SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

VERITAS — Auditing the Ground Truth, Not Just the Model

AKASH STEPHEN, Dr. Somervell Memorial CSI Medical College & Hospital

Briefly describe what attendees will see during your demo.
I walk through two things. First, an interactive "artifact floor" explorer: I move sliders for reader accuracy and subgroup prevalence, and the audience watches a fairness gap appear for a model that is identical in both groups, then watch it shrink as I add readers to the panel. I show the real number from LIDC-IDRI (four radiologists, 1,016 CT scans): a 0.239 apparent sensitivity gap where the true gap is zero. Second, the VERITAS harness itself: I ingest a small study, run a blinded multi-reader adjudication, launch the metrics job, and open the STARD-style report and results view, showing per-subgroup inter-reader agreement (Krippendorff's α) and the tamper-evident provenance ledger that records how the reference standard was built, with one-click chain verification. The through-line, in one sentence: every other tool audits the model; this one audits the ground truth the model is graded against. (Live explorer: artifact-floor.pages.dev)

Problem & Motivation: What clinical or operational problem does your MVP address?
Medical-AI validation grades a model against a reference standard, the labels treated as ground truth. But that ground truth is built by humans who disagree, then collapsed into one consensus label, and it is imperfect. Two consequences follow, and current tools are blind to both. First: when you compare a model across demographic subgroups (a fairness audit), an imperfect reference interacting with differing disease prevalence produces an apparent gap even for a model that is exactly equally good in both groups. Prevalence routinely differs twofold or more across age, sex, and ancestry so the field risks reporting "bias" that belongs to the labels, not the model. The effect can even reverse sign: a model genuinely better for one group can be measured as worse. Second: almost every validation pipeline discards the individual reader reads and keeps only the collapsed consensus label. That makes it impossible, after the fact, to estimate how good the reference was, let alone correct for it or audit its construction. I kept hitting this in published fairness claims: a subgroup gap gets pinned on the model, but nobody reports how noisy the labels were or how prevalence differed across the groups. Fairness toolkits (Fairlearn, AIF360, Aequitas) have no concept of an imperfect reference at all. The one thing that decides whether a fairness finding is real — the quality of the ground truth — is the thing nobody measures, or keeps the data to measure.

What You Built: Describe your MVP, tool, or workflow in plain language
VERITAS is an open validation harness that treats the reference standard as a measured instrument, not as truth. Inputs: a CSV of model outputs (a score/prediction per case), the individual reads from multiple human readers, and optional subgroup attributes (e.g., sex, site). It runs a blinded multi-reader workflow — readers can't see model outputs or each other's reads — and an adjudication step to resolve disagreements. Outputs: standard classification metrics (sensitivity, specificity, PPV/NPV, AUC) with bootstrap confidence intervals; subgroup fairness gaps with Benjamini–Hochberg multiplicity correction; and — the distinctive part — a quality report on the reference standard itself: how much the readers agreed (Krippendorff's α, overall and per subgroup), how many cases were excluded and why, and how the ground truth was constructed. Every read,

disagreement, and adjudication decision is written into a tamper-evident, hash-chained provenance ledger, so you can later prove what the labels were and how they were made. It exports a STARD-style report. A companion piece — the artifact floor — quantifies how large a fairness gap an imperfect reference can fabricate from prevalence alone, so a reported gap can be checked against the floor below which it carries no information. In plain terms: you feed it model predictions and messy human labels; it tells you how good the model is, how fair it is, how trustworthy the labels underneath were, and it keeps an auditable record of the whole thing.

Tech Stack & Development Approach: What tools, frameworks, or methods did you use to build it?

Backend: Python — FastAPI (REST), PostgreSQL (data + provenance ledger), Redis + RQ (long-running metric/bootstrap jobs), SQLAlchemy/Alembic, NumPy/scikit-learn (metrics). The ledger is a per-study SHA-256 hash chain with database triggers that reject updates/deletes to ledger rows, plus write-once analysis payloads. Frontend: Next.js 15 (App Router) / React / TypeScript, API client generated from an OpenAPI schema. Runs locally via `docker compose` (an on-prem profile, no external calls). ~264 automated tests. The artifact-floor analysis is separate, pure Python (NumPy/SciPy): closed-form derivations, a latent-class EM fit for reader accuracy, and a goodness-of-fit test, checked by a 400,000-sample Monte Carlo. The interactive demo is one self-contained HTML/JS file on Cloudflare Pages. Development approach, honestly: this was built with heavy AI assistance (Anthropic's Claude, via Claude Code) under my direction — I'm a final-year medical student, not a trained software engineer. I set the architecture and its invariants (provenance-by-construction; blinding-as-authorization; undefined metrics report as *n/e*, never 0), specified the statistics, and verified the results; the AI did much of the implementation and drafting. The math is independently reproducible from open, unit-tested code — that's how I keep "AI-assisted" from meaning "unverified." A constant discipline was staying honest about what's genuinely built (agreement, ledger, metrics) versus still research-only (the floor).

Validation & Evidence: Have you done any informal testing or validation? If so, what did you find?

Two kinds. The finding - The artifact floor is derived in closed form and confirmed by a 400,000-sample Monte Carlo that re-derives it from simulated readers, so an algebra error can't pass silently. On LIDC-IDRI (four thoracic radiologists, 1,016 CT scans — a genuinely public multi-reader dataset) I fit reader accuracy by latent-class analysis (mean sensitivity 0.948, specificity 0.787) and get an artifact floor of 0.239 for a 30%-vs-10% prevalence contrast — larger than most published imaging-fairness gaps, for a model identical across groups. On the same data, the conditional-independence assumption every consensus label silently relies on is rejected (G^2=44.1, 6 df, $p \approx 7\times10^{-8}$). All of it reproduces from open, unit-tested code (github.com/akashstephen/artifact-floor). The tool - ~264 automated tests, including release-gate tests that enforce the core guarantees — the ledger's immutability and hash-chain verification, and a blinding suite that sweeps every API response to assert no model output leaks to a blinded reader. A 32-check end-to-end smoke harness runs the full pipeline on a seeded study. What I have NOT done: validated the tool in a real clinical or research study, or run a published model through it end-to-end. No institution has used it. So the evidence is "the math is right and the software works as specified" — not "this changed a real validation outcome," yet.

Feedback You're Seeking: What specific feedback, help, or collaboration are you hoping to get from the CAIMI26 community?

Three things. 1. A clinical/research champion with real multi-reader data - The whole approach needs the individual reader reads that most pipelines discard. I'd love to run VERITAS on a real reader study — even a small one — and show what happens to a reported subgroup gap once the reference standard's own

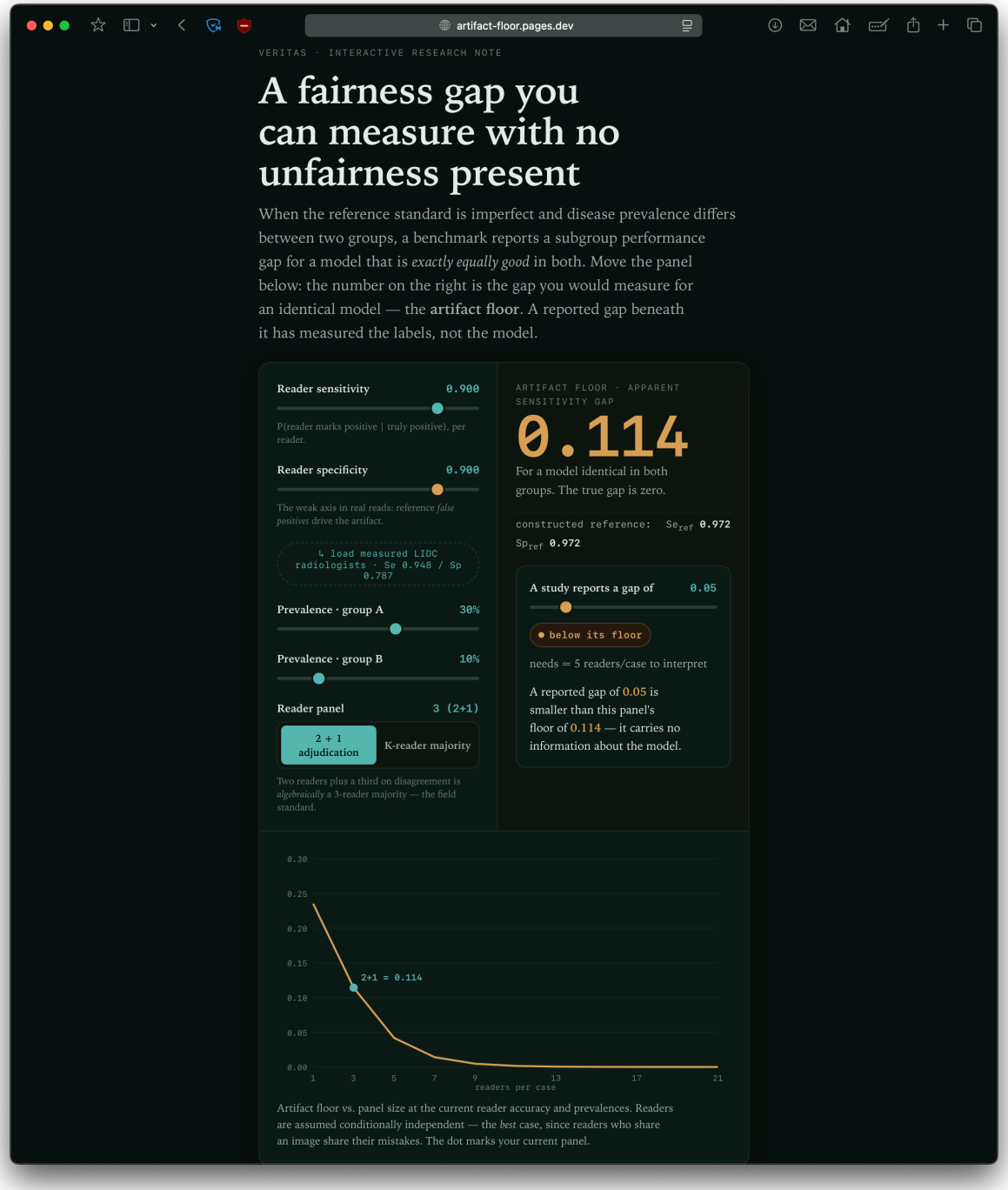
accuracy is measured. If your group keeps per-reader labels, I want to talk. 2. What belongs on the report face : For a regulatory or peer-review reviewer, which reference-standard-quality signals actually change your trust in a fairness number — agreement (α)? the artifact floor? the adjudication rate per subgroup? I don't want to build signals nobody reads. 3. How to present a bound you know is conservative - The floor currently assumes readers err independently; in imaging they don't (shared image-driven error), which only makes the true floor worse — so every number I show is a lower bound. How should a tool communicate "this is conservative, reality is worse" without either alarming people or being ignored? Bigger picture: I'm a medical student who built this partly to find out whether it's genuinely useful or just clever. An honest "this wouldn't survive contact with our workflow, because X" is exactly as valuable to me as encouragement.

Known Limitations & Honest Failures: What isn't working yet, or what have you already tried that failed?

The biggest limitation is intellectual, and I lead with it: the artifact-floor math assumes readers make independent errors. They don't — radiologists reading the same ambiguous image tend to err the same way. Correlated error makes the constructed reference worse, so every floor I report is a lower bound; the real problem is bigger than my numbers. Quantifying correlated error from real data is genuinely unsolved and beyond what I've done. I state this on the demo itself rather than bury it. Built vs. claimed: per-subgroup inter-reader agreement is integrated end-to-end in the harness today. The artifact floor is not yet wired into the live tool — it's validated research plus an interactive demo, but the results view doesn't yet annotate each subgroup gap with its floor. That's the next build, and it's non-trivial: estimating reader accuracy honestly needs ≥3–4 readers per case, which my synthetic demo study doesn't have. A retracted mistake: an earlier plan made MRMC statistics the headline; I dropped it after realizing it answers a different question (reader performance, not reference-standard error) — a clean example of pattern-matching a method before stating the estimand. Not addressed at all: images/DICOM (tabular only), correlated error, and any real deployment. And the ledger is honestly *tamper-evident*, not *tamper-proof* — a database admin who rewrites rows could recompute a self-consistent chain. I don't claim otherwise.

Potential for Impact: If this MVP were fully developed and deployed, who would benefit and how?

Patients are the ultimate stakeholders - Fairness audits exist to catch models that underperform for under-served groups. If some reported "gaps" are artifacts of noisy labels and differing prevalence — and some real gaps are masked by the same mechanism (the effect can reverse sign) — then acting on the raw numbers misdirects scarce mitigation effort. Getting the reference standard right is upstream of getting fairness right. Radiologists and reader-study teams get credit for a truth most pipelines erase: that they disagree, and that the disagreement is information. VERITAS keeps and reports it instead of averaging it away. AI developers and validators get a way to separate "my model is biased" from "my labels are noisy and my subgroups differ in prevalence" — before publishing a fairness claim or submitting to a regulator. Regulators and journals - Subgroup validation and traceable data lineage are now required (EU AI Act Article 12; FDA AI-device guidance; TRIPOD+AI). A tool that records how the ground truth was constructed, under a tamper-evident ledger, and flags fairness claims the labels can't support, aims squarely at that bar. Trainees and small groups who can't afford commercial validation infrastructure get an open, on-prem option. Honest scope: this is upstream plumbing — it makes fairness and validation claims trustworthy; it doesn't invent new fairness methods. That layer underneath is exactly what's missing today.



GitHub Repo:

- https://github.com/akashstephen/artifact-floor
- https://artifact-floor.pages.dev/