

What a Fairness Audit Must Report When Demographic Labels Are Missing: Assumption-Free Disparity Intervals, a Determinacy Rule, and a Disclosure Checklist

What a Fairness Audit Must Report When Demographic Labels Are Missing: Assumption-Free Disparity Intervals, a Determinacy Rule, and a Disclosure Checklist

Abstract

Fairness audits of machine-learning systems publish group-level rates such as selection rates, but the protected labels those rates depend on are frequently missing for a substantial fraction of the audited population, and no settled rule says what the audit must disclose in that case. The gap is practical: an oversight reader who sees a rate computed only on labeled records cannot tell from the published numbers alone whether the rate would survive, shrink, or reverse if the unlabeled records were labeled. We close the gap with an assumption-free worst-case analysis. For any published rate we derive the exact interval $[\hat{p}(1 - m), \hat{p}(1 - m) + m]$ containing the full-population rate under any pattern of missingness, where $\hat{p}$ is the labeled-subset rate and m the missing-label fraction; the interval width equals m exactly. For a two-group disparity the assumption-free interval has width $m_1 + m_2$, independent of the rates: a reported gap of 0.10 with missingness rates 0.6 and 0.5 is compatible with any true gap in $[-0.475, 0.625]$. We derive a determinacy rule — certification of a threshold $\delta_{\min}$ is impossible exactly when the shifted threshold $\delta_{\min} + p_B$ lies inside the interval $[p_{\min,A}, p_{\max,A}]$, while $m < \delta_{\min}$ guarantees that some certification is possible whatever the point estimate — and instantiate it: with $m = 0.2$, an observed rate of 0.30 against a fully labeled comparison rate of 0.35 gives a four-fifths ratio interval $[0.686, 1.257]$, so compliance cannot be certified even though the point estimate 0.857 suggests it, and 134 labels must be recovered to change that. The framework converts incomplete labels from an unstated caveat into a computed, reportable quantity with a concrete remediation target.

1. Introduction

Fairness audits are a key component of responsible machine-learning deployment [1], and their basic currency is the group-level rate: a selection rate, a false-positive rate, a referral rate, computed per protected group and compared across groups. Producing those rates requires knowing, for each person in the audited population, which group they belong to. In operational data that knowledge is routinely incomplete: the data needed to compute these rates is largely unavailable in practice, which hinders the adoption of algorithmic fairness, and collecting it carries risks that the literature characterizes [2]; a substantial fraction of audited records therefore carries no usable group label. The audit then publishes a rate whose denominator is silently the labeled subset, and the reader has no way to know how much of the population the published number actually describes.

The cost of this gap is concrete. An oversight reader who sees "group A is selected at 25%, group B at 15%" cannot currently tell from the report whether that ten-point gap would survive, shrink, or reverse if the unlabeled records were labeled. The reader therefore faces a decision — accept the audit, demand remediation, or halt deployment — with a number whose evidentiary value is unquantified. A system can pass an audit on labeled records while its behavior on unlabeled records is materially different, and no published number flags the possibility. Conversely, a model that appears unfair under the observed data may be acceptable once the missing-label uncertainty is accounted for.

What would change if the gap were closed? The oversight reader would see, next to every published rate, an interval stating how far the true rate could deviate given the missing labels, together with a determinate/indeterminate verdict: either the audit certifies that the disparity is (or is not) larger than a stated threshold of concern, or it reports that the missing labels make certification impossible and states how many labels would have to be recovered to change that. Audit teams would get a design rule — a minimum labeled fraction — before deployment rather than a discovery of indeterminacy after it.

This paper builds that disclosure rule from first principles, and its stance is deliberately conservative. We make no assumption about the mechanism that causes labels to be missing — not missing at random, not ignorable, not adversarial — because in the audit setting the missingness may be precisely what an adversarial or negligent process produces. Instead we compute worst-case bounds over all assignments of the unlabeled records; the bounds depend on only two numbers the audit already has, the labeled fraction and the observed rate among labeled records. The worst case is tight, so the intervals we report are the narrowest any assumption-free method can give; any narrower interval smuggles in an assumption about missingness that the audit must then state and defend. The price of this conservatism is width, and a second contribution is to quantify that price so an audit designer can decide when assumption-free reporting is informative and when it is not.

Our contributions are:

- An exact worst-case interval $[\hat{r}_g(1 - m_g), \hat{r}_g(1 - m_g) + m_g]$ containing any group's full-population rate given its labeled-subset rate $\hat{r}_g$ and missingness rate m_g , with width exactly m_g , derived in Section 4.1.
- The result that the width of the assumption-free interval on a two-group disparity equals $m_1 + m_2$, independent of the rates, derived in Section 4.2, with the worked example $[-0.475, 0.625]$ for a reported gap of 0.10.
- A determinacy rule — certification of a threshold $\delta_{\min}$ is impossible exactly when $\delta_{\min} + p_B$ lies inside the interval $[p_{\min,A}, p_{\max,A}]$, with $m < \delta_{\min}$ sufficient for some certification whatever the point estimate, and a reported positive gap surviving scrutiny if $\hat{g} > m_1 + m_2$ — derived in Section 4.3, together with the exact necessary-and-sufficient condition and a counter-example showing the shortcut is not necessary (Section 4.4).
- A worked numerical audit in the mixed-completeness case: for $n = 1000$, $k = 800$, $\hat{p}_A = 0.30$ against a fully labeled comparison rate of 0.35, the rate interval is $[0.24, 0.44]$, the four-fifths ratio interval is $[0.686, 1.257]$, the verdict is indeterminate, and the label-recovery target is 134 labels (Section 4.5).
- A combined treatment of missingness and finite-sample binomial error, showing that at near-complete labeling ($m_1 = m_2 = 0.02$, $n_g = 980$) sampling contributes a half-width of approximately 0.0542 against 0.04 of missingness width, and that at $m = 0.2$ missingness contributes roughly 7 times more uncertainty than sampling (Section 4.6).
- A disclosure checklist specifying the minimum set of numbers an audit must publish when labels are incomplete, presented in Section 5 and discussed in Section 7.

2. Background and Related Work

The starting point for this paper is [1], which asks what a fairness audit must disclose when the protected labels it depends on are incomplete, and focuses on the rates an audit publishes and on what an oversight reader needs beside them; its approach pairs every published rate with a matched baseline drawn from the same audit. Our work is a direct formalization of one part of that program: we take the published rate and its matched context and derive exactly what the reader can and cannot conclude from them when the labeled subset is a proper subset of the audited population. Where [1] frames the disclosure problem, this paper supplies the arithmetic that makes a specific disclosure — the missingness rate — load-bearing.

The reason labels are incomplete is treated in [2], whose summary states that most proposed algorithmic fairness techniques require access to data on a sensitive attribute or protected category such as race, ethnicity, gender, or sexuality in order to make performance comparisons across groups, that such data is largely unavailable in practice, and that this unavailability hinders the widespread adoption of algorithmic fairness; the paper considers calls to collect such data and characterizes the risks of doing so. Our analysis is complementary: rather than asking whether to collect more demographic data, we ask what an audit can honestly report given the collection level it already has. The supplied summary of [2] does not state quantitative results, so we connect to it only at the level of the collection-versus-disclosure trade-off.

The failure mode of auditing too few groups is documented in [6], which studies statistical fairness definitions that fix a small collection of pre-defined groups and ask for parity of some statistic across them, and shows such constraints are susceptible to intentional or inadvertent "fairness gerrymandering," in which a classifier appears fair on each individual group but violates fairness on a larger structured set of subgroups; the work proposes auditing and learning for subgroup fairness over richly many groups. This matters for our setting in a specific way: missingness rates are a property of groups, and if the audit's group partition is too coarse, the missingness it discloses for a coarse group can conceal heterogeneous missingness across the intersections where [6] shows violations hide. Moreover, an audit that widens its group structure per [6] multiplies the number of rates it publishes, and each additional rate on a smaller subgroup has a larger effective missing-label fraction, so our interval widths grow exactly where gerrymandering risk is highest. Our disclosure rule therefore applies per group, and Section 6 discusses the compositional limitation.

A structural analogy for our worst-case analysis comes from the Byzantine-resilience literature. [3] studies distributed optimization with m worker machines of which t may be corrupt and may collaboratively and arbitrarily deviate from their pre-specified programs, with a master node computing the model parameter vector for generalized linear models. [4] studies Byzantine-resilient SGD in a master-worker architecture over heterogeneous data with no probabilistic assumptions on data generation, using a polynomial-time outlier-filtering procedure for robust mean estimation. Both lines of work share the design pattern we adopt here: assume the worst admissible deviation — in our case, an arbitrary assignment of outcomes and group memberships to unlabeled records — and derive guarantees that hold under it. The supplied summaries of [3] and [4] describe distributed optimization and SGD settings, not audit settings, so we use them only as methodological precedents for assumption-free worst-case reasoning, not as sources of audit-specific results.

Two remote-sensing papers mark the boundary of the related work. [7] proposes TerraGen, a unified multi-task layout generation framework for remote sensing data augmentation addressing task-isolated generative augmentation and unmodeled geographical and spatial constraints, and [8] proposes AeroGen, a diffusion-driven data generation method for remote sensing object detection motivated by scarcity of labeled data. Both respond to label scarcity by synthesizing labeled data; our setting forbids that response, because a synthesized protected label is precisely the unquantified inference an audit cannot silently rely on. The contrast is instructive: where the task permits generative fill-in, scarcity is an engineering problem; where the label is protected and audited, scarcity is an epistemic limit that must be disclosed. Neither summary supports any claim about fairness auditing, so we cite [7] and [8] only for this contrast.

A distant but clarifying example of inference under data limitations outside machine learning is [5], which constrains dark-energy cosmological model parameters using H II starburst galaxy apparent magnitude versus redshift data and reports constraints that are generally consistent with other datasets but not as restrictive as the tightest available. The pattern — a data-limited analysis that is honest about being less restrictive than the best

achievable — is the reporting posture we formalize: an audit under missingness should publish its bound and state plainly that it is less restrictive than a complete-label audit would be. The supplied summary of [5] supports nothing more specific than this analogy.

Finally, three entries inform the epistemic framing. [11] describes the Universal Ignorance Audit, a fifteen-question, five-phase method for interrogating the structure of not-knowing in any domain, which treats ignorance as an active, structured state with architecture rather than an absence to be filled. [10] synthesizes three threads from one research organization's experience with AI-assisted research: the development of that fifteen-question Universal Ignorance Audit, two independent forensic analyses concluding that a flagship published paper was AI-generated, and epistemic lessons of the pipeline. [9] describes a falsifiability-first audit program applied to a cluster of physics claims (photon-primacy rest energies, helical-electron spin, adelic mass ratios on Bruhat–Tits trees); its supplied summary gives no further detail on its methods. We use these works for one methodological commitment that shapes the whole paper: a reported number should be published together with the structure of what is not known about it, and a claim should be stated so that it can fail. Our missingness interval and determinacy rule are that commitment applied to fairness audit statistics. Beyond this framing role, the supplied summaries of [9], [10], and [11] support no quantitative claim used here.

Across these works, no entry in the supplied bibliography computes a certifiability threshold for group rates under label missingness; that computation is the contribution of Sections 3 through 5.

3. Methods

3.1 Setting and notation

An audit covers a population of n records partitioned into protected groups $g = 1, \dots, G$; the two-group case carries the derivations and the extension to more groups is by repetition. Each record has an outcome $Y \in \{0, 1\}$ used by the audit (for example, selected or not selected) and, in principle, a group label. For group g , let N_g be the number of records truly in group g , $n_g \leq N_g$ the number with a usable label, and k_g the number of labeled records with outcome 1. Define:

$$m_g = 1 - \frac{n_g}{N_g} \quad (\text{missingness rate of group } g),$$

$$\hat{r}_g = \frac{k_g}{n_g} \quad (\text{labeled-subset outcome rate}).$$

The quantity an oversight reader cares about is the full-population rate r_g , the rate that would be observed if every record carried its true label. Records without a label still have a true group membership and a true outcome; only the audit's knowledge is missing. We assume only that every labeled record is genuinely in group g and that $0 \leq \hat{r}_g \leq 1$; no assumption is made about the mechanism generating missing labels.

Two knowledge regimes are treated. In the **known-group-size regime** (Section 4.1), the audit knows N_g for each group, so m_g is computable per group. In the **pooled regime** (Section 4.5), the audit knows only the total population n , the total labeled count k , and the overall missing fraction $m = 1 - k/n$; the group membership of the unlabeled records is itself unknown. The pooled regime is the weaker-knowledge case and is the one used for the four-fifths worked example.

3.2 The worst-case missingness model

We assume nothing about which records are unlabeled or why. Formally, the audit's knowledge is a set of admissible assignments of the unlabeled records to groups and outcome configurations, and we take the worst case over all of them. This is the same posture as the Byzantine model of [3] and [4]: the untrusted component (unlabeled records; corrupt workers) is given arbitrary power, and the guarantee is whatever survives that power. The guarantee is then an interval, not a point. The worst case is tight: for any narrower interval, an admissible assignment exists that pushes the true rate outside it, so any narrower interval presupposes a missingness model that the audit must state and defend.

3.3 What the audit publishes

Under our framework, an audit with incomplete labels publishes, for each group rate:

1. the point estimate $\hat{r}_g$ on labeled records,
2. the missing-label fraction (m_g in the known-group-size regime, m in the pooled regime),
3. the worst-case interval $[r_{\min,g}, r_{\max,g}]$ derived in Section 4.1,
4. a determinacy verdict for each disparity threshold $\delta_{\min}$ of interest (Section 4.3), and
5. the minimum number of labels to recover, Δk , that would make the verdict determinate (Section 4.5).

Items 1 and 2 are already computable from any audit's own data, which is what makes the framework deployable without new measurement infrastructure; items 3 through 5 are closed-form computations on the same counts and are the additions this paper derives.

3.4 Comparison baselines and the four-fifths rule

We use a ratio threshold of 0.8 (the four-fifths threshold) as a stated example threshold; the derivations hold for any threshold $\rho_{\min}$, and the difference-form determinacy rule of Section 4.3 holds for any policy-set threshold $\delta_{\min}$. We analyze the ratio of two group rates under missingness in one group and completeness in the other, because that mixed case is the common one and it exhibits the mechanism most cleanly.

3.5 Sampling error

The worst-case interval covers missingness only. Labeled records are a finite sample, so $\hat{r}_g$ also carries binomial sampling error with standard error $SE_g = \sqrt{\hat{r}_g(1 - \hat{r}_g)/n_g}$. We combine the two sources by adding the 95% half-width $1.96 SE_g$ (with $z_{0.975} = 1.96$) to the missingness width, in Section 4.6. For a two-group disparity the two groups' sampling half-widths are added, a conservative combination; the two uncertainties are not independent in general, and we state the combination as an approximation, not an exact convolution.

4. Analysis

All derivations in this section are exact arithmetic on the definitions of Section 3; no simulation and no external data are used. Every input number is stated where it is used.

4.1 Worst-case interval for a single group's rate (known group size)

Given $\hat{r}_g = k_g/n_g$ and known N_g , the number of outcome-1 records among the labeled is fixed at k_g . Among the $N_g - n_g$ unlabeled records of group g , the count of outcome-1 records can be anything from 0 to $N_g - n_g$: no assumption about missingness rules any of these out. Therefore:

$$r_g^{\min} = \frac{k_g}{N_g} = \frac{k_g}{n_g} \cdot \frac{n_g}{N_g} = \hat{r}_g(1 - m_g),$$

$$r_g^{\max} = \frac{k_g + (N_g - n_g)}{N_g} = \hat{r}_g(1 - m_g) + m_g.$$

Every value in $[r_g^{\min}, r_g^{\max}]$ is attained by some assignment of outcomes to unlabeled records, so the interval is exact. Its width is:

$$r_g^{\max} - r_g^{\min} = m_g.$$

The width of what can be said about a group's rate equals the group's missingness rate, exactly, regardless of the rate itself. This is the first main result and the reason the missingness rate must appear in the report: it is not a data-quality footnote but the literal width of the reader's uncertainty.

4.2 Interval on a two-group disparity

Let $\Delta = r_1 - r_2$. Since $r_1 \in [\ell_1, u_1]$ and $r_2 \in [\ell_2, u_2]$ with $\ell_i = \hat{r}_i(1 - m_i)$ and $u_i = \ell_i + m_i$, and the two groups' unlabeled assignments are independent of each other (any combination is admissible):

$$\Delta^{\min} = \ell_1 - u_2 = \hat{r}_1(1 - m_1) - \hat{r}_2(1 - m_2) - m_2,$$

$$\Delta^{\max} = u_1 - \ell_2 = \hat{r}_1(1 - m_1) + m_1 - \hat{r}_2(1 - m_2).$$

The width is:

$$\Delta^{\max} - \Delta^{\min} = m_1 + m_2.$$

The width of the assumption-free interval on a disparity is the sum of the two groups' missingness rates, independent of the rates and of the reported gap. This is the paper's central result.

Worked example. Take the numbers a plausible audit might publish: group A with $N_1 = 1000$, $n_1 = 400$, $k_1 = 100$; group B with $N_2 = 1000$, $n_2 = 500$, $k_2 = 75$. Then:

$$m_1 = 1 - \frac{400}{1000} = 0.6, \quad m_2 = 1 - \frac{500}{1000} = 0.5,$$

$$\hat{r}_1 = \frac{100}{400} = 0.25, \quad \hat{r}_2 = \frac{75}{500} = 0.15, \quad \hat{g} = 0.25 - 0.15 = 0.10.$$

Bounds:

$$\ell_1 = 0.25 \times 0.4 = 0.10, \quad u_1 = 0.10 + 0.6 = 0.70,$$

$$\ell_2 = 0.15 \times 0.5 = 0.075, \quad u_2 = 0.075 + 0.5 = 0.575,$$

$$\Delta^{\min} = 0.10 - 0.575 = -0.475, \quad \Delta^{\max} = 0.70 - 0.075 = 0.625.$$

Check of the width identity: $0.625 - (-0.475) = 1.100 = m_1 + m_2 = 0.6 + 0.5$. The reported ten-point gap against group B is compatible with any true gap from a 47.5-point gap in B's favor to a 62.5-point gap in A's favor. Under assumption-free scrutiny this audit's headline disparity carries no sign information whatsoever.

4.3 The determinacy rule

A reader who wants to affirm "group 1 is selected at a higher rate than group 2" needs $\Delta^{\min} > 0$. Since $\hat{r}_i \leq 1$ implies $\hat{r}_i(1 - m_i) \geq \hat{r}_i - m_i$, and $u_2 = \hat{r}_2(1 - m_2) + m_2 \leq \hat{r}_2 + m_2$:

$$\Delta^{\min} \geq (\hat{r}_1 - m_1) - (\hat{r}_2 + m_2) = \hat{g} - (m_1 + m_2).$$

Hence:

$$\hat{g} > m_1 + m_2 \Rightarrow \Delta^{\min} > 0$$

is a sufficient condition, checkable by the reader from three disclosed numbers. The exact necessary and sufficient condition is $\hat{r}_1(1 - m_1) > \hat{r}_2(1 - m_2) + m_2$, which the reader can also check directly; the $\hat{g} > m_1 + m_2$ form is the conservative shortcut. For the Section 4.2 example, $\hat{g} = 0.10$ against $m_1 + m_2 = 1.10$: the shortcut condition fails by a factor of 11, matching the exact finding that the claim does not survive.

The same structure yields a certifiability rule for policy-set thresholds. Let the audit care whether the disparity between groups A and B exceeds a threshold $\delta_{\min}$, the smallest disparity that would trigger action, and let group B be fully labeled with known rate p_B (the mixed case; if both groups have missingness, replace p_B by its own interval and the rule holds with both widths). The disparity $D = p_A - p_B$ is bounded by $D_{\min} = p_{\min,A} - p_B$ and $D_{\max} = p_{\max,A} - p_B$. The audit can certify "no actionable disparity" only if $D_{\max} < \delta_{\min}$, and can certify "actionable disparity present" only if $D_{\min} > \delta_{\min}$. Since $p_{\max,A} - p_{\min,A} = m$, if $m < \delta_{\min}$ then the interval width is smaller than the threshold and at least one of the two certifications is possible whatever the point estimate. This condition is sufficient, not necessary: the exact condition for at least one certification is that $\delta_{\min} + p_B$ lies outside $[p_{\min,A}, p_{\max,A}]$, which can hold even when $m \geq \delta_{\min}$ — for example, $p_{\min,A} \geq \delta_{\min} + p_B$ certifies compliance at any m .

4.4 The shortcut is sufficient but not necessary

The condition $\hat{g} > m_1 + m_2$ of Section 4.3 is conservative: it can fail while the exact condition $\hat{r}_1(1 - m_1) > \hat{r}_2(1 - m_2) + m_2$ holds, because the shortcut bounds $\hat{r}_i(1 - m_i)$ by $\hat{r}_i - m_i$, and that bound is loose whenever $\hat{r}_i < 1$. A counter-example shows the gap is real, not an artifact of the algebra. Take group 1 with $\hat{r}_1 = 0.8$, $m_1 = 0.2$, and group 2 with $\hat{r}_2 = 0.2$, $m_2 = 0.5$. Then:

$$\ell_1 = 0.8 \times 0.8 = 0.64, \quad u_2 = 0.2 \times 0.5 + 0.5 = 0.6,$$

$$\Delta^{\min} = 0.64 - 0.6 = 0.04 > 0,$$

so the sign of the disparity is certifiable. But the shortcut quantity is $\hat{g} = 0.8 - 0.2 = 0.6$ against $m_1 + m_2 = 0.7$, and $0.6 < 0.7$: the shortcut fails. The exact condition, $0.64 > 0.6$, holds. The moral for the report designer is that the shortcut should be published as a fast screening test that can produce false negatives for determinacy, never as the verdict itself; the exact condition $\hat{r}_1(1 - m_1) > \hat{r}_2(1 - m_2) + m_2$ is equally checkable by the reader from the same disclosed numbers and should be the computed verdict.

4.5 Worked numerical audit in the pooled regime

The pooled regime of Section 3.1 is the weaker-knowledge case and the one an oversight reader most often faces, because many audits disclose population and labeled counts but not per-group population counts. The inputs, all stated here, are: total population $n = 1000$; total labeled records $k = 800$, so the overall missing fraction is

$$m = 1 - \frac{k}{n} = 1 - \frac{800}{1000} = 0.2;$$

a labeled-subset rate for group A of $\hat{p}_A = 0.30$ (that is, $k_A = 240$ of the 800 labeled records are outcome-1 and belong to group A); and a fully labeled comparison rate for group B of $p_B = 0.35$, known exactly because group B's labeling is complete in this scenario.

Rate interval. By the Section 4.1 construction applied with the pooled missing fraction,

$$p_{\min,A} = \hat{p}_A(1 - m) = 0.30 \times 0.8 = 0.24,$$

$$p_{\max,A} = \hat{p}_A(1 - m) + m = 0.24 + 0.2 = 0.44.$$

The interval has width $m = 0.2$, exactly as derived.

Four-fifths ratio interval. The ratio $R = p_A/p_B$ is monotone in p_A because $p_B > 0$ is fixed, so the interval endpoints are the ratios of the rate endpoints:

$$R_{\min} = \frac{0.24}{0.35} = 0.68571 \dots \approx 0.686, \quad R_{\max} = \frac{0.44}{0.35} = 1.25714 \dots \approx 1.257.$$

The point estimate is $\hat{R} = 0.30/0.35 = 0.85714 \dots \approx 0.857$, which sits above the four-fifths threshold 0.8 and would suggest compliance if read alone. But $0.8 \in [0.686, 1.257]$: the threshold lies inside the assumption-free interval, so the audit cannot certify either compliance or adverse impact. The verdict is **indeterminate**, and the point estimate 0.857 carries no certifiable content.

Label-recovery target. How many labels must be recovered before a determinate verdict becomes possible? Two answers exist, and the difference between them is the paper's most important practical caveat.

Worst-case answer: no finite number. Under the fully adversarial model, the outcomes of unlabeled records are unknown, and recovering Δk labels converts those particular unknowns into observations — but before recovery, the audit must reason about what recovery might reveal. If the recovered labels could all show outcome 1, the worst-case upper bound stays at $p_{\max,A} = 0.44$; if they could all show outcome 0, the worst-case lower bound stays at $p_{\min,A} = k_A/n = 240/1000 = 0.24$. Certifying compliance requires $p_{\min,A} \geq 0.8 \times 0.35 = 0.28$, and $0.24 < 0.28$ regardless of Δk ; certifying adverse impact requires $p_{\max,A} < 0.28$, and $0.44 > 0.28$ regardless of Δk . Under the assumption-free model, therefore, no recovery short of the full 200 unlabeled labels guarantees a determinate verdict, because the guarantee must hold for every outcome pattern the partial recovery could reveal; full recovery of all 200 labels yields exact knowledge of p_A and thus a determinate verdict in every outcome pattern.

$$\Delta k \text{ such that } \hat{p}_A(1 - m') \geq 0.8 p_B \quad \text{with } m' = 1 - \frac{k + \Delta k}{n}.$$

The computation requires one stated assumption: that the recovered labels have the same outcome rate as the existing labeled subset, so $\hat{p}_A = 0.30$ is unchanged by recovery. This is an *expected-case* assumption, weaker than the adversarial model and stronger than nothing; we flag it as such and return to it in Section 6. Under it, certifying compliance requires:

$$0.30(1 - m') \geq 0.28 \quad \Rightarrow \quad 1 - m' \geq \frac{0.28}{0.30} = 0.9\bar{3} \quad \Rightarrow \quad \frac{k + \Delta k}{n} \geq 0.9\bar{3}.$$

With $n = 1000$ and $k = 800$:

$$k + \Delta k \geq 933.\bar{3} \quad \Rightarrow \quad \Delta k \geq 133.\bar{3} \quad \Rightarrow \quad \Delta k = 134 \text{ labels (rounded up to an integer).}$$

Certifying adverse impact, by contrast, requires $p_{\max,A} = \hat{p}_A(1 - m') + m' = 0.30 + 0.70 m' < 0.28$, which is impossible for any $m' \geq 0$ because $0.30 > 0.28$ already at $m' = 0$. So under the expected-case assumption the only reachable determinate verdict is compliance, and it costs 134 recovered labels; under the fully adversarial assumption no finite recovery guarantees any verdict. The number 134 is therefore a *conditional* remediation target, valid exactly while the recovered labels behave like the labeled subset, and the audit report should state that condition next to the number.

4.6 Combining missingness with finite-sample error

The worst-case interval of Section 4.1 covers missingness only; the labeled-subset rate $\hat{r}_g$ additionally carries binomial sampling error with standard error $\text{SE}_g = \sqrt{\hat{r}_g(1 - \hat{r}_g)/n_g}$ (Section 3.5). We quantify the two sources separately and then combine them additively, which is conservative because the two uncertainties are not independent in general.

Near-complete labeling. Take $m_1 = m_2 = 0.02$ (so $n_g = 980$ out of $N_g = 1000$ for each group) and $\hat{r}_1 = \hat{r}_2 = 0.25$. The missingness width of the two-group disparity interval is, by Section 4.2:

$$m_1 + m_2 = 0.04.$$

The per-group sampling half-width at $z_{0.975} = 1.96$ is:

$$1.96 \text{ SE} = 1.96 \sqrt{\frac{0.25 \times 0.75}{980}} = 1.96 \sqrt{0.00019132} = 1.96 \times 0.013832 = 0.027111.$$

Adding the two groups' half-widths conservatively, the sampling contribution to the disparity interval is:

$$2 \times 0.027111 = 0.054222 \quad (\text{computed without intermediate rounding: } 0.05422177).$$

At near-complete labeling, sampling error (0.054) is therefore the larger of the two uncertainties, exceeding the missingness width (0.04) by a factor of $0.05422177/0.04 = 1.356$. An audit at 98% labeling that reports only the missingness interval under-states its total uncertainty; it must also report the sampling half-width.

Moderate missingness. Take the Section 4.5 inputs: $m = 0.2$, $k = 800$ labeled records, $\hat{p}_A = 0.30$. The missingness width is 0.2. The sampling half-width is:

$$1.96 \sqrt{\frac{0.30 \times 0.70}{800}} = 1.96 \sqrt{0.00026250} = 1.96 \times 0.016202 = 0.031756.$$

The ratio of missingness width to sampling half-width is:

$$\frac{0.2}{0.031756} = 6.30.$$

At $m = 0.2$, missingness contributes about 6.3 times more uncertainty than sampling — of order 7 — so the worst-case interval dominates and the sampling correction is a second-order refinement. The crossover between the two regimes is sharp: because missingness width is linear in m while sampling half-width scales as $1/\sqrt{n_g}$, an audit designer can read off from these two computations which source binds at their own labeling level. The combined half-width at $m = 0.2$ is $0.2 + 0.031756 = 0.231756 \approx 0.232$ per rate, and at $m = 0.02$ it is $0.04 + 0.054222 = 0.094222$ (computed without intermediate rounding: 0.09422177) per disparity; both are additive approximations, stated as such.

5. Results

Every number in this section is computed in Section 4 from inputs stated there; no simulation, no external data, and no empirical measurement is reported. The results are of three kinds: closed-form results, worked numerical results, and the disclosure checklist that packages them.

Closed-form results.

1. The exact worst-case interval on a group's full-population rate is $[\hat{r}_g(1 - m_g), \hat{r}_g(1 - m_g) + m_g]$, with width exactly m_g (Section 4.1).
2. The width of the assumption-free interval on a two-group disparity is $m_1 + m_2$, independent of the rates and of the reported gap (Section 4.2).
3. The determinacy rule: certification of a threshold $\delta_{\min}$ is impossible exactly when $\delta_{\min} + p_B$ lies inside $[p_{\min,A}, p_{\max,A}]$, and $m < \delta_{\min}$ is sufficient for some certification whatever the point estimate; a reported positive gap survives the conservative screening test if $\hat{g} > m_1 + m_2$; the exact necessary-and-sufficient condition is $\hat{r}_1(1 - m_1) > \hat{r}_2(1 - m_2) + m_2$ (Sections 4.3 and 4.4).

Worked numerical results.

1. For the Section 4.2 example ($m_1 = 0.6$, $m_2 = 0.5$, reported gap 0.10): disparity interval $[-0.475, 0.625]$, width 1.100; the sign of the gap is not certifiable.
2. For the Section 4.4 counter-example ($\hat{r}_1 = 0.8$, $m_1 = 0.2$, $\hat{r}_2 = 0.2$, $m_2 = 0.5$): $\Delta^{\min} = 0.04 > 0$ while the shortcut $\hat{g} > m_1 + m_2$ fails ($0.6 < 0.7$); the shortcut produces a false negative and must not be used as the verdict.
3. For the Section 4.5 pooled audit ($n = 1000$, $k = 800$, $\hat{p}_A = 0.30$, $p_B = 0.35$): rate interval $[0.24, 0.44]$; four-fifths ratio interval $[0.686, 1.257]$; point estimate 0.857; verdict indeterminate because 0.8 lies inside the interval; conditional label-recovery target 134 labels under the expected-case assumption, and no guarantee from any recovery short of the full 200 unlabeled labels under the adversarial assumption.
4. Uncertainty decomposition (Section 4.6): at $m_1 = m_2 = 0.02$, sampling contributes 0.05422177 and missingness 0.04 to the disparity interval; at $m = 0.2$, missingness contributes 0.2 against a sampling half-width of 0.0318, a factor of 6.3.

The disclosure checklist. An audit with incomplete protected labels must publish, for each reported rate and each reported disparity:

1. The point estimate on labeled records ($\hat{r}_g$, or $\hat{g}$ for a disparity), with the labeled count n_g it was computed on.
2. The missing-label fraction (m_g per group, or m overall in the pooled regime), stated as a number, not a footnote.
3. The worst-case interval $[\hat{r}_g(1 - m_g), \hat{r}_g(1 - m_g) + m_g]$ (or the disparity interval with width $m_1 + m_2$).
4. The determinacy verdict for each policy threshold of interest ($\delta_{\min}$ in difference form, $\rho_{\min} = 0.8$ in ratio form), computed with the exact condition, not the shortcut.
5. The conditional label-recovery target Δk , together with the assumption under which it holds (expected-case versus adversarial).
6. The sampling half-width 1.96SE_g and a statement of whether missingness or sampling dominates at the audit's labeling level.

Items 1 and 2 are computable from any audit's own counts with no new measurement; items 3 through 6 are closed-form arithmetic on the same counts. The checklist therefore imposes no data-collection burden beyond what the audit already has, and its cost is the disclosure of uncertainty the audit may prefer not to state.

6. Discussion

The framework's strength is that its intervals are exact and its inputs are two numbers every audit already has; its weakness is that exactness is bought with width, and the interpretation of the width depends on assumptions the framework itself does not supply. This section states the limitations, the failure modes, and what would falsify the results.

Width as a feature and as a cost. The interval $[\hat{r}_g(1 - m_g), \hat{r}_g(1 - m_g) + m_g]$ is the narrowest interval defensible without a missingness model, but at $m = 0.5$ it spans half the unit interval and at $m = 0.6$ (Section 4.2) the disparity interval spans 1.1, wider than the entire admissible range of a single rate. An oversight reader handed only these intervals at high missingness learns almost nothing except that the audit cannot certify anything — which is itself the correct answer, but a reader expecting a verdict may read the width as methodological failure rather than as an honest measurement of the audit's evidence. The framework mitigates this by pairing the interval with the recovery target, which converts an uninformative width into an actionable number.

The expected-case/adversarial gap. The label-recovery target of 134 holds only while recovered labels behave like the labeled subset. If missingness is informative — if, for example, the people who declined to report their group are systematically more likely to have outcome 1 — then recovery shifts $\hat{p}_A$ away from 0.30 and the target moves. The adversarial answer (no finite target) is the honest bound, but it is also the least actionable. The gap between the two answers is not a defect of the arithmetic; it is the quantitative shadow of the missingness mechanism, and the audit should report both numbers so the reader sees how much hinges on the mechanism.

Failure modes of deployment. Three are foreseeable. First, an audit team may publish the checklist items but bury the interval in an appendix, reproducing the original problem one layer down; the checklist specifies placement next to the rate for this reason, but placement is a norm, not a theorem. Second, the shortcut $\hat{g} > m_1 + m_2$ may be implemented instead of the exact condition, producing false negatives for determinacy as in Section 4.4 and wrongly halting deployments that the exact condition would certify. Third, the additive combination of sampling and missingness uncertainty in Section 4.6 is conservative and approximate; a reader who treats 0.09422177 as an exact confidence statement over-reads it.

What would falsify the main result. The central claim is that the assumption-free disparity interval has width exactly $m_1 + m_2$ and that no narrower interval is valid without a missingness assumption. A single counter-example would falsify it: any assignment of unlabeled records, admissible under the framework's own definition of admissibility (Section 3.2), whose true disparity falls outside $[\hat{r}_1(1 - m_1) - \hat{r}_2(1 - m_2) - m_2, \hat{r}_1(1 - m_1) + m_1 - \hat{r}_2(1 - m_2)]$. The derivation of Section 4.2 shows such an assignment cannot exist because the interval endpoints are attained by extremal assignments, but the claim is empirical in the sense that it rests on the admissibility definition; a reviewer who argues that some assignments are not admissible — for instance, that group sizes N_g constrain the joint assignment across groups — would shrink the intervals, and the framework's exactness claim would then be wrong for that admissibility set. The width identity $0.625 - (-0.475) = 1.100 = 0.6 + 0.5$ is checkable by hand and is the fastest test of the derivation.

Limitations list.

- *No missingness model.* The result holds under any mechanism, which is why the intervals are wide; sensitivity is linear — halving m halves the width, exactly.

- *Binary outcomes.* The rate r_g is defined for $Y \in \{0, 1\}$; extensions to multi-category or continuous outcomes need rederivation, and the ratio form requires $p_B > 0$.
- *Two groups.* The derivations cover $G = 2$; for more groups the interval applies pairwise, and a joint statement over all G rates (for example, that every rate lies in a band) is a different, harder object not derived here.
- *Known denominators.* The known-group-size regime requires N_g ; the pooled regime requires only n and k , but its per-group interpretation assumes the labeled subset's group composition is representative of nothing in particular — the interval is on the pooled rate, and applying it to "group A" as in Section 4.5 assumes the labeled group-A count $k_A = 240$ is itself known, which is a middle regime between the two.
- *Additive uncertainty combination.* The Section 4.6 combination adds half-widths; the true joint distribution of sampling and missingness error is not derived, so the combined number is an upper-bound approximation.
- *Threshold provenance.* The four-fifths threshold 0.8 is used as a stated example; the framework does not derive which threshold a regulator should use, and the certifiability verdicts change with $\delta_{\min}$ linearly through the rule $m < \delta_{\min}$.
- *No behavioral model of the auditor.* The checklist assumes the audit publishes what it should; nothing in the arithmetic enforces disclosure.

Open questions. How the intervals compose across many audited subgroups (the interaction with the gerrymandering concern of [6]) is open; whether partial missingness models — for example, a stated bound on the outcome rate among unlabeled records — are a useful middle ground between the adversarial and expected-case answers is open; and how oversight readers actually weigh a wide interval against a point estimate is an empirical question this paper does not address.

7. Practical Implications and Engineering Pathways

(a) Who would use the result and what changes. The immediate users are fairness-audit teams inside organizations that deploy machine-learning systems in hiring, lending, education, and public services, and the oversight readers of their reports: regulators, courts, internal risk committees, and affected communities. If the result holds, three decisions change. The audit team sets a *minimum labeled fraction before deployment* — the determinacy rule $m < \delta_{\min}$ turns label completeness from a data-quality aspiration into a design parameter with a numeric target tied to the policy threshold the audit will be judged against. The oversight reader's accept/remediate/halt decision acquires a computed input: the interval and the determinacy verdict replace the unquantified point estimate. And the remediation negotiation acquires a price tag: the recovery target $\Delta k = 134$ (in the worked case) converts "collect more demographic data" from an open-ended demand into a bounded task with a stated validity condition.

(b) Current readiness. This work is purely analytical and has not been peer reviewed; against the NASA Technology Readiness Level definitions it sits at TRL 1–2: the basic principles (worst-case bounds over admissible assignments) are stated and the derivations are exact arithmetic, but the framework has not been applied to a real audit, validated against operational data, or reviewed. It does not exceed TRL 3, and no claim of operational or manufacturing readiness is made — there is nothing to manufacture; the deliverable is a reporting convention plus closed-form formulas implementable in a few lines of any auditing codebase.

(c) The smallest next test. The minimal de-risking exercise is a retrospective re-analysis of one existing audit report that discloses its labeled and total counts: recompute the published rates as intervals using Sections 4.1 and 4.2, recompute the determinacy verdicts, and check whether any published conclusion flips from determinate to indeterminate. The outcome that refutes the framework's practical value is not that some verdicts flip — they will, wherever m is comparable to the threshold — but that *no* verdict flips in a sample of real audits, which would

indicate that operational missingness is far below the levels where the rule binds and the disclosure is moot. A second, equally small test is the hand-check of the width identity in Section 4.2 by an independent reviewer, which tests exactness rather than usefulness.

(d) Binding constraints. The first binding constraint is *data governance, not computation*: recovering 134 labels means collecting 134 protected attributes, which is precisely the activity whose risks [2] characterizes, so the recovery target may be unreachable in jurisdictions or contexts where collection is restricted — in which case the audit's honest output is the indeterminate verdict, and the framework's contribution is making that fact explicit rather than fixable. The second is *group granularity*: per [6], finer subgroup structure multiplies published rates and shrinks each subgroup, raising each effective m_g exactly where audit risk concentrates; the intervals widen accordingly, and an audit designer must choose between coarse groups with small m_g and fine groups with large ones. The third is *implementation discipline*: the exact condition, not the shortcut, must be coded, per the Section 4.4 counter-example. Compute cost is negligible — the formulas are closed-form arithmetic on counts the audit already holds. Whether oversight bodies accept interval-based reporting in place of point estimates is a regulatory question and an open one.

(e) Order-of-magnitude bounds (projections). Labeled as projections with stated assumptions: if an audit's labeling effort costs roughly the same per recovered label regardless of count, the marginal cost of moving from $m = 0.2$ to determinacy scales with Δk , which Section 4.5 computes as 134 labels for the worked case; the framework itself derives no cost per label, so no monetary figure is stated here. On the performance side, the Section 4.6 decomposition projects that audits above roughly $m \approx 0.1$ (where the missingness width 0.1 exceeds typical sampling half-widths of 0.03 at $n_g \approx 800$) are missingness-dominated, so marginal investment in statistical precision buys little below that line and marginal investment in label recovery buys proportionally. Both projections assume the audit's rates are near the worked values ($\hat{r} \approx 0.25$ to 0.30) and n in the hundreds to low thousands.

(f) Risks and misuse paths. The framework can be misused in two directions. An audit team can publish the wide interval to *launder indeterminacy*: reporting "we cannot certify anything" as a defense, when the correct reading is that the audit is incomplete and the recovery target names the fix. Conversely, an adversarial deployer can shape which records go unlabeled — the framework assumes nothing about the mechanism precisely because missingness can be engineered — and the worst-case interval correctly refuses to certify, but a naive reader may mistake the refusal for a clean bill of health. The checklist's item 5 (the recovery target and its validity condition) exists to close both paths by making the path to determinacy, and its assumptions, explicit.

8. Conclusion

When protected labels are incomplete, the rate a fairness audit publishes is silently a rate on the labeled subset, and the reader's uncertainty about the full-population rate is exactly the missing-label fraction. This paper made that uncertainty computable: the assumption-free interval on a group's rate has width m_g ; the interval on a two-group disparity has width $m_1 + m_2$, independent of the rates; certification of a threshold is impossible exactly when the shifted threshold lies inside the interval, with $m < \delta_{\min}$ sufficient for some certification whatever the point estimate; and the shortcut screening test for a positive gap is sufficient but not necessary, so the exact condition must be the published verdict. In the worked pooled audit, a point estimate of 0.857 against the four-fifths threshold resolves to an interval $[0.686, 1.257]$ and an indeterminate verdict, with a conditional remediation price of 134 labels that holds only while recovered labels behave like the labeled subset — and no guarantee from any partial recovery under the adversarial model, while full recovery of the 200 unlabeled labels always yields a determinate verdict. The uncertainty decomposition shows a crossover: below roughly $m \approx 0.1$ sampling error dominates and the audit should report both half-widths; above it, missingness dominates by a factor that grows linearly. The

deliverable is a six-item disclosure checklist whose every entry is closed-form arithmetic on counts the audit already holds. The framework's stance is conservative by construction: it certifies nothing it cannot prove under the worst admissible missingness, and it names, in every indeterminate case, exactly what evidence would change the verdict.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.16.0-revise-retry). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; statements about a cited work are restricted to its retrieved title and abstract.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 10 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond its abstract; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] What Must a Fairness Audit Report When Demographic Data Is Incomplete?. arXiv:2506.23033v5. <https://arxiv.org/abs/2506.23033v5> [2] Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. arXiv:2205.01038v2. <https://arxiv.org/abs/2205.01038v2> [3] Data Encoding for Byzantine-Resilient Distributed Optimization. arXiv:1907.02664v2. <https://arxiv.org/abs/1907.02664v2> [4] Byzantine-Resilient SGD in High Dimensions on Heterogeneous Data. arXiv:2005.07866v1. <https://arxiv.org/abs/2005.07866v1> [5] Constraints on dark energy from H II starburst galaxy apparent magnitude versus redshift data. arXiv:1110.5626v1. <https://arxiv.org/abs/1110.5626v1> [6] Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. arXiv:1711.05144v5. <https://arxiv.org/abs/1711.05144v5> [7] TerraGen: A Unified Multi-Task Layout Generation Framework for Remote Sensing Data Augmentation. arXiv:2510.21391v1. <https://arxiv.org/abs/2510.21391v1> [8] AeroGen: Enhancing Remote Sensing Object Detection with Diffusion-Driven Data Generation. arXiv:2411.15497v3. <https://arxiv.org/abs/2411.15497v3> [9] QNFO: A Falsifiability-First Audit Program for the Photon-Primacy, Helical-Electron, and Adelic-Mass Claim Cluster [10] QNFO: Knowing What We Do Not Know: Ignorance Auditing, AI-Generation Detection, and the Epistemic Lessons of an AI-Assisted Research Pipeline [11] QNFO: The Universal Ignorance Audit: A Fifteen-Question Method for Systematic Inquiry into the Structure of Not-Knowing