beta = 1.8$	72.9% (2)	90.4% (2)	89.6% (3)	85.3% (7)

that reports themselves reflect true harm. With the benefit of hindsight (and access to the full dataset), we are able to calculate "ground truth" incidence rates. It turns out that a relative risk of 1.5 corresponds approximately to the 75th percentile of $\Pr[Y_i \mid G]$, and is comprised of entirely Black or Latino groups (across age and gender). On the other hand, a relative risk of 1.2 is actually around the 50th percentile—that is, half of the 115 groups have a (true) relative risk that should have been sufficiently high to trigger the alarm. This discrepancy can be explained entirely by the relative sizes of each group; the larger groups, which correspond to majority or more privileged groups, also have relative risk that is lower than average.

In Table 2, we show results for the asymptotically-valid Z-test for a variety of reporting models. We highlight two notable phenomena here. First, when $\beta = 1.4$—which means that the test itself should provide no guarantees about identifying groups with $\text{RR}_G > 1.5$, even if reporting across groups exactly equal—the test still identifies a $\mathcal{G}^{\text{Flag}}$ that is comprised mostly of groups that do experience a true $\text{RR}_G > 1.5$; this is consistent with the theory in that the more severely-impacted groups will be identified sooner. On the other hand, though there are around 25 groups with $RR_G > 1.5$, there are far fewer than 25 groups in $\mathcal{G}^{\text{Flag}}$ regardless of the β at which the test was run. The likely explanation for this is that when groups themselves are very small, it is more difficult to gather sufficient samples to determine that a gap exists; and it is possible that running the test for more than 10,000 steps would have identified them.

6 Discussion

This work is an initial approach to using incident databases for post-deployment auditing; we believe there is a rich range of future work that develops the ideas in this paper, both technically and conceptually.

On the statistical and algorithmic side, because our framework allows for plugging in any existing sequential test, new methods that control for multiple hypothesis testing both over time and over the number of distinct hypotheses would be directly beneficial for this application. On the other hand, one might hope for online methods that do not require pre-specifying hypotheses and instead develops them sequentially in a quasi-unsupervised fashion, or that improve guarantees by exploiting relationships across hypotheses, as has proven useful in multi-objective learning.

More conceptually, while the application examples in Section 5 are somewhat stylized, they demonstrate that incident databases can be promising starting points for new types of post-deployment evaluation. For incident databases to be practically useful, there are a plethora of additional considerations to incorporate from a variety of disciplines. For instance, if a reporting system was available, how would individuals engage with them in theory, and in practice? To what extent do, and should, individual incentives affect the database, and how it is designed? How can the result of a test (a null hypothesis rejection) be contextualized by existing and emerging legal frameworks?

To the best of our knowledge, we are the first to propose individual incident reporting to identify patterns of disproportionate harm in interactions with a particular system; more generally, however, one might imagine that similar reporting systems can be developed to provide insights about concerns beyond fairness. In fact, while the framework introduced in our work is not intrinsically about algorithmic deployments, it is one way to operationalize recent regulatory movement in AI policy towards allowing for or requiring individual reports. Any way to make such reports actionable at large scale must, to some extent, aggregate of individual reports to develop more systematic evaluations of an underlying algorithm. We therefore see our work as one step towards giving voice to individual experiences—and towards having them make a difference.

Acknowledgements

We are grateful to Ian Waudby-Smith, Kevin Jamieson, and Robert Nowak for helpful discussions in developing this work.

JD is supported in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. JD also thanks the AI Policy Hub at UC Berkeley for funding support in the 2023-2024 academic year. BR is generously supported in part by NSF CIF award 2326498, NSF IIS Award 2331881, and ONR Award N00014-24-1-2531. DIR is supported by the Mozilla Foundation.

References

Hammaad Adam, Fan Yin, Huibin Hu, Neil Tenenholtz, Lorin Crawford, Lester Mackey, and Allison Koenecke. Should i stop or should i go: Early stopping with heterogeneous populations. *Advances in Neural Information Processing Systems*, 36, 2024.

Gabriel Agostini, Emma Pierson, and Nikhil Garg. A bayesian spatial model to correct under-reporting in urban crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21888–21896, 2024.

Marco Almada. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on artificial intelligence and law*, pages 2–11, 2019.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Jay Bartroff and Jinlin Song. Sequential tests of multiple hypotheses controlling type i and ii familywise error rates. *Journal of statistical planning and inference*, 153:100–114, 2014.

Yewon Byun, Dylan Sam, Michael Oberst, Zachary Lipton, and Bryan Wilder. Auditing fairness under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 4339–4347. PMLR, 2024.

Sarah H Cen and Rohan Alur. Ai auditing and the access question: Exploring black-box auditing and its connection to hypothesis testing.

Robert T Chen, Suresh C Rastogi, John R Mullen, Scott W Hayes, Stephen L Cochi, Jerome A Donlon, and Steven G Wassilak. The vaccine adverse event reporting system (vaers). *Vaccine*, 12(6):542–550, 1994.

John J Cherian and Emmanuel J Candès. Statistical inference for fairness auditing. *arXiv preprint arXiv:2305.03712*, 2023.

Ziyu Chi, Aaditya Ramdas, and Ruodu Wang. Multiple testing under negative dependence. *arXiv preprint arXiv:2212.09706*, 2022.

Brian Cho, Kyra Gan, and Nathan Kallus. Peeking with PEAK: Sequential, Nonparametric Composite Hypothesis Tests for Means of Multiple Data Streams, June 2024.

Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas. Auditing fairness by betting. *Advances in Neural Information Processing Systems*, 36, 2024.

Thomas M Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529. PMLR, 2018.

Jessica Dai, Nika Haghtalab, and Eric Zhao. Learning with multi-group guarantees for clusterable subpopulations. *arXiv preprint arXiv:2410.14588*, 2024.

European Parliament. Eu ai act: First regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2024-12-17.

Michael Feffer, Nikolas Martelaro, and Hoda Heidari. The ai incident database as an educational tool to raise awareness of ai harms: A classroom exploration of efficacy, limitations, & future improvements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11, 2023.

Jean Feng, Alexej Gossmann, and Gene Pennello. Monitoring machine learning-based risk prediction algorithms in the presence of performativity.

Annelise Gilbert. Workday ai biased against black, older applicants, suit says, February 2023. URL https://news.bloomberglaw.com/daily-labor-report/workday-ai-biased-against-black-disabled-applicants-suit-says.

Ira Globus-Harris, Michael Kearns, and Aaron Roth. An algorithmic framework for bias bounties. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1106–1124, 2022.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Naveena Karusala, Sohini Upadhyay, Rajesh Veeraraghavan, and Krzysztof Z Gajos. Understanding contestability on the margins: Implications for the design of algorithmic decision-making in public services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.

Martin Kulldorff, Robert L Davis, Margarette Kolczak, Edwin Lewis, Tracy Lieu, and Richard Platt. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential analysis*, 30(1):58–78, 2011.

Susan Landau, James X Dempsey, Ece Kamar, Steven M Bellovin, and Robert Pool. Challenging the machine: Contestability in government ai systems. *arXiv preprint arXiv:2406.10430*, 2024.

Katharine MN Lee, Tamara Rushovich, Annika Gompers, Marion Boulicault, Steven Worthington, Jeffrey W Lockhart, and Sarah S Richardson. A gender hypothesis of sex disparities in adverse drug events. *Social Science & Medicine*, 339:116385, 2023.

Zhi Liu and Nikhil Garg. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1016–1017, 2022.

Zhi Liu, Uma Bhandaram, and Nikhil Garg. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 4(1):57–65, 2024.

Emmanuel Martinez and Lauren Kirchner. The secret bias hidden in mortgage-approval algorithms. *The Markup*, 2021.

Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021.

Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling. *arXiv preprint arXiv:2402.17861*, 2024.

Matthew E Oster, David K Shay, John R Su, Julianne Gee, C Buddy Creech, Karen R Broder, Kathryn Edwards, Jonathan H Soslow, Jeffrey M Dendy, Elizabeth Schlaudecker, et al. Myocarditis cases reported after mrna-based covid-19 vaccination in the us from december 2020 to august 2021. *Jama*, 327(4):331–340, 2022.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021.

Divya Shanmugam, Kaihua Hou, and Emma Pierson. Quantifying disparities in intimate partner violence: a machine learning method to correct for underreporting. *npj Women's Health*, 2(1):15, 2024.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32, 2019.

Tom T Shimabukuro, Michael Nguyen, David Martin, and Frank DeStefano. Safety monitoring in the vaccine adverse event reporting system (vaers). *Vaccine*, 33(36):4398–4405, 2015.

Violet Turri and Rachel Dzombak. Why we need to know more: Exploring the state of ai incident documentation practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 576–583, 2023.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

Kristen Vaccaro, Karrie Karahalios, Deirdre K Mulligan, Daniel Kluttz, and Tad Hirsch. Contestability in algorithmic systems. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, pages 523–527, 2019.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Heather P Whitley and Wesley Lindsey. Sex-based differences in drug activity. *American family physician*, 80(11):1254–1258, 2009.

Ziwei Wu and Jingrui He. Fairness-aware model-agnostic positive and unlabeled learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1698–1708, 2022.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Additional Related Work

A.1 Technical discussions

Relationship to Chugg et al. [2024] Perhaps the most closely related work is that of Chugg et al. [2024], who propose applying a sequential hypothesis test with the explicit goal of quickly identifying bias in deployed systems in real time; they also draw from the testing-by-betting framework, and our e-value algorithm utilizes the betting strategy (i.e., λ updates via ONS) given in their work. However, we propose fundamentally different tests—their test is of equality of means across different groups, while ours compares within groups—leading to a few key differences. Perhaps most significantly, their test assumes access to one sample from each group at each timestep; the test must wait until a new sample arrives from each group. This naturally extends the time horizon, possibly dramatically, when groups are imbalanced. More precisely, in expectation, each timestep would require waiting for $\sum_{G\in\mathcal{G}} \frac{1}{\mu_G}$ samples, which is lower bounded by $|\mathcal{G}|^2$.[8] Additionally, the method assumes that the labels and predictions of samples X are known. Our test, on the other hand, is designed explicitly to bypass the fundamental problem that, in many cases, labels (or predictions) are unknown.

Relationship to Adam et al. [2024] The work of Adam et al. [2024] provides a method for *early stopping* in (sequential) RCTs in the case that there exists some subgroup with a negative (average) treatment effect. From an algorithmic point of view, this is a very similar problem setup: we have a problem where data arrives sequentially, and the goal is to stop early if disparate harm is detected. However, their application context is quite different; other key differences include that algorithm proceeds in batch-like phases, and provide primarily asymptotic theoretical guarantees.

A.2 Policy and application context

Comparison to current practice in VAERS. As outlined in Shimabukuro et al. [2015], reports from the Vaccine Adverse Event Reporting System (VAERS) are used to "detect vaccine safety signals," but not to rigorously *determine* safety. VAERS functions as a means to continuously monitor vaccines that are licensed in the U.S., with emphasis on vaccines that are high use (e.g. flu), newly-approved vaccines, and new rollout policies/recommendations for existing vaccines.

The core statistical component of VAERS data analysis relies primarily on descriptive analysis and comparison to historical trends, specifically via disproportionality analysis. In particular, VAERS uses a "proportional reporting ratio" (PRR), defined $\text{PRR} = \frac{V_iE_j/(V_iE_j+V_iE_x)}{V_xE_j/(V_xE_j+V_xE_x)}$, where V_iE_j indicates the number of reports of adverse event j for vaccine i, and V_xE_x indicates the number of reports of any other adverse event for any other vaccine. Qualitatively, the numerator captures how frequent event E_j was relative to all adverse reports for vaccine V_i, while the denominator captures how frequently event E_j is reported for all other vaccines V_x. PRR therefore represent how much *more* event E_j tends to happen for the particular vaccine V_i, as compared to other vaccines.

B Practical considerations

We conclude this section with a brief discussion of modeling decisions that are necessary for practical implementations of our proposals.

Choosing $\mathcal{G}$. In our experiments in Section 5, we choose to define subgroups as all possible combinations of available demographic characteristics. That said, a practitioner may seek to define $\mathcal{G}$ more carefully in accordance with their application. For instance, if the goal is to illustrate discrimination in a legal sense, $\mathcal{G}$ should be defined with respect to (protected) demographic features, rather than arbitrary combinations of covariates. On the other hand, $\mathcal{G}$ could include which batch of a medication an individual received; our tests could then help identify whether some batches were improperly manufactured.

[8] Technically the sum is multiplied by $\max_G \mu_G$, i.e. scaled by frequency relative to the biggest group.

Baseline rates $\{\mu_G^0\}_{G\in\mathcal{G}}$. A natural question that arises from the modeling in this section is how $\{\mu_G^0\}_{G\in\mathcal{G}}$ can be determined, or if Assumption 2.1 is strictly necessary. Practically speaking, these base preponderances may be estimated, possibly with some amount of noise. However, for ease of exposition, we assume we have access to the true, underlying values of $\{\mu_G^0\}_{G\in\mathcal{G}}$, as the estimation problem can be addressed with standard techniques and is not core to our contribution. Perhaps more significantly, in practice these baseline preponderances may change over time (e.g. if some subgroups increased uptake of a vaccine, or applied for loans more frequently, over time).

Note that testing against base preponderances of the reference population (i.e., to compare μ_G to μ_G^0) is a new test proposed by this work, and the analysis in Sections 3.1 and 3.2 is specific to this test. Existing approaches to monitoring in incident databases compare to different baselines, most commonly the historical overall incidence rate for the specific symptom, sometimes by subgroup [Shimabukuro et al., 2015, Kulldorff et al., 2011, Oster et al., 2022]. These baselines could, in principle, be plugged into the algorithms in Section 4, but new analysis for (possibly group-varying) reporting rates would be necessary to draw inferences about analogous quantities of interest (e.g., RR or IR), as current approaches do not generally consider reporting behavior. In contrast, our modeling allows us to identify what quantities may affect the true incidence rate even if they may be unmeasurable.

Setting β. Finally, to run the test proposed in Equation (1), it is necessary to determine how to set the value of β. As we will see in Section 4, when β is set too high, then the test may never identify problematic groups, or identify them more slowly; on the other hand, as is clear from the previous subsections, rejecting the null hypothesis for a smaller β requires more stringent assumptions on reporting behavior. Thus, we suggest a procedure to set β as follows: (1) choose a relative risk or incidence rate threshold where it would be problematic for any group if RR_G or IR_G surpassed that threshold; (2) make the corresponding assumptions about reporting behavior; (3) use these quantities to compute a reasonable β. We give some example computations in Section 5. Due to an equivalence between hypothesis testing and confidence intervals, it is statistically valid to rerun tests with different βs once data collection has begun. Thus, it may be prudent to begin by setting the lowest β that reporting assumptions would allow; then, if the tests appear to be stopping very quickly, to re-run them at higher βs, which would allow a practitioner to get a better sense of the severity of the harm.

C Additional Experimental Details

C.1 Additional information on VAERS experiments

Data sources. The Vaccine Adverse Event Reporting System (VAERS) is a national adverse event incident database for U.S.-licensed vaccines, co-managed by the Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA) Chen et al. [1994], Shimabukuro et al. [2015]. For this case study, we focus on reports of myocarditis after receiving a COVID-19 vaccine; we thus filter the set of publicly-available reports from VAERS accordingly. To determine per-demographic base rates, i.e. to compute $\{\mu_G^0\}_{G\in\mathcal{G}}$, we utilize VaxView, a government dataset tracking national vaccine coverage (publicly accessible here).

Defining $\mathcal{G}$. For this application, we consider (intersections of) sex and age buckets to be the subgroups of interest. In particular, age buckets are discretized into 0-4, 5-11, 12-17, 18-29, 30-39, 40-49, 50-64, 65-74, and 75+; the sex categories represented in the data are female, male, and unknown. In total, after removing groups for which no vaccines were known to have been given, $\mathcal{G}$ contains 29 total groups. While in principle it would have been interesting to also consider race/ethnicity, we are limited by the availability (and granularity) of the data given in VaxView.

C.2 Additional information on HMDA experiments

Data sources. We use the data (and preprocessing code) of Martinez and Kirchner [2021], which uses 2019 data from the HMDA.[9] The analysis of Martinez and Kirchner [2021] used the full year of data from 2019, which included over 183k denials.

Defining $\mathcal{G}$. While Martinez and Kirchner [2021] analyzed disparities with respect to race, we define groups as all possible intersections of demographic features across gender, race, and age. The available race categories include Native, Asian, Black, Pacific Islander, White, and Latino; sex categories include female, male, and unknown/nonbinary; and age categories include <25, 25-34, 35-44, 45-54, 55-64, and 65+. In total, after removing groups which comprise less than 0.1% of all loan applicants, $\mathcal{G}$ contains 115 groups.

[9]The Consumer Financial Protection Bureau (CFPB) collects and publishes this data from financial institutions annually, with a two-year lag; the report (and our work) uses 2019 data which is finalized as of Dec. 31, 2022. The most recent year for which data is available is 2022, though it is available for edits through 2025.