Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SynthGuard-ReleaseBench: Locked-Audit Evidence for Synthetic Tabular Data Releases

Jeffery Opoku

jeffery.opoku01@utrgv.edu

University of Texas Rio grande valley

David

Florida International University

Research Article

Keywords: synthetic data, data release, audit design, differential privacy, membership inference, reproducibility, statistical benchmarking

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10705803/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

SynthGuard-ReleaseBench: Locked-Audit Evidence for Synthetic Tabular Data Releases

Jeffery Opoku David Banahene

School of Mathematical and Statistical Science, University of Texas Rio Grande Valley
opokujeffery5@gmail.com

Robert Stempel College of Public Health and Social Work, Florida International University
abanahene54@gmail.com

Preprint, July 2026

Abstract

Synthetic tabular data are often judged by realism, privacy, or downstream-task scores. Those scores do not answer whether a proposed release is supported for a named use, population, and threat model. We introduce SynthGuard-ReleaseBench, an audit framework that locks the use, candidate panel, tolerances, and audit schedule before evaluation. It compares real-trained and synthetic-trained workflows on protected data, gives simultaneous finite-sample bounds for bounded loss gaps, requires controls, and keeps utility, empirical privacy risk, mechanism claims, and human release authority separate.

Across four American Community Survey studies, five non-ACS records, two chronological diagnostics, and a sealed prototype, the benchmark retains favorable, unfavorable, and excluded outcomes. Transparent baselines pass some locked audits; compact learned models fail under the declared budgets; a health-table case is excluded because its negative control passes. A post-audit scaling arm, repeated across three generation seeds, shows the same locked criterion admitting those learned models once they are fit on enough data while still rejecting a dependence-destroying control at every size, so the criterion discriminates rather than merely rejects; the same repetition withdraws a finer single-seed ordering.

The theory adds a pre-audit sample-size rule, variance-adaptive and anytime-valid certificates that tighten the bound two to ten times on the same locked evidence, a temporal certificate for time-ordered audits, and two lower bounds: ordinary bounded queries reconstruct a protected audit once the query budget reaches its size, and the panel-size correction is necessary rather than conservative. The contribution is a reproducible workflow for use-specific release evidence, not a claim that any generator is private, safe, or deployment-ready.

Keywords: synthetic data; data release; audit design; differential privacy; membership inference; reproducibility; statistical benchmarking

1 Introduction

Synthetic tabular data are now used for prototyping, testing, model development, data sharing, and controlled analysis. Modern generators can be sophisticated, including conditional adversarial models [33], classical partially synthetic-data methods [14], and differentially private mechanisms [4, 8]. Yet a release decision cannot be reduced to the question "which generator gets the best average score?" The established distinction between general and analysis-specific utility already

points in this direction [31]. A data user may need a reliable calibration curve, a subgroup-specific model assessment, or a fixed descriptive estimate. A data steward may also need evidence about a concrete attack scenario. These are different questions and they require different measurements.

This paper introduces SynthGuard-ReleaseBench, an audit protocol and open implementation for evidence about a *named synthetic-data use*. The protocol does not propose a new universal synthetic-data score or a new generator. Its contribution is to connect four pieces that are usually reported separately: (i) a declared analytical use and tolerance; (ii) a protected, pre-specified audit; (iii) simultaneous uncertainty accounting across a panel of checks; and (iv) a distinct, threat-model-specific privacy-risk report. A separate release gate then asks whether the audit had the required integrity conditions. A passing metric grid is deliberately insufficient for an automatic release decision.

The distinction matters. An observed low membership-inference score does not establish general privacy: attacks depend on the adversary, information, and access pattern [30, 32]. Conversely, a differential privacy claim is a formal property of a specified mechanism and adjacency relation, not a statement that every intended analysis has small error [4]. Existing synthetic-data tools and benchmarks provide valuable generation and evaluation infrastructure, including NIST's SDNist tools and the SDGym benchmark environment [10, 16]. SynthGuard complements rather than replaces them by requiring a locked, use-level evidence packet.

The contributions are as follows.

1. We formalize a use-specific audit target and simultaneous finite-sample certificates for a fixed panel of bounded, paired loss gaps, together with a sufficient audit-size rule for a declared certification margin. We add variance-adaptive and anytime-valid versions of the certificate, which tighten the reported half-width by factors of two to ten on our own locked evidence at no cost in validity, and a buffered-block certificate that extends the framework to time-ordered audits under a declared dependence horizon. Two lower bounds delimit the framework: ordinary bounded queries reconstruct a protected audit exactly once the query budget reaches the audit size, so unrestricted reuse fails operationally and not merely in principle; and the panel-size correction is necessary, since a width that ignores it loses simultaneous coverage as the panel grows.

2. We specify a benchmark protocol with separate development, selection, and audit roles; negative and positive controls; pre-registered candidate and analysis panels; full-grid reporting; and an explicit distinction between public measurement and sealed local auditing.

3. We provide four completed public ACS studies, five non-ACS row-level technical records, two deliberately non-certifying chronological tracks, and a sealed-audit prototype. The full inventory and claim boundaries are moved to the supplement.

4. We provide an installable command-line package, hash-checked evidence registries, a container, continuous integration, a documented comparison with established evaluation packages, and an outcome-free industrial challenge for an unrelated external team.

Figure 1 summarizes the workflow. The protected audit produces statistical and threat-model-specific evidence; it does not inherit the organization's authority to release data.

Scope. The public studies in this paper use ACS Public Use Microdata Sample files. The Census Bureau documents their variable coding, sample design, and weights [1]. Our public results use a row-level independent-and-identically-distributed (iid) working assumption to demonstrate the

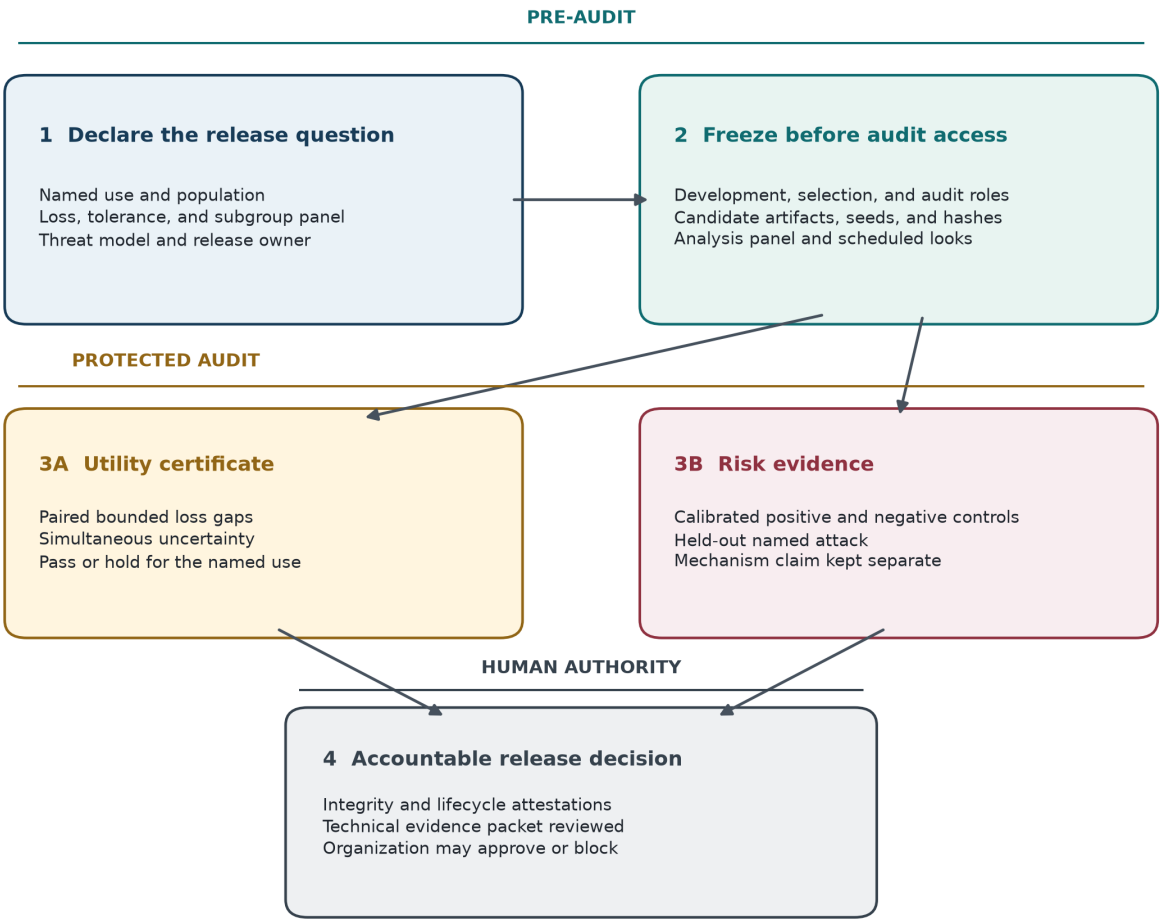


Figure 1: SynthGuard-ReleaseBench decision flow. The declared use, tolerance, candidate artifacts, and analyses are frozen before protected-audit access. Utility certification and empirical risk evidence remain separate, and neither can bypass the accountable release decision.

technical protocol. They are not design-based survey inference, are not evidence about household clustering, and are not a substitute for a live confidential audit. The paper makes no universal claim that any method is safe, private, or ready for deployment.

2 Related Work and Positioning

Partially synthetic data have a long history in disclosure control and official statistics [14]. Practical sequential-regression synthesis is available in widely used tools such as synthpop [11]. More recent tabular generators include deep conditional models such as CTGAN [33]. Differential privacy provides a formal framework in which a mechanism's output distribution has a bounded change under a stated neighboring-dataset relation [4]; private Bayesian-network and private-GAN approaches illustrate the diversity of possible generators [36, 6]. None of these approaches makes the analytical evaluation problem disappear: the usefulness of a synthetic release remains use-dependent.

Benchmarking libraries are essential for comparing generators. SDGym provides a general

benchmarking environment, while SDNist provides data and metrics developed for structured tabular synthetic-data evaluation [16, 10]. SDMetrics reports validity, structure, marginal, pairwise, and task-oriented measurements [26]; SynthEval combines configurable fidelity, utility, and privacy metrics [28]; and Anonymeter provides attack-based singling-out, linkability, and inference-risk evaluations [29]. The closest prior work in spirit is Houssiau et al.'s framework for auditable synthetic data generation [17]. There a generator is restricted to a small set of well-chosen statistics and the release carries a *generator card* declaring which information was used; a companion auditing procedure, reduced to a hypothesis test in a regression problem, lets a data holder verify that the card is accurate and that only approved statistics were consumed. The same group's TAPAS toolbox supplies adversarial privacy auditing for synthetic data [18]. Both share our conviction that a release should arrive with checkable evidence, not assurances.

The audits answer different questions, and the two are complementary rather than competing. Houssiau et al. audit the *generator*: given a declared card, did the synthesis use only the approved inputs? SynthGuard audits the *proposed release for a declared downstream use*: on a protected evaluation that was locked before it was opened, does a synthetic-trained workflow match its real-trained reference within a pre-committed tolerance, simultaneously across a whole candidate-by-analysis panel, with a declared negative control required to fail and a governance gate that can withhold a technically passing release? A generator card certifies provenance of information; a use-specific certificate bounds downstream equivalence with finite-sample uncertainty. Neither implies the other: a generator can consume only approved statistics and still be useless for the named analysis, and a release can pass our equivalence audit while its card is silent about what it consumed. An organization plausibly wants both.

Systematic assessments of tabular synthesis also exist: Du and Li evaluate eight synthesizer families on twelve real data sets and propose improved fidelity, privacy, and utility metrics [25]. That work and ours ask different questions. A systematic assessment compares synthesizers under a shared measurement suite; SynthGuard instead fixes a single named use and asks whether one proposed release has evidence for it. The difference is procedural, not a matter of which metric is used: we require the candidate panel, analysis panel, tolerances, and audit schedule to be committed and hashed before the protected audit is opened, we report a simultaneous finite-sample bound over the whole candidate-by-analysis panel rather than per-metric point estimates, and we require a declared negative control to fail before any result is admitted. These systems are valuable and have different aims. NIST's public challenge experience also illustrates the value of explicit evaluation design when comparing privacy-preserving synthesis methods [15]. Earlier statistical work framed utility evaluation as a comparison of inferences from protected and original data, often alongside disclosure risk [7]. Privacy evaluation is an active and necessary area. Membership inference asks whether a record was in training data [30]; broad empirical studies have shown that synthetic-data privacy and utility can be difficult to predict across models and data sets [32]. SynthGuard is designed around that uncertainty. It does not turn one attack result into a global privacy score.

Table 1 states the narrower position of this work. The comparison describes the roles played by common artifact types, not an exhaustive claim about every existing package or paper. In particular, SynthGuard does not compete with a generator or a formal DP proof. It makes a release-evidence workflow auditable.

The point of table 2 is not that the existing tools are incomplete implementations. They serve different purposes. The accompanying five-scenario audit makes the distinction executable: audit reuse, a passing negative control, unregistered post-hoc metrics, and a missing attestation trigger a SynthGuard hold, while a matching Anonymeter attack remains the more specialized privacy-risk instrument. SynthGuard requires these outputs and controls to travel together in one reviewable record, and it keeps a technical certificate, empirical attack evidence, and human release authority

Table 1: Scope of common synthetic-data artifacts and SynthGuard-ReleaseBench. "Usually" describes the usual role of the artifact type, not every implementation.

Artifact type	Named analytical use	Locked held-out audit	Simultaneous panel decision	Threat-model-specific risk report
Generator paper or package	Sometimes	Rarely	Rarely	Sometimes
Generic benchmark or leaderboard	Sometimes	Usually public	Rarely	Sometimes
Formal DP mechanism	No utility target implied	Not required	No	Mechanism-level guarantee
SynthGuard-ReleaseBench	Yes	Yes, for a live sealed audit	Yes	Yes, reported separately

Table 2: Documented-scope comparison with established evaluation packages. "Outside default scope" is not a defect; it means that a surrounding workflow must supply the named release control. The machine-readable scenario matrix is included in the release.

Tool	Fidelity and utility metrics	Dedicated privacy attacks	Detects reuse of a protected audit	Requires negative-control rejection	Governance gate and artifact hashes
SDMetrics [26]	Yes	Metric dependent	Outside default scope	Outside default scope	Outside default scope
SynthEval [28]	Yes	Yes, configurable	Outside default scope	Outside default scope	Outside default scope
Anonymeter [29]	Limited	Yes, core role	Outside default scope	Outside default scope	Outside default scope
SynthGuard-ReleaseBench	Declared use panel	Named attack record	Blocks invalid audit reuse	Required	Required together

separate.

Methodological novelty. The new object is not another realism score or generator leaderboard; it is a fail-closed statistical decision record that couples a predeclared use, a protected audit, simultaneous uncertainty, calibrated controls, threat-model-specific risk evidence, and accountable release authority without allowing any one component to stand in for the others.

3 Problem Setup

Let D_{dev} be a development sample, D_{sel} a generator-selection sample, and $D_{\text{audit}} = \{Z_i\}_{i=1}^{n}$ an audit sample. A candidate generator $g \in \mathcal{G}$ is fit using only the data allowed by its declared protocol. A synthetic data set S_g is produced before audit evaluation. For a declared analysis $a \in \mathcal{A}$, let $f_{R,a}$ be the workflow trained or calibrated on its permitted real-data reference and let $f_{S,g,a}$ be the corresponding workflow trained or calibrated on S_g. The exact construction of each workflow is part of the locked record.

The primary estimand is a use-level expected loss gap,

$$\Delta_{g,a} = \mathbb{E}\left[\ell_a(f_{S,g,a}, Z) - \ell_a(f_{R,a}, Z)\right], \tag{1}$$

where the loss is transformed in advance to a declared bounded scale. The audit estimator is

$$\widehat{\Delta}_{g,a} = \frac{1}{n_a} \sum_{i \in I_a} \left\{\ell_a(f_{S,g,a}, Z_i) - \ell_a(f_{R,a}, Z_i)\right\}. \tag{2}$$

The set I_a may be a locked subgroup. Thus a subgroup is not a post-hoc plot: it is a separate analysis with its own sample size, tolerance, and uncertainty bound.

The statistical target is equivalence within a declared tolerance, not superiority of a generator and not rejection of a point null. The logic is confidence-set inclusion, as in classical equivalence testing [27]. A candidate is technically certified for analysis a only when the simultaneous upper bound satisfies $|\widehat{\Delta}_{g,a}| + h_a \le \tau_a$. Otherwise the outcome is a *hold*: it may reflect a material gap, insufficient audit information, or both. Unless a separate lower bound excludes τ_a, a hold is not evidence that $|\Delta_{g,a}| > \tau_a$. Throughout the results, "fails a check" means "fails to certify under this rule."

The public protocol locks the data source, splits, generator versions, generation seeds, candidate grid, analysis panel, tolerances, and scheduled audit looks before headline evaluation. The sealed version additionally keeps audit records unavailable to generator developers. This separation is essential. Once an audit is repeatedly inspected to choose models or analyses, the simple finite-sample statement below no longer describes the full adaptive process; holdout reuse and adaptivity require their own validity analysis [5].

4 A Simultaneous Use-Specific Certificate

This section states the primary certificate, then delimits it. Section 4.1 gives the certificate and the planning rules that follow from it. Section 4.2 establishes what no version of it can support: an audit opened to unrestricted reuse is reconstructible, and the panel-size correction cannot be dropped. Section 4.3 sharpens the reported width without weakening the guarantee, and section 4.4 relaxes the row-level independence assumption for clustered and time-ordered records.

4.1 The primary certificate and its planning rules

We state the primary certificate in a simple form so that its guarantees and its limits are visible. It uses standard concentration, not a new probability inequality.

Proposition 1 (Scheduled simultaneous bounded-loss audit)**.** *Fix a finite candidate panel $\mathcal{G}$, analysis panel $\mathcal{A}$, and a pre-specified sequence of audit looks $l \in \{0, \dots, L-1\}$. Conditional on all information fixed before the audit, suppose that for every candidate, analysis, and scheduled look, the paired losses*

$$X_{i,g,a} = \ell_a(f_{S,g,a}, Z_i) - \ell_a(f_{R,a}, Z_i)$$

are iid and lie in $[-B_a, B_a]$. Let $q_l > 0$ be fixed spending weights with $\sum_l q_l \le 1$, and let $n_{a,l}$ be the scheduled audit size for analysis a at look l. Define

$$h_{a,l} = B_a \sqrt{\frac{2\log\{2|\mathcal{G}||\mathcal{A}|/(\alpha q_l)\}}{n_{a,l}}}. \tag{3}$$

Then, with probability at least $1-\alpha$, all declared candidates, analyses, and scheduled looks satisfy

$$\left|\widehat{\Delta}_{g,a,l} - \Delta_{g,a}\right| \le h_{a,l}.$$

Consequently, the rule $|\widehat{\Delta}_{g,a,l}| + h_{a,l} \le \tau_a$ implies $|\Delta_{g,a}| \le \tau_a$ on that event.

Proof. For a fixed (g, a, l), Hoeffding's inequality for variables in $[-B_a, B_a]$ gives

$$\Pr\left(\left|\widehat{\Delta}_{g,a,l} - \Delta_{g,a}\right| > h_{a,l}\right) \le \frac{\alpha q_l}{|\mathcal{G}||\mathcal{A}|}.$$

A union bound over candidates, analyses, and scheduled looks gives a total error probability at most $\alpha \sum_l q_l \leq \alpha$. The final statement follows from the triangle inequality. □

No independence is required across candidates or analyses: the same audit records may induce strong dependence among their paired losses. The union bound only requires the stated iid condition across audit units for each fixed bounded loss sequence.

Corollary 2 (Selection before audit)**.** *Under the conditions of proposition 1, let $\widehat{g}$ be selected by any rule measurable with respect to development data, selection data, pre-audit randomness, and the frozen candidate artifacts, but not the protected audit outcomes. Then the certificate in proposition 1 holds for every declared analysis of $\widehat{g}$ with probability at least $1 - \alpha$.*

Proof. Conditional on all pre-audit information, $\widehat{g}$ is one member of the fixed panel $\mathcal{G}$. The simultaneous event in proposition 1 holds for every member of that panel, so it holds for the selected member. □

Proposition 3 (Sufficient audit size for a certification margin)**.** *Under the conditions of proposition 1, fix a candidate, analysis, and scheduled look. Suppose the true gap lies inside its tolerance by a known planning margin $0 < \eta_a \leq \tau_a$,*

$$|\Delta_{g,a}| \leq \tau_a - \eta_a.$$

If

$$n_{a,l} \geq \frac{8B_a^2}{\eta_a^2} \log\left\{\frac{2|\mathcal{G}||\mathcal{A}|}{\alpha q_l}\right\}, \tag{4}$$

then the candidate passes the declared check on the joint $1 - \alpha$ event from proposition 1.

Proof. Condition (4) makes $h_{a,l} \leq \eta_a/2$. On the joint concentration event,

$$|\widehat{\Delta}_{g,a,l}| + h_{a,l} \leq |\Delta_{g,a}| + 2h_{a,l} \leq \tau_a - \eta_a + \eta_a = \tau_a.$$

□

The bound is a pre-audit planning result, not observed power and not a promise that an unknown candidate will pass. It makes the cost of a larger candidate panel, a larger analysis panel, or more scheduled looks explicit. The command-line tool implements the equal-spending version of (4).

Proposition 4 (Valid correction for controlled audit adaptation)**.** *Suppose that, before audit access, the protocol enumerates a finite menu $\mathcal{T}$ of at most M permitted adaptation transcripts. Each transcript fixes a bounded gap function for every declared candidate, analysis, and look, but the final transcript may be selected after inspecting the audit. Replacing* (3) *by*

$$h_{a,l}^{(M)} = B_a \sqrt{\frac{2\log\{2M|\mathcal{G}||\mathcal{A}|/(\alpha q_l)\}}{n_{a,l}}}$$

gives simultaneous coverage for every transcript and therefore for the audit-selected transcript.

Proof. Apply Hoeffding's inequality to every transcript, candidate, analysis, and scheduled look, allocating $\alpha q_l/(M|\mathcal{G}||\mathcal{A}|)$ to each event. A union bound gives total error probability at most $\alpha \sum_l q_l \leq \alpha$. Because the event holds for all $t \in \mathcal{T}$, it also holds for the transcript chosen after audit inspection. □

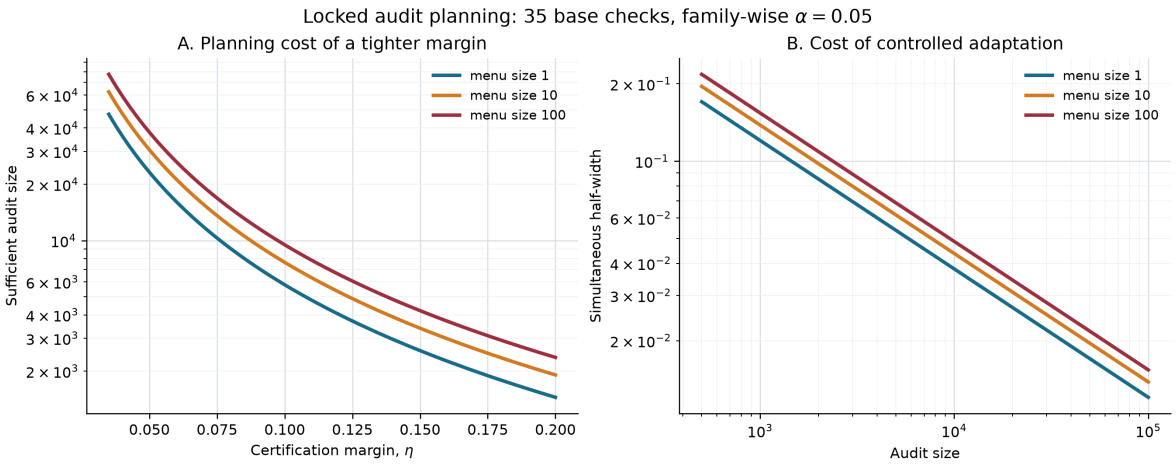


Figure 2: Pre-audit planning under bounded paired losses. Panel A gives the sufficient audit size in proposition 3 for 35 base checks as the certification margin changes. Panel B shows the simultaneous half-width when the fixed adaptation menu in proposition 4 has size 1, 10, or 100. The logarithmic adaptation cost is much smaller than the quadratic cost of halving the margin, but it is not zero.

Proposition 4 provides a narrow correction, not permission for open-ended tuning. The adaptation branches must be concrete enough to enumerate and reproduce before audit access. Figure 2 shows the resulting planning cost for 35 base checks.

For the completed public ACS studies, losses were scaled so $B_a = 1$. We used the pre-specified spending schedule $q_l = 6/\{\pi^2(l+1)^2\}$, which sums to one over infinitely many planned looks. The public runs report the whole candidate grid. Selecting a candidate on D_{sel} is permitted by the protocol; selecting it after observing D_{audit} is not.

Remark 1 (What the certificate means)**.** *The certificate is a statement about the declared loss, tolerance, population represented by the iid audit assumption, and frozen workflow. It is not a confidence statement about arbitrary downstream analyses. It does not establish that the synthetic rows are private, non-identifying, legally releasable, survey-design valid, or appropriate for an untested subgroup. A failure can arise from real utility loss, an overly strict tolerance, a small audit sample, or a conservative bound. In all of these cases, the correct outcome is* hold or uncertified*, not a forced pass/fail narrative about a generator.*

4.2 Two lower bounds on what an audit can support

Proposition 5 (Sharp failure under unrestricted audit reuse)**.** *Let P be a non-atomic audit distribution on a standard Borel space and let $S = (Z_1, \ldots, Z_n)$ be the observed audit sample. If an analyst may choose an arbitrary bounded gap function after seeing S, there is a measurable choice $X_S : \mathcal{Z} \to [0,1]$ with empirical audit mean zero and population mean one almost surely. Hence, for any $h \le \tau < 1$, a rule that accepts when $|\widehat{\Delta}| + h \le \tau$ can accept while the true gap exceeds tolerance.*

Proof. After observing S, define $X_S(z) = 0$ when $z \in \{Z_1, \ldots, Z_n\}$ and $X_S(z) = 1$ otherwise. Its empirical mean on S is zero. Because P is non-atomic, the finite observed set has probability zero, so $\mathbb{E}_P[X_S(Z)] = 1$ almost surely. Therefore $|\widehat{\Delta}| + h = h \le \tau$, while $|\Delta| = 1 > \tau$. □

Proposition 5 is an elementary but sharp release-audit lower bound. It does not say that all adaptive analysis is impossible. It says that unrestricted, unrecorded reuse cannot inherit the locked-audit guarantee. A finite, pre-declared adaptation menu can be handled by including every permitted

branch in the simultaneous panel; richer reuse needs a reusable-holdout or adaptive-inference design [5].

One objection to proposition 5 is that its witness function is pathological: a real analyst does not write down the indicator of the observed audit set. The next result removes that objection. It shows the same failure arises from ordinary bounded queries under a realistic query budget, so the danger is operational, not measure-theoretic.

Proposition 6 (Reconstruction from bounded audit queries)**.** *Assume the analyst knows which records constitute the protected audit, that is their identifiers, but not their hidden binary labels $b \in \{0,1\}^n$, and may submit bounded gap functions and read back only their reported empirical audit means to full precision. For $j = 1, \dots, m$, let $s_j \in \{-1,1\}^n$ be a publicly seeded pseudorandom sign vector indexed by identifier, and let query j report the bounded statistic*

$$y_j = \frac{1}{n} \sum_{i=1}^{n} s_{ji} \, (2b_i - 1) \in [-1, 1].$$

Write $S \in \{-1,1\}^{m \times n}$ for the stacked signs. If $m \geq n$, then S has full column rank with probability at least $1 - (1/2 + o(1))^n$, and on that event b is determined exactly by $y = S(2b - \mathbf{1})/n$.

Proof. The map $b \mapsto S(2b - \mathbf{1})/n$ is affine and injective whenever S has full column rank, in which case inverting recovers $2b - \mathbf{1}$ and hence b. For $m \geq n$ it suffices that the leading $n \times n$ block be non-singular, and the probability that a square Rademacher matrix is singular is at most $(1/2 + o(1))^n$ [19]. □

The identifier assumption is what makes the attack an ordinary one, and it is the standard setting of database reconstruction [20]: audit membership is often administratively known, for example a published list of sampled record keys, while the sensitive labels are exactly what the audit protects. Note also that full column rank is a high-probability event, not a certain one; a Rademacher design is singular with positive probability, so the statement cannot be strengthened to hold almost surely.

The consequence for a release audit is a certificate that accepts a memorising candidate. Let the declared analysis be zero-one prediction loss, let f_R be the real-trained reference with audit loss $\widehat{L}(f_R)$, and let f_S return the recovered label on any audit identifier and a fixed class elsewhere. On the reconstruction event f_S has audit loss zero, so its empirical gap is $\widehat{\Delta} = -\widehat{L}(f_R)$ and the rule $|\widehat{\Delta}| + h \leq \tau$ accepts whenever $\widehat{L}(f_R) \leq \tau - h$, which is the ordinary situation of a competent reference. Its population gap is $\Delta = L(\text{const}) - L(f_R)$, the excess loss of the constant rule, which is strictly positive whenever the reference beats that rule and exceeds τ for any reference worth auditing. The audit therefore certifies a lookup table.

Proposition 6 is the release-audit specialization of classical database reconstruction [20] and of the adaptive data analysis lower bounds [21, 22]. Figure 3 reports the empirical threshold. With $n = 400$ audit rows, mean recovery is 0.64 at a budget of $0.1n$ queries and 0.84 at $0.5n$, with no exact recoveries; exact recovery occurs in half of replicates at $0.9n$ and in every replicate once the budget reaches n. The transition sits where the theory places it, at a query budget comparable to the audit size. Two operational consequences follow. Query budgets against a protected audit must be counted and capped, not left implicit; and an audit that has answered on the order of n adaptive queries should be treated as spent, exactly as a locked panel is spent once inspected.

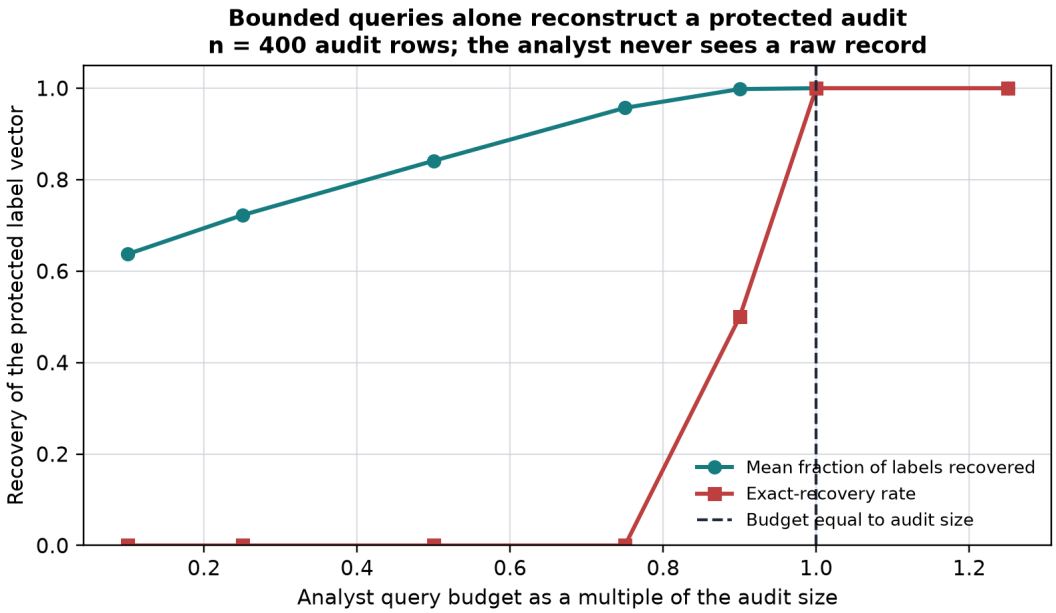


Figure 3: Operational audit-reuse attack. The analyst sees only reported bounded query means, never a raw audit row. Recovery of the protected label vector is partial while the query budget is below the audit size and becomes exact at a budget comparable to it, matching proposition 6. The dashed line marks a budget equal to the audit size.

A companion question is whether the multiplicity correction in proposition 1 is an artifact of the union bound. It is not.

Proposition 7 (Necessity of the panel correction)**.** *Fix $n \geq 1$ and $\alpha \in (0, 1/2)$. Let the panel consist of K candidates whose paired gaps are iid Rademacher with true gap zero, independent across candidates. Any procedure that reports a common half-width h for every candidate and attains simultaneous coverage $1 - \alpha$ must satisfy*

$$h \;\geq\; c_\alpha \min\Big\{1,\; \sqrt{\tfrac{\log K}{n}}\Big\},$$

where $c_\alpha > 0$ depends only on α. In particular a half-width that ignores K cannot retain simultaneous coverage as K grows.

Proof. Write $\bar{X}_k = n^{-1} \sum_i \varepsilon_{ik}$ for the kth candidate mean and $p = \Pr(|\bar{X}_1| > h)$. The K means are independent, so simultaneous coverage requires $(1 - p)^K \geq 1 - \alpha$, hence $p \leq 1 - (1 - \alpha)^{1/K} \leq \log\{1/(1 - \alpha)\}/K$.

A Rademacher average obeys a matching anti-concentration bound: there are absolute constants $c_1, c_2 > 0$ with $\Pr(|\bar{X}_1| > t) \geq c_1 \exp(-c_2 n t^2)$ for all $0 \leq t \leq 1/2$, which follows from the central limit theorem with a Berry–Esseen correction, or from a direct binomial tail estimate. Suppose first that $h \leq 1/2$. Combining the two displays gives $c_1 \exp(-c_2 n h^2) \leq \log\{1/(1 - \alpha)\}/K$, so

$$h^2 \;\geq\; \frac{1}{c_2 n} \log \frac{c_1 K}{\log\{1/(1 - \alpha)\}},$$

which is of order $\log(K)/n$ once K exceeds a constant depending on α. If instead $h > 1/2$, the claimed bound holds trivially because its right-hand side is at most c_α. Taking c_α small enough to cover both cases and the finitely many small K gives the statement. □

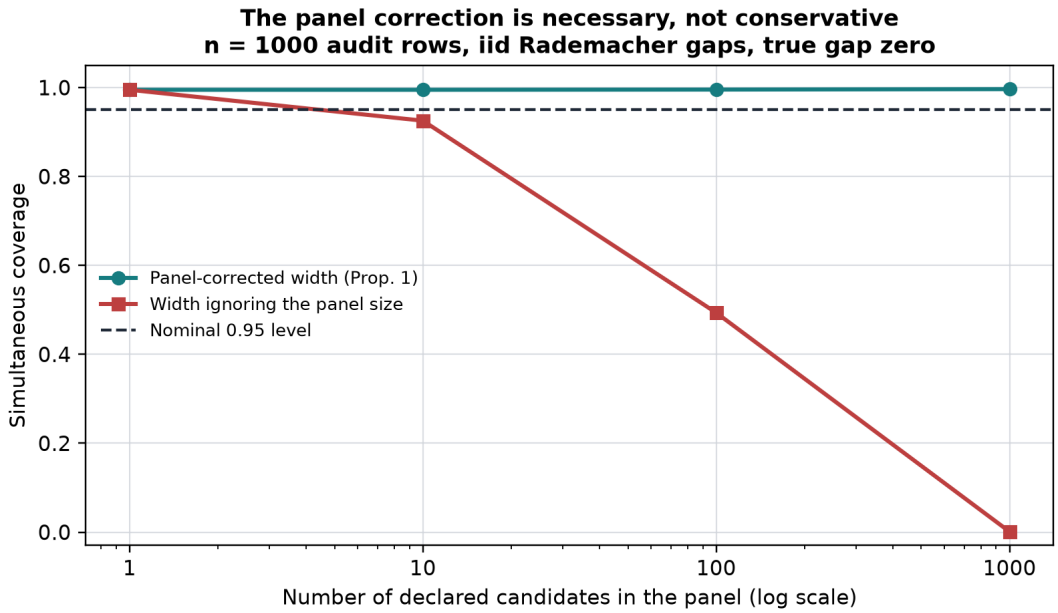


Figure 4: Necessity of the panel correction at $n = 1{,}000$ with iid Rademacher gaps and true gap zero. Ignoring the panel size loses simultaneous coverage as the panel grows; the corrected width in proposition 1 holds its nominal level.

The truncation at one matters. the gaps lie in $[-1, 1]$, so $h = 1$ always covers and no lower bound may exceed it. The bound therefore bites in the regime that matters, $\log K \lesssim n$, and says that within it the logarithmic panel cost is unavoidable, not an artifact of the union bound.

Figure 4 confirms this numerically at $n = 1{,}000$: a half-width computed for a single candidate retains 0.995 simultaneous coverage at $K = 1$, but falls to 0.926 at $K = 10$, 0.494 at $K = 100$, and 0.001 at $K = 1{,}000$, while the panel-corrected width from proposition 1 stays at or above 0.995 throughout. The logarithmic cost of a large declared panel is therefore real, and it is the price of being allowed to lock many candidates at once rather than a conservatism that a sharper analysis could remove.

4.3 Variance-adaptive instruments for the same audit

Proposition 1 pays for the full declared range $[-B_a, B_a]$ whatever the data look like. Paired loss gaps in practice are concentrated near zero, because the real-trained and synthetic-trained workflows agree on most records and disagree on a minority. The range is then a poor proxy for the uncertainty, and the certificate is wider than the evidence requires. Two standard instruments fix this without weakening the guarantee.

Proposition 8 (Variance-adaptive scheduled certificate)**.** *Under the conditions of proposition 1, let $\widehat{V}_{g,a,l}$ be the sample variance of the paired gaps in the cell (g, a, l) and set $\delta_{a,l} = \alpha q_l/(|\mathcal{G}||\mathcal{A}|)$. Replacing* (3) *by the empirical Bernstein half-width*

$$h^{\mathrm{EB}}_{g,a,l} = \sqrt{\frac{2\widehat{V}_{g,a,l}\log(4/\delta_{a,l})}{n_{a,l}}} + \frac{14B_a\log(4/\delta_{a,l})}{3(n_{a,l}-1)} \tag{5}$$

preserves the conclusion of proposition 1 with the same $1-\alpha$ simultaneous guarantee.

Proof. Maurer and Pontil's inequality is stated for observations in an interval of unit length [23]. Rescale the gaps by their range $R_a = 2B_a$, so that X/R_a lies in an interval of unit length with sample variance $\widehat{V}/R_a^2$. Applying the one-sided inequality to each tail at level $\delta_{a,l}/2$ and multiplying through by R_a leaves the variance term unchanged, because $R_a\sqrt{\widehat{V}/R_a^2} = \sqrt{\widehat{V}}$, and multiplies the second term by $R_a = 2B_a$, which is the source of the factor $14 = 7 \times 2$ in (5). Union bounding over candidates, analyses, and looks exactly as in proposition 1, the spending weights again give total error at most $\alpha \sum_l q_l \leq \alpha$. □

The range factor matters in practice as well as in principle: with $B_a = 1$ the second term is $28\log(4/\delta)/\{3(n-1)\}$, roughly 0.0025 at the final Adult look, so dropping the factor of two would understate the reported half-width by about a fifth. The implementation and every number below use the range $R_a = 2$.

A second instrument removes the look schedule altogether. A predictable plug-in empirical Bernstein confidence sequence [24] is valid simultaneously at every sample size, so it may be evaluated at any stopping time, including a data-dependent decision to stop the audit. Multiplicity across the panel is still paid by dividing α, but no alpha spending across looks is needed, because time-uniformity is built into the construction.

Both instruments must be pre-registered like any other element of the locked panel. Neither may be applied after the fact to revise a recorded outcome: choosing the bound family after seeing which one passes is precisely the adaptation that proposition 5 forbids. We therefore report them as a measurement on the already-locked Adult evidence, with every recorded pass and fail left untouched, and we recommend them for pre-registration in future panels.

Figure 5 reports that measurement. The gap vectors are recomputed from the byte-reproducible locked artifacts, and the recomputed Hoeffding half-widths reproduce the locked certificate to within 10^{-9}, which is the gate that makes this a comparison of instruments on identical evidence, not a new experiment. At the final 16,000-record look the Hoeffding half-width is 0.0316 for every candidate. The scheduled empirical Bernstein half-widths range from 0.0067 for the Gaussian copula to 0.0138 for TVAE, a tightening of 2.3 to 4.7 times; the anytime confidence sequence ranges from 0.0032 to 0.0082, a tightening of 3.9 to 9.8 times. The ordering is informative: the copula tightens most because its gaps are the most concentrated, and TVAE least because its gaps are the most dispersed, so the variance-adaptive width reports something about each candidate that the range-based width cannot.

The practical consequence is a much larger effective audit. Under proposition 3 the audit size needed for a target margin scales as the squared half-width, so a fourfold tightening corresponds to roughly a sixteenfold reduction in required audit rows at fixed margin, or a proportionally finer certifiable tolerance at fixed sample size. For the locked Adult panel the Hoeffding width consumed slightly more than half of the 0.060 tolerance before any evidence was considered; the empirical Bernstein width consumes about a ninth of it.

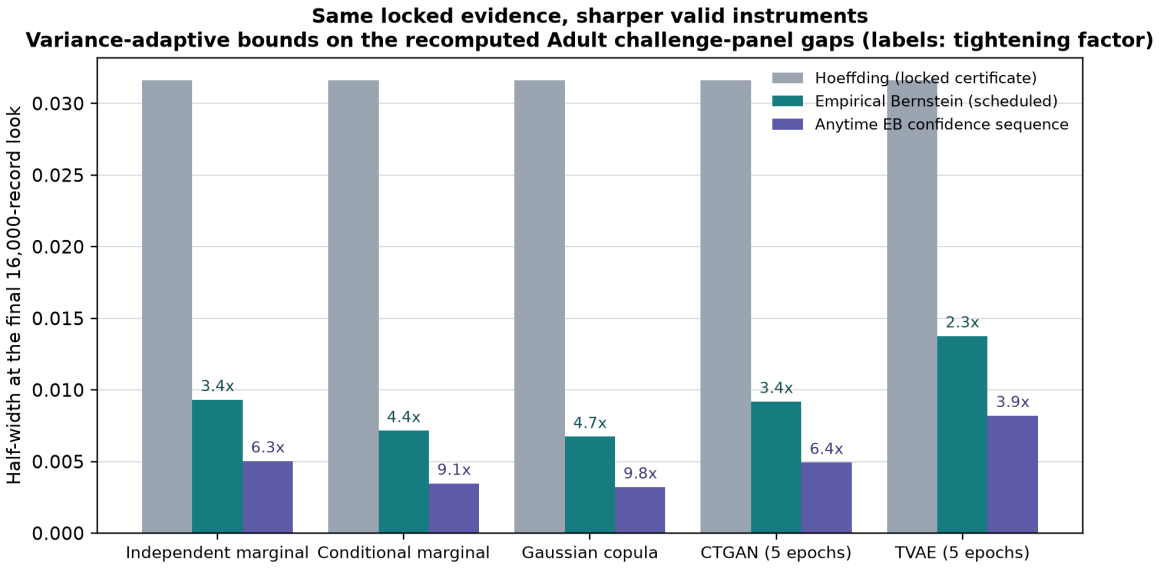


Figure 5: Same locked evidence, sharper valid instruments. Half-widths at the final 16,000-record look of the Adult challenge panel, recomputed from the locked artifacts. Labels give the tightening factor relative to the Hoeffding certificate. The comparison revises no recorded outcome; it quantifies what a pre-registered variance-adaptive instrument would have delivered.

4.4 Audit units beyond the row

The certificate so far assumes the audit rows are exchangeable. Two common designs break that assumption in different ways, and each needs its own audit unit rather than a correction factor.

Proposition 9 (Scheduled simultaneous independent-block audit)**.** *Fix the same candidate, analysis, and scheduled-look panels. For each analysis a, suppose that its pre-specified audit blocks $b = 1, \ldots, m_{a,l}$ are independent and identically distributed conditional on the locked development process, and that the within-block mean paired loss gap $\bar{X}_{b,g,a}$ lies in $[-B_a, B_a]$. Replacing $n_{a,l}$ in equation (3) by $m_{a,l}$ gives a simultaneous bound for the equally block-weighted estimand $\Delta^{\text{block}}_{g,a} = \mathbb{E}[\bar{X}_{b,g,a}]$. Thus the rule $|\widehat{\Delta}^{\text{block}}_{g,a,l}| + h^{\text{block}}_{a,l} \leq \tau_a$ is valid on the same $1-\alpha$ event.*

Proof. For each fixed candidate, analysis, and look, the independent bounded block means satisfy Hoeffding's inequality. Applying the same alpha-spending allocation and union bound as in proposition 1 proves the claim. □

Remark 2 (Block scope)**.** *The block certificate changes the estimand: every independent block has equal weight, not every row. It requires the same named blocks for every candidate in an analysis. The implementation records both the number and sizes of blocks, but no algorithm can infer independence from a CSV file. In particular, the public BRFSS and Chicago temporal tracks below are not retroactively certified by this proposition.*

Stress test of the audit unit. We ran a fixed-seed 2,000-replicate diagnostic with 50 independent blocks of 200 perfectly dependent rows each. Within a block, all rows shared one Rademacher loss gap, so the true equally block-weighted gap was zero. Treating all 10,000 rows as iid gave empirical coverage of only 0.0975 (Monte Carlo standard error 0.0066) for the nominal 0.95 zero-gap target. Applying proposition 9 to the 50 independent block means gave coverage 0.9965 (Monte Carlo

standard error 0.0013). Figure 6 is deliberately stylized: it validates the audit-unit distinction in a known dependence setting, not the dependence structure of any public data set.

Time-ordered audits. Proposition 9 needs blocks whose independence follows from the sampling design. Time-ordered audit records rarely offer that: consecutive rows are dependent, and the dependence decays rather than vanishing at a known boundary. This left the BRFSS and Chicago tracks outside the certificate, which was an honest gap but a gap nonetheless. The following fills it under a declared, and clearly falsifiable, dependence horizon.

Proposition 10 (Scheduled buffered-block temporal certificate)**.** *Fix the same candidate, analysis, and scheduled-look panels. Suppose the audit rows for analysis a arrive in a fixed time order and the paired gap process is m_a-dependent, meaning gaps separated by more than m_a positions are independent, with identically distributed block means across the window. Partition the window into non-overlapping blocks of length L_a separated by buffers of length at least m_a, and let $\bar{X}_{b,g,a}$ denote the retained block means, $K_{a,l}$ their number at look l. Then the retained block means are iid and bounded, so applying* (3) *or* (5) *with $n_{a,l}$ replaced by $K_{a,l}$ yields a simultaneous $1-\alpha$ bound for the equally block-weighted estimand $\Delta^{\mathrm{tmp}}_{g,a} = \mathbb{E}[\bar{X}_{b,g,a}]$.*

Proof. Two retained blocks are separated by at least m_a positions by construction, so under m_a-dependence their gap vectors are independent; hence the retained block means are independent, and identically distributed by assumption. Each is an average of gaps in $[-B_a, B_a]$ and so lies in the same interval. The cited inequalities apply to these $K_{a,l}$ bounded iid summands, and the alpha-spending union bound of proposition 1 carries over unchanged. □

Three limitations are structural, not incidental. The estimand changes: retained blocks are weighted equally and rows inside buffers are discarded, so the implementation reports both the number of retained blocks and the discarded row count. The horizon m_a is declared, never inferred; no algorithm can read a dependence horizon off a CSV file, exactly as none can verify exchangeability for the row certificate. And a mis-declared horizon invalidates the certificate, so the honest posture is to declare m_a generously.

A declared horizon is the bluntest available response to structured dependence, and we chose it on purpose. it buys exact independence between retained blocks at the cost of discarding buffer rows and of resting on an assumption no data can confirm. Finer instruments exist for the related problem of calibrated prediction under seasons, recurring regimes, and other structured non-exchangeability, for example spectral weighting of calibration residuals with an online correction to the miscoverage target [13]. Adapting that machinery to a release audit is not immediate, because the audit must commit to its uncertainty statement before the protected data are opened, whereas an online correction learns from the stream it is calibrating on. We therefore keep the blunt instrument here and flag the adaptive route as the natural next step for a temporal certificate.

Temporal coverage stress test. We validated the audit unit on a fixed-seed 2,000-replicate design with known dependence: 10,000 time-ordered rows in regimes of 200 consecutive rows sharing a single Rademacher gap, so the true row-mean and block-mean gap are both exactly zero. Regime boundaries are unknown to the auditor and blocks are deliberately misaligned with them, so independence comes only from the buffer exceeding the horizon. Treating all rows as iid gave empirical coverage 0.1105 (Monte Carlo standard error 0.0070) against the nominal 0.95. Proposition 10 with blocks and buffers of 250 rows gave coverage 0.9995 under the Hoeffding form and 1.0000 under the empirical Bernstein form. The temporal instrument is conservative here by construction, since the regime design makes block means maximally dispersed; the point is that the

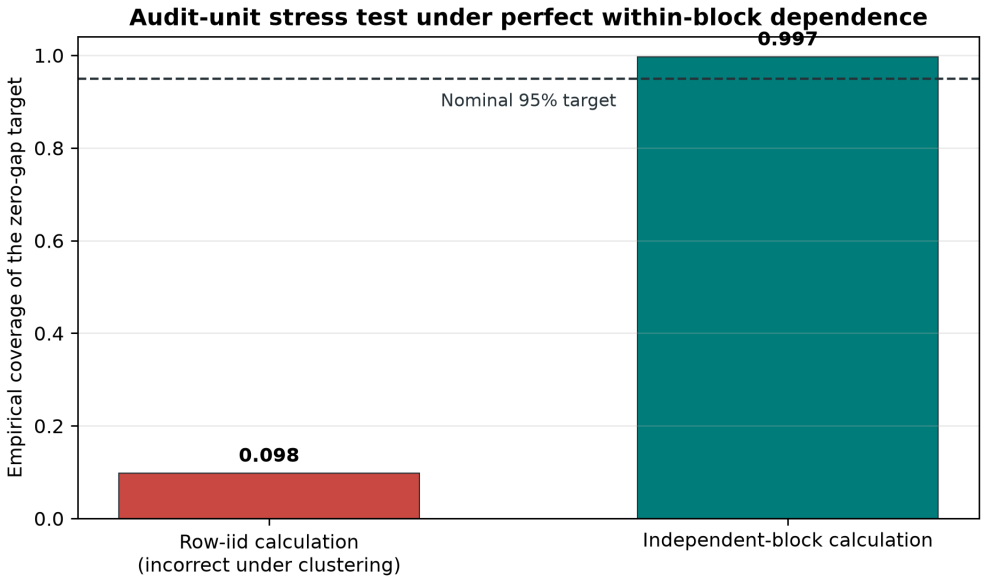


Figure 6: Audit-unit stress test under perfect within-block dependence. The row-iid calculation is intentionally invalid in this setup and severely undercovers. The independent-block calculation uses 50 block means and is conservative here. This is a reproducible diagnostic for the protocol, not a substantive data result.

row-level calculation is not merely conservative but wrong, missing its nominal level by a factor of eight.

4.5 Setting a tolerance, not merely locking one

Locking a tolerance prevents it from being tuned after the audit is opened. It does not make the number defensible, and every conclusion in this paper depends on that number. We therefore add a pre-audit protocol for choosing it, using development data only, so that running the protocol consumes no protected audit.

The protocol rests on a *real-versus-real floor*. Fit the declared workflow twice on two disjoint real samples of the size the synthetic candidate will receive, and score both against the same held-out development slice. The resulting absolute paired gap is irreducible: it is the disagreement the workflow already shows between two honest real datasets, with no synthesis involved. A tolerance below that floor asks a synthetic candidate to agree with the reference more closely than a second real sample of the same size does, which mostly measures sampling noise. The recommended tolerance is then the larger of this floor and a declared decision-cost anchor, the smallest loss difference the named use would act on, so that the tolerance is never tighter than the noise and never looser than the decision requires.

On the Adult case, with 60 replicates at the fitted size of 1,000 rows, the 95th-percentile real-versus-real gap is 0.0052. The declared decision-cost anchor for the prototyping use is 0.02, which therefore binds, and the protocol recommends 0.02. The locked tolerance was 0.060: about 11.6 times the noise floor and three times the recommended value. We report it instead of quietly restating the locked panel, because it means the Adult tolerance was more permissive than a principled procedure would have chosen, and the reader should discount the corresponding passes accordingly.

The consequence is sharper than a caution, and it connects directly to section 4.3. At the

recommended tolerance of 0.02 the range-based Hoeffding half-width of 0.0316 is by itself larger than the whole tolerance, so under that instrument *no* candidate can be certified at this audit size whatever the observed gaps. Empirical Bernstein narrows the widths to 0.0067–0.0138 but still certifies nothing, the selected copula missing at 0.0208. The anytime confidence sequence certifies exactly one candidate, the Gaussian copula at 0.0173, while the negative control, both learned models, and even the target-conditional marginal at 0.0201 remain uncertified. A defensible tolerance at this sample size is thus reachable only with a variance-adaptive instrument, and when it is reached it admits one candidate rather than four. This is the clearest case for the sharper bounds. without them, an auditor must choose between an indefensibly loose tolerance and certifying nothing at all.

Figure 7 reports the protocol and the resulting sensitivity. Panel B makes the fragility of a single locked number visible directly: each candidate's decision flips exactly at its own reported upper gap, so a reader can see how far each pass sat from the threshold, not only that it passed. The selected copula flips at 0.0457 and the target-conditional marginal at 0.0482, both comfortably below the locked 0.060 but far above the recommended 0.02; the two learned models flip near 0.15 and the control at 0.0959, so those failures are robust to any plausible choice.

Two limits apply. The floor is estimated for one workflow, one loss, and one fitted size, and must be recomputed for any other; and the decision-cost anchor is a judgement about the declared use that no data can supply. The protocol states both openly instead of leaving the tolerance unexplained.

The first limit is an instance of a general problem: a quantitative audit is only meaningful over the range of conditions its evidence actually covers, and a threshold calibrated at one scale silently stops applying at another. The same difficulty arises outside synthetic data, for example in prospectivity modelling, where predictions must be audited against the range of evidence that supports them rather than extrapolated across scales [12]. Our floor inherits that constraint directly: it is calibrated at 1,000 fitted rows, and section 7.6 shows the learned candidates behaving quite differently at 4,000 and 16,000, so a tolerance imported across sizes without recomputation would be unsupported by its own evidence.

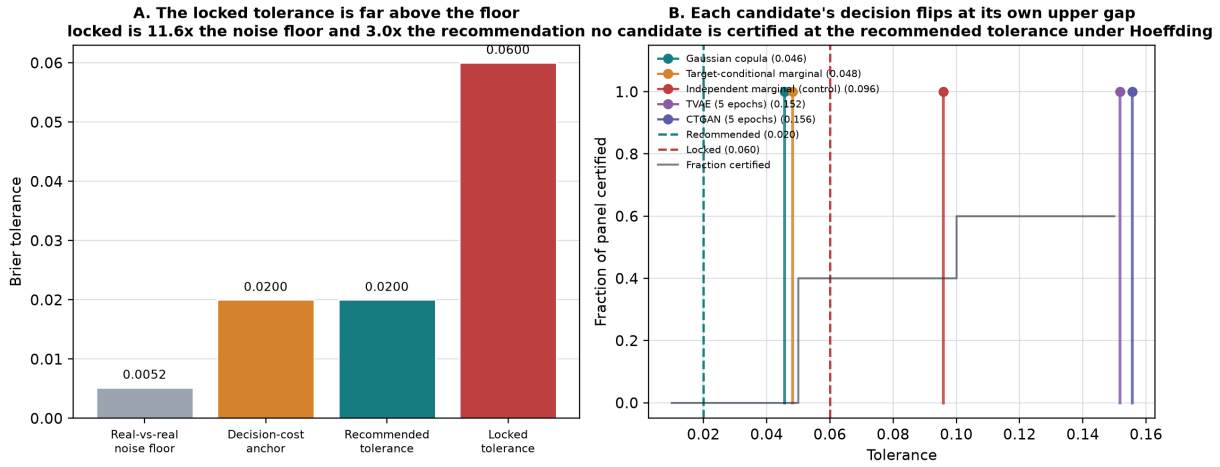


Figure 7: Pre-audit tolerance protocol on the Adult case. Panel A places the locked tolerance against the development-only real-versus-real noise floor and the declared decision-cost anchor; the locked value is about eleven times the floor. Panel B sweeps the tolerance and marks the point at which each candidate's decision flips, which is its reported upper gap. Failures of the negative control and both learned models are robust across the whole range; the two passes are not.

5 Release Evidence Is More Than a Certificate

SynthGuard has four evidence layers. First, the utility layer records every locked loss gap, tolerance, sample size, and simultaneous bound. Second, the privacy layer records a named attacker, its information and access, calibration data, and held-out evaluation result. Third, the mechanism layer records any formal privacy statement, including its adjacency relation and composition accounting. Fourth, the governance layer records whether the audit was independent of fitting and whether the panel was locked before evaluation.

This structure prevents a common but harmful shortcut: treating a favorable utility plot, a low empirical attack score, or an ε value as a universal release authorization. The output is an evidence packet, not a single scalar rank. A release gate may remain blocked even when all declared technical utility checks pass.

6 Benchmark Design

6.1 Public and sealed tracks

The public measurement track is intentionally reproducible. Its data source, splits, code, seeds, and output artifacts are available for scrutiny. This makes it useful for method comparison and regression testing, but it also means that it cannot stand in for a confidential audit of a future release. The sealed-audit track addresses that gap: an organization runs the same evaluator locally; it keeps audit rows inside the organization; and it exports only aggregate receipts, a technical certificate, and a blocked-or-not-blocked governance result.

6.2 Candidate families and controls

The completed studies use transparent baselines instead of claiming to establish a current best generator contest. The negative control independently resamples each feature and deliberately

destroys dependence. The non-private utility comparator samples features conditionally on the target but makes no privacy claim. The DP candidates use a target-conditional, independent-histogram mechanism with Laplace noise. Under add/remove-one-record adjacency and fixed public domains, each mechanism releases one target histogram and ten target-by-feature contingency tables; basic sequential composition gives pure ε-DP with $\varepsilon/11$ allocated to each histogram release. Clipping, normalization, and sampling are post-processing.

This deliberately narrow generator set makes the protocol auditable and exposes a basic failure mode. The record-disjoint California extension adds a regularized Gaussian-copula comparator. A fresh UCI Online Shoppers stress test further adds one compact, non-private CTGAN under a fixed CPU budget; it is a deliberately limited learned baseline, not a neural-model contest. A newly locked UCI Adult case then runs the full challenge panel, adding a matched small TVAE next to CTGAN under the identical budget [33]. A post-audit Adult sensitivity arm repeats only the learned slots at their library-default 300-epoch budgets; it addresses compute sensitivity but cannot be added to the locked certificate panel after audit access. A fresh patient-unique UCI Diabetes 130-US Hospitals panel applies the same matched learned slots; because its negative control passes the final grid, it is retained as excluded calibration evidence rather than added as a certificate case [40]. A prior exploratory CTGAN experiment is retained in the repository only as an invalidated record because its purported fresh replication was not record-disjoint. It is not used as primary evidence in this paper. This correction is documented, not silently removed.

6.3 Completed studies

California DP frontier. From the 2024 ACS California public microdata, we formed a fixed 80,000-row cohort: 48,000 development rows, 16,000 selection rows, and 16,000 audit rows. The candidates were the independent-marginal negative control, the target-conditional non-private comparator, and three pure-DP conditional histograms with $\varepsilon \in \{0.5, 1, 4\}$. Five analyses were locked: predictive Brier loss, clipped log loss, normalized income mean-squared error, and Brier loss in two pre-specified sex groups. The candidate selected on selection Brier loss was the $\varepsilon = 0.5$ DP histogram.

California generator-breadth extension. We constructed a second 80,000-row California cohort from source rows disjoint from the frontier cohort: 48,000 development rows, 16,000 selection rows, and 16,000 protected audit rows. The fixed six-candidate panel added a target-conditional regularized Gaussian copula to the two non-private marginal candidates and the three DP histograms. The extension used the same five locked analyses, tolerances, and scheduled audit sizes. The Gaussian-copula comparator was selected on the separate selection Brier score; it is a non-private statistical comparator, not a privacy mechanism.

New York robustness grid. A fresh 100,000-row New York ACS cohort used 50,000 development, 20,000 selection, and 30,000 audit rows. We ran a fully pre-specified 33-candidate grid: three training sizes (2,500, 10,000, and 50,000), the two non-private controls, and the three DP budgets at three fixed seeds. The rare-group diagnostic was fixed before the run. No candidate was selected after audit inspection.

Florida RiskLab. A 50,000-row Florida ACS cohort was divided into a 20,000-row generator-training partition, 5,000-member and 5,000-nonmember calibration partitions, 15,000-member and 10,000-nonmember evaluation partitions, and a 15,000-row unused reserve. The named membership threat model uses a full-record nearest-neighbor score under a documented public-domain encoding.

The attack threshold is calibrated without the held-out evaluation records. We also report an attribute-prediction generalization-gap diagnostic. A row-resampling positive control is deliberately included so that a weak attack battery cannot look reassuring merely because it cannot detect an obvious leak.

Non-ACS control calibration. The UCI Bank Marketing pilot used 27,000 development rows, 9,000 selection rows, and 9,000 audit rows, with pre-contact predictors only. It is retained as a threshold-calibration failure because its independent-marginal negative control passed the final technical grid. A separately locked UCI Default of Credit Card Clients study used 15,000 development rows, 6,000 selection rows, and 9,000 audit rows. It excluded the source sex field, used one primary overall-Brier analysis, and required rejection of the independent-marginal control. These public historical records are used only to test the audit protocol; they are not used to make or validate credit decisions, eligibility decisions, fairness claims, or deployment decisions [9, 37].

Fresh fixed-budget learned baseline. The UCI Online Shoppers study used 5,000 development sessions, 1,330 selection sessions, and 6,000 audit sessions from a fixed public historical table. Four non-private candidates were fixed: independent marginal, target-conditional marginal, a regularized Gaussian copula, and CTGAN. Each used the same 1,000-session random development subsample. CTGAN was fixed at five CPU epochs, batch size 500, PAC 10, and 64-dimensional embedding and hidden widths. The runner acquired an exclusive process lock, wrote candidate artifacts and hashes, and wrote its locked manifest before reading selection or audit data. This technical stress test neither supports advertising or consumer targeting nor supports privacy, legal, or deployment claims [38].

Fresh five-role challenge panel. The UCI Adult study is the first case run under the multi-domain challenge panel rule. The locked public cohort has 48,000 sampled records: 24,000 for development, 8,000 for selection, and 16,000 protected audit rows. All five required candidates were fixed before the audit: the independent-marginal negative control, the target-conditional marginal, the regularized Gaussian copula, and small CTGAN and TVAE models under the shared CPU-only budget of five epochs, batch size 500, 64-dimensional embeddings, and two 64-unit hidden layers [39, 33]. Every candidate was fit on the same random 1,000-record development subsample. Two earlier UCI Adult attempts are kept in the repository as invalidated records: one stopped before the audit, and one was ruined when two processes wrote into the same folder. This study therefore uses its own unique result folder, refuses to start while any earlier runner lock exists, writes its compressed artifacts with a fixed timestamp so their hashes can be reproduced byte for byte, keeps wall-clock timings out of the locked evidence files, and publishes an automatic consistency check across its configuration, metadata, artifact hashes, and summary.

Post-audit compute sensitivity. After inspecting the locked Adult audit, we ran a separate sensitivity arm on the identical 1,000-record development subsample. CTGAN and TVAE each used 300 CPU epochs and batch size 500. CTGAN used a 128-dimensional embedding, two 256-unit generator and discriminator layers, and PAC 10; TVAE used a 128-dimensional embedding and two 128-unit compression and decompression layers. The selection partition, audit rows, schedule, and tolerance were reused only to measure how the learned-model gaps changed with a larger budget. Because the protected audit had already been inspected, this arm was not pre-registered, its bounds are labeled nominal-only, and it makes no certificate or candidate-selection claim.

Fresh patient-unique health calibration. The UCI Diabetes 130-US Hospitals study retains one source record per patient and uses 8,000 development, 2,000 selection, and 8,000 protected-audit records. It excludes direct identifiers, race, gender, weight, and diagnosis-code fields. Its five candidates, the independent marginal, target-conditional marginal, regularized Gaussian copula, compact CTGAN, and compact TVAE, all share a 1,000-record development subsample and the same CPU-only five-epoch, batch-500 learned-model budget. Candidate artifacts and hashes were written before the selection and audit files were read. This is a research-only public historical health-table audit; it is not clinical prediction, medical advice, treatment recommendation, fairness, privacy-compliance, legal, or deployment evidence [40].

Outcome-free external industrial challenge. On 2026-07-22 we froze a separate protocol around the UCI Steel Industry Energy Consumption table, a 35,040-row record from a South Korean steel plant [41]. The first 60% of the chronologically sorted source is reserved for development, the next 20% for selection, and the last 20% for an external protected audit. The fixed panel adds TabDDPM and TabSyn to CTGAN, TVAE, the Gaussian copula, and two transparent controls [34, 35]. Learned families receive the same one-hour accelerator budget at training sizes 2,000, 8,000, and 20,000 for three fixed seeds. Timeouts and incompatibilities must be reported. Because the source is time ordered, a row-iid certificate is prohibited. At manuscript freeze, no audit outcome had been returned; this is a pre-specified external challenge, not a completed result.

7 Results

The studies are ordered by what each one demonstrates, not by data source: a locked utility frontier, then generator breadth on a record-disjoint cohort, then external-domain control calibration and the fixed-budget learned-model cases, then scale and seed robustness, then empirical privacy risk, and finally the governance gate. The ACS and UCI families are therefore interleaved by purpose, not grouped by origin.

Reading the decisions. Every reported pass is an equivalence certificate for the named loss, tolerance, population, and frozen workflow under the stated audit assumption. Every reported failure or hold means that this certificate was not obtained; it is not, by itself, a test establishing that the true gap exceeds tolerance. Negative controls are expected to be held, and a negative control that certifies equivalence invalidates the case's calibration role.

7.1 California: a locked utility frontier

The selection Brier losses were 0.2502 for the independent-marginal control, 0.1703 for the conditional non-private comparator, and 0.1701, 0.1708, and 0.1702 for DP budgets 0.5, 1, and 4, respectively; the real-development reference was 0.1562. At the final 16,000-row scheduled audit look, the independent-marginal negative control failed the five-analysis grid. The conditional non-private comparator and all three DP candidates passed. Table 3 gives the complete final-look summary, Figure 8 separates selection from audit outcomes, and Figure 9 shows the fixed-budget selection comparison. This is the intended behavior of the audit: it rejects a dependence-destroying baseline and reports a bounded comparison for a declared use. It is not evidence that the passing candidates are safe to release.

Table 3: California final-look summary. "Worst upper gap" is the largest $|\widehat{\Delta}| + h$ among the five declared checks; direct comparisons across analyses should still use each analysis's own tolerance.

Candidate	Selection Brier	All five checks pass?	Worst upper gap
Independent marginal, non-private	0.2502	No	0.1471
Target conditional, non-private	0.1703	Yes	0.0621
DP conditional histogram, $\varepsilon = 0.5$	0.1701	Yes	0.0622
DP conditional histogram, $\varepsilon = 1$	0.1708	Yes	0.0624
DP conditional histogram, $\varepsilon = 4$	0.1702	Yes	0.0621

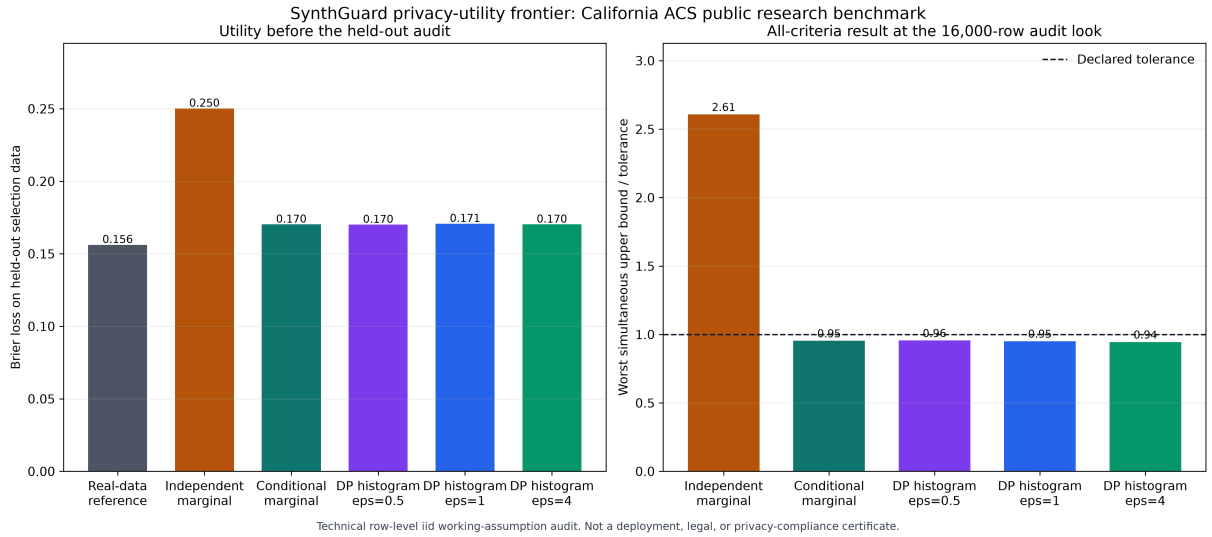


Figure 8: California locked-audit overview. The negative control fails while the target-conditional and DP candidates pass the declared public technical checks at the final scheduled look. This figure is a use-specific diagnostic under the study assumptions, not a release authorization.

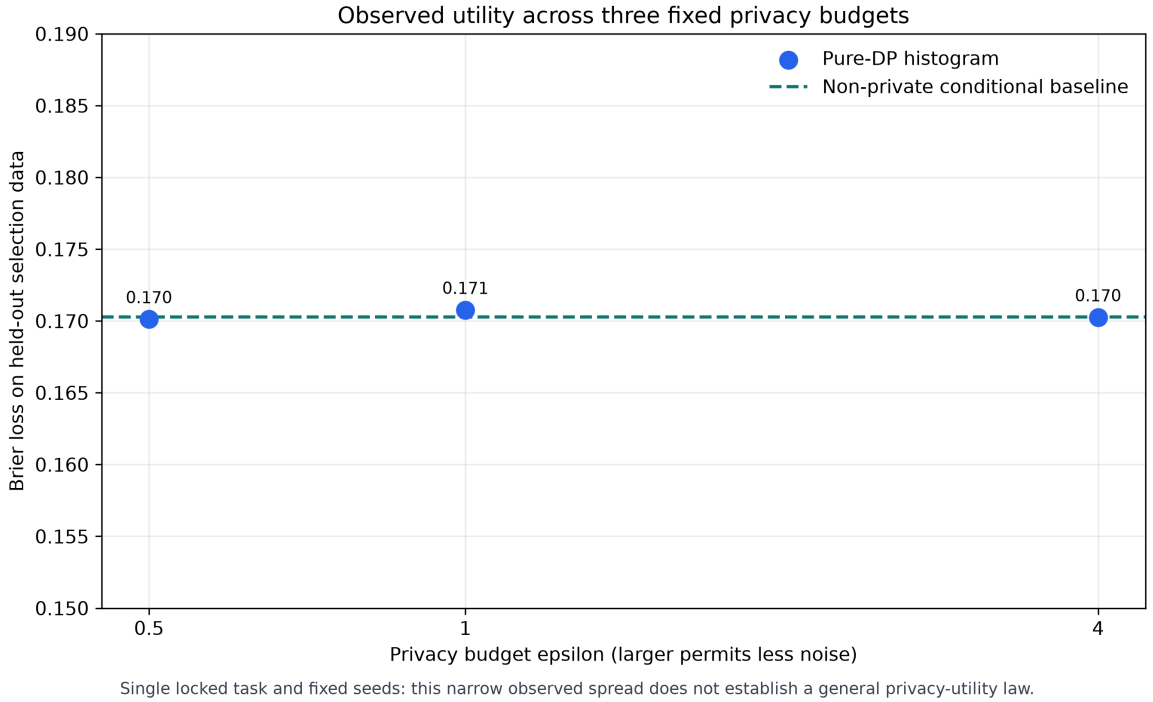


Figure 9: California selection utility across the fixed DP budgets. The plot is descriptive for this study and is not a general law relating ε to utility.

7.2 California record-disjoint extension: adding a dependency-aware comparator

The extension tests a limited but important alternative to featurewise marginal synthesis. The source-row-overlap check between the extension and the original California frontier returned zero. The Gaussian-copula comparator had the lowest pre-audit selection Brier loss (0.1637), compared with 0.1726–0.1734 for the three DP histograms, 0.1728 for target-conditional marginals, and 0.2514 for the independence-breaking control. At the final 16,000-row scheduled look, the copula and every conditional candidate passed all five technical checks; the negative control failed. The selected copula's worst upper gap was 0.0544. Table 4 reports the full candidate grid, and Figure 10 shows the separate selection and audit stages. This is evidence that the declared audit can evaluate an additional transparent dependency model in this public workflow. It is not evidence that a Gaussian copula is generally preferable, and it makes no privacy claim.

Table 4: California record-disjoint generator-breadth extension. Selection occurs before audit. The final technical decision is separate from the blocked release-governance decision.

Candidate	Selection Brier	All pass?	Worst upper gap
Independent marginal, non-private	0.2514	No	0.1436
Target conditional marginal, non-private	0.1728	Yes	0.0627
Target conditional Gaussian copula, non-private	0.1637	Yes	0.0544
DP conditional histogram, $\varepsilon = 0.5$	0.1734	Yes	0.0638
DP conditional histogram, $\varepsilon = 1$	0.1726	Yes	0.0624
DP conditional histogram, $\varepsilon = 4$	0.1728	Yes	0.0629

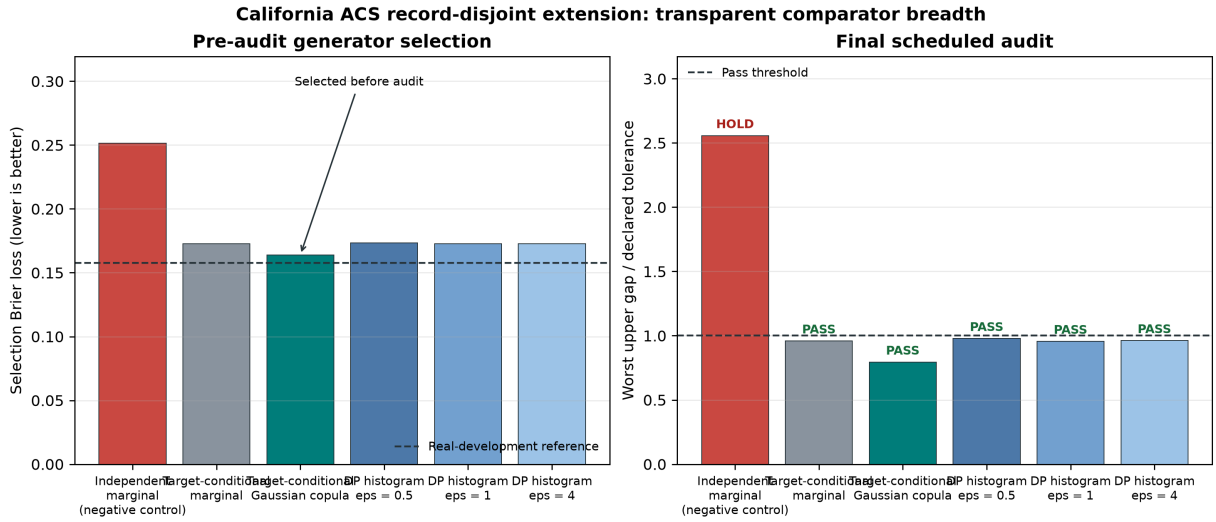


Figure 10: Record-disjoint California extension. The left panel fixes the selected candidate using selection data only; the right panel reports the final simultaneous audit grid. A bar below the threshold passed the declared technical checks. The plot does not turn the non-private copula comparator into a privacy result or a release approval.

7.3 Non-ACS external-domain control calibration

The external-domain record is intentionally mixed, not uniformly favorable. The UCI Bank Marketing pilot used a fixed three-candidate, five-analysis panel. At the final 9,000-record look, its independent-marginal negative control passed every declared check, including an overall-Brier upper gap of 0.0635 against a tolerance of 0.075. That outcome means the pilot's thresholds were not sufficiently discriminating for the intended control role. We retain the complete record as `EXCLUDED_NEGATIVE_CONTROL_PASSED`; we do not count it as favorable external-domain evidence or revise its tolerances after seeing the audit.

The UCI Default of Credit Card Clients case is a separate protocol fixed before its protected audit. It uses the same transparent three-candidate family, one primary Brier-loss check with tolerance 0.055, a 15,000 / 6,000 / 9,000 development-selection-audit split, and a control-rejection rule. The source sex field is excluded. The copula won selection Brier (0.1485; target-conditional marginal 0.1653; independent marginal 0.1739; real-development reference 0.1387). At the final scheduled look, Brier upper gaps were 0.0762, 0.0678, and 0.0528, respectively, so only the preselected copula passed. Figure 11 places this control-sensitive outcome beside the excluded Bank Marketing calibration. This is a technical result, not a credit-risk model or a release approval.

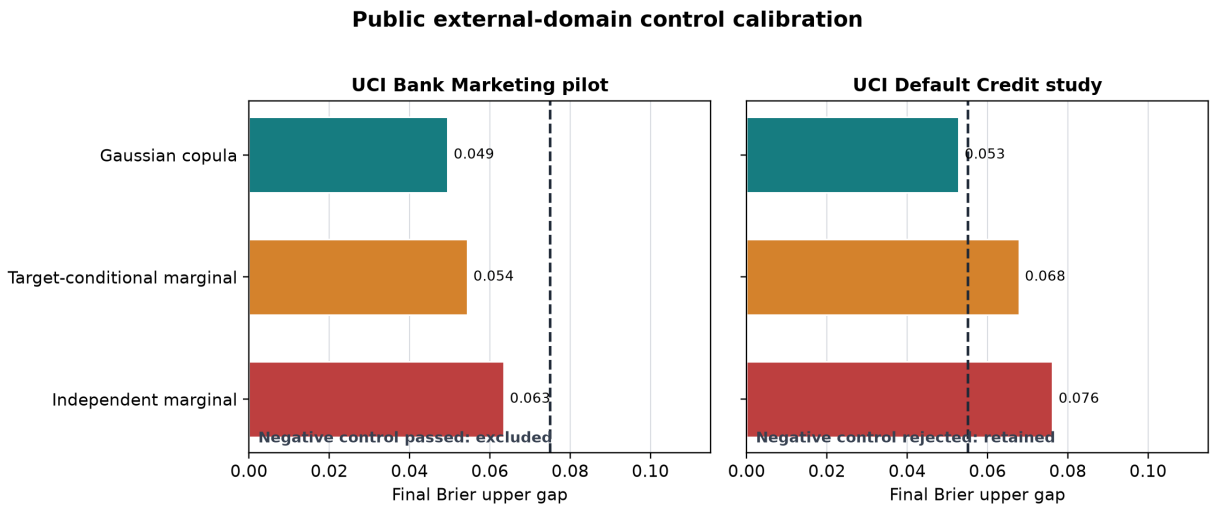


Figure 11: Public external-domain control calibration. Each bar is the final scheduled-look upper bound for the overall Brier gap; the dashed line is that case's locked tolerance. The Bank Marketing pilot is retained but excluded because its negative control passed. In the separately locked Default Credit study, both marginal comparators fail and the preselected non-private Gaussian copula passes. Neither case supports a credit, privacy, fairness, legal, or deployment claim.

7.4 Fresh UCI Online Shoppers fixed-budget learned-generator stress test

The Online Shoppers case tests the learned-baseline slot in a way that does not hide its compute budget. All four candidates were fit from the same fixed 1,000-session development subsample. Candidate artifacts and hashes were recorded before the selection and audit files were read. On the 1,330-session selection partition, Brier losses were 0.1413 for independent marginals, 0.1413 for target-conditional marginals, 0.1133 for the Gaussian copula, and 0.3184 for compact CTGAN, compared with 0.0947 for the real-development reference. The Gaussian copula was therefore selected before the final audit.

At the final 6,000-session scheduled look, Brier upper gaps were 0.1001, 0.0892, 0.0676, and 0.2842 for the same four candidates, respectively, against the locked tolerance 0.075. The independent negative control, target-conditional marginal, and compact CTGAN failed; only the Gaussian copula passed. CTGAN fit and sampled on CPU in 31.6 seconds. Figure 12 shows the pre-audit selection and final audit comparison. This is useful negative evidence: under this declared compact budget and this public e-commerce table, the learned generator did not improve the audit result. It is not a claim that CTGAN fails under every training budget, data set, or implementation. The public run's release packet remains blocked because the human integrity attestation is not asserted.

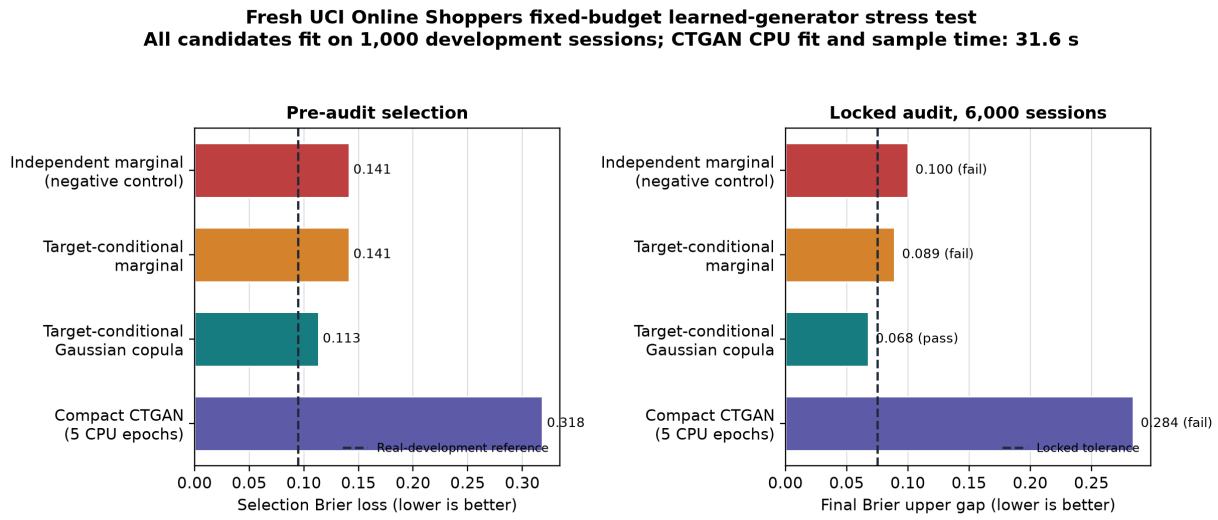


Figure 12: Fresh UCI Online Shoppers fixed-budget learned-generator stress test. All candidates fit on the same 1,000-session development subsample. The left panel is the pre-audit selection comparison; the right panel is the final simultaneous Brier upper bound at 6,000 audit sessions. The compact CTGAN and the independence-breaking control fail the locked technical check. This is a compute-limited research measurement, not a general neural-model ranking or a release authorization.

7.5 UCI Adult: the pre-registered five-role challenge panel

The Adult case is the first study to run the challenge protocol's complete candidate panel, with CTGAN and TVAE trained under the identical CPU budget. On the 8,000-record selection partition, the Brier losses were 0.1750 for the independent marginal, 0.1254 for the target-conditional marginal, 0.1206 for the Gaussian copula, 0.2323 for the small CTGAN, and 0.2333 for the small TVAE, compared with 0.1084 for the real-development reference. The Gaussian copula had the best selection score, so it was chosen before the final audit.

At the final 16,000-record audit look, the simultaneous Brier upper gaps were 0.0959, 0.0482, 0.0457, 0.1556, and 0.1518 for the same five candidates, against the locked tolerance of 0.060 (Figure 13). The negative control failed, which the certificate requires. The target-conditional marginal and the selected Gaussian copula passed. Both small learned models failed, with estimated gaps of +0.124 for CTGAN and +0.120 for TVAE. The two learned models land close to each other and clearly apart from the simple baselines, which is exactly the comparison the matched-slot rule is meant to show. On CPU, CTGAN took 58.5 seconds to fit and sample; TVAE took 4.4 seconds. As with the Online Shoppers test, this is negative evidence under one small declared budget. It does not say that either model fails with more training, other data, or other implementations. The case is control-sensitive and its release packet stays blocked until a person makes the required integrity attestations.

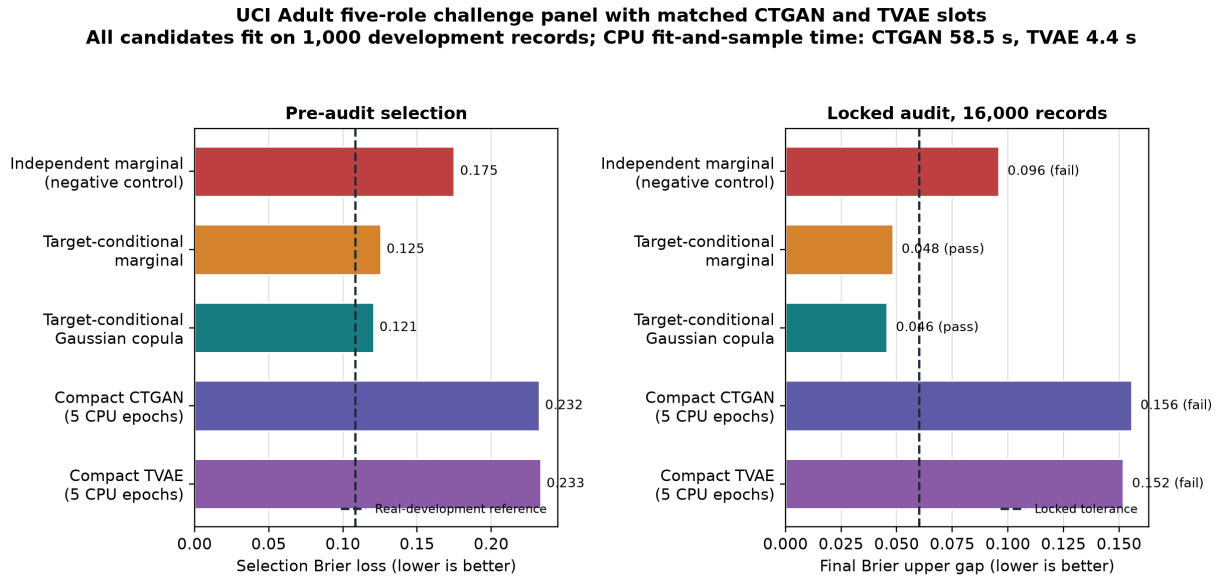


Figure 13: UCI Adult five-role challenge panel with matched CTGAN and TVAE slots. Every candidate was fit on the same 1,000-record development subsample under the shared five-epoch CPU budget. Left: the selection comparison before the audit. Right: the final simultaneous Brier upper bound at 16,000 audit records. The negative control and both small learned models fail the locked check; the simple conditional baselines pass. This is a small-budget research measurement, not a model ranking or permission to release data.

Post-audit budget sensitivity. The larger training budget improves both learned-model measurements without reversing the main comparison. CTGAN's selection Brier loss falls from 0.2323 to 0.1846 and its final nominal upper gap falls from 0.1556 to 0.1068. TVAE's selection Brier loss falls from 0.2333 to 0.1669 and its final nominal upper gap falls from 0.1518 to 0.0834. Both final values remain above 0.060. Table 5 summarizes the compact and 300-epoch measurements. The comparison answers a limited concern about the five-epoch stress test: the compact budget explains part, but not all, of the observed gap on this 1,000-record fit sample. It does not repair the already-opened audit or support a general CTGAN-versus-TVAE ranking.

Post-audit data-scaling arm. Budget alone does not settle the question, because every measurement so far used a 1,000-record fit sample. We therefore held the learned budget at its library default and refit the whole five-role panel at three development sizes, 1,000, 4,000, and 16,000 records, scoring each with the same fixed workflow, audit rows, schedule, and tolerance. The result is a clear crossover (Figure 14 and Table 6). At 1,000 records the Gaussian copula has the smallest final nominal upper gap, 0.0394, while CTGAN reaches 0.0833 and TVAE 0.1093, so both learned models fail the 0.060 tolerance. At 4,000 records CTGAN falls to 0.0556 and passes, while TVAE at 0.0932 still fails. At 16,000 records both learned models pass, at 0.0476 for CTGAN and 0.0461 for TVAE against 0.0519 for the copula. The independent-marginal negative control fails at every size, with gaps of 0.1137, 0.1091, and 0.1129, so the audit remains correctly calibrated across the whole sweep. Section 7.6 repeats the sweep at three generation seeds and reports which parts of this pattern survive; the apparent ordering of the learned models against the copula at 16,000 records does not.

Two things follow, and we state both. First, the compact-budget failure reported above is

Table 5: UCI Adult learned-model compute sensitivity. The 300-epoch arm is post-audit and nominal-only.

Budget	Candidate	Selection Brier	Final upper gap
Locked, 5 epochs	CTGAN	0.2323	0.1556
Locked, 5 epochs	TVAE	0.2333	0.1518
Post-audit, 300 epochs	CTGAN	0.1846	0.1068
Post-audit, 300 epochs	TVAE	0.1669	0.0834

mainly a small-sample effect rather than a general verdict on either architecture: give these models enough data and the same locked criterion admits them. Second, and more important for the framework, this is the behavior a useful audit should have. A criterion that could never pass a learned generator would be a broken instrument. The same fixed rule that rejects an intentionally dependence-destroying control at every sample size, and that rejects small-sample neural fits, also accepts those fits once the evidence supports them. This arm is post-audit and its bounds are nominal-only; it cannot enter the locked certificate panel, and it is a descriptive scaling measurement on one public table, not a general sample-size threshold for synthesis.

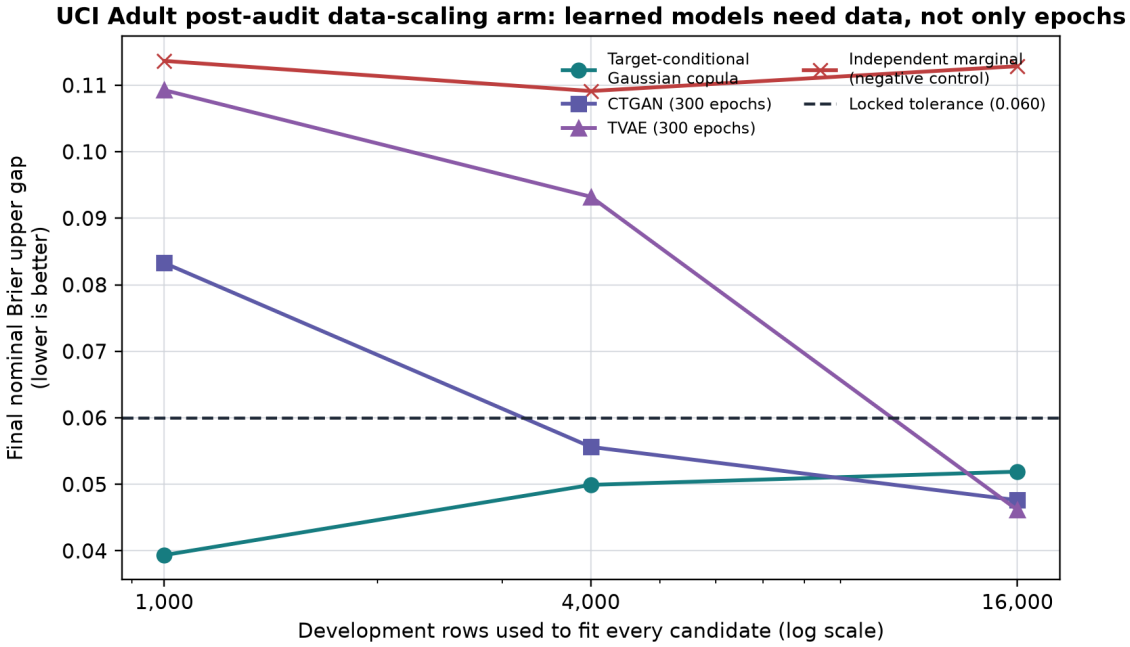


Figure 14: UCI Adult post-audit data-scaling arm. Every candidate is refit at each development size with the learned budget held at its library default; bounds are nominal-only because the audit had already been inspected. The transparent copula wins at 1,000 records, CTGAN crosses the locked tolerance by 4,000, and both learned models pass and overtake the copula by 16,000 records. The negative control fails at every size, so the criterion stays calibrated throughout.

Table 6: UCI Adult post-audit data-scaling arm: final nominal Brier upper gaps against the locked 0.060 tolerance. Pass or fail is nominal only; this arm makes no certificate claim.

Candidate	1,000 rows	4,000 rows	16,000 rows
Independent marginal (control)	0.1137 fail	0.1091 fail	0.1129 fail
Target-conditional Gaussian copula	0.0394 pass	0.0499 pass	0.0519 pass
CTGAN, 300 epochs	0.0833 fail	0.0556 pass	0.0476 pass
TVAE, 300 epochs	0.1093 fail	0.0932 fail	0.0461 pass

7.6 Which parts of the crossover survive independent seeds

A crossover read from one generation seed per cell cannot be distinguished from seed noise. We therefore repeated the whole sweep at three independent generation seeds for every training size, fixing the question in advance: do both learned models fail unanimously at 1,000 records, pass unanimously at 16,000, and beat the copula at 16,000 under every seed?

The first two hold and the third does not. At 1,000 records both learned models fail under all three seeds, with mean upper gaps of 0.0902 for CTGAN and 0.1158 for TVAE. At 16,000 records both pass under all three seeds, at means of 0.0491 and 0.0430. The negative control fails under every seed at every size, means 0.1192, 0.1109, and 0.1111, so the calibration property is robust. But the apparent ordering against the copula at 16,000 records is not: TVAE beats the copula in all three seeds, while CTGAN does so in only two, reaching 0.0556 against the copula's 0.0521 under one seed. The single-seed reading that both learned models overtake the copula was therefore an artifact, and we withdraw it.

Dispersion is itself informative and varies by cell (table 7). The copula is highly stable, with a standard deviation of 0.0008 to 0.0022 across sizes, because its fitting procedure is nearly deterministic given the sample. The learned models are far noisier, and most so where the evidence is thinnest: TVAE has a standard deviation of 0.0202 at 1,000 records and 0.0184 at 4,000, against 0.0027 at 16,000. That intermediate cell is where a single seed misleads most, and indeed TVAE passes in only one of three seeds at 4,000 records, where the single-seed run reported a clean failure.

The point carries beyond this table. A locked panel fixes the candidate list, the analyses, the tolerances, and the audit schedule, but a stochastic generator still has a seed, and the certificate as stated conditions on the candidate artifact, not on the training procedure that produced it. Reporting one seed per learned candidate therefore understates uncertainty about the procedure, even though it correctly describes the artifact that was locked. We recommend that future locked panels declare a seed budget per learned candidate and report the resulting dispersion alongside the certificate, and we treat the single-seed learned results elsewhere in this paper as evidence about specific artifacts, not about the generators that produced them.

Table 7: Seed stability of the post-audit scaling arm: mean final nominal upper gap across three generation seeds, with the number of seeds passing the locked 0.060 tolerance. Bounds are nominal and no certificate is claimed.

Candidate	1,000 rows	4,000 rows	16,000 rows
Independent marginal (control)	0.1192 (0/3)	0.1109 (0/3)	0.1111 (0/3)
Target-conditional Gaussian copula	0.0391 (3/3)	0.0503 (3/3)	0.0522 (3/3)
CTGAN, 300 epochs	0.0902 (0/3)	0.0534 (3/3)	0.0491 (3/3)
TVAE, 300 epochs	0.1158 (0/3)	0.0718 (1/3)	0.0430 (3/3)

7.7 UCI Diabetes 130-US Hospitals: an excluded health calibration

The Diabetes 130-US Hospitals study deliberately records a calibration failure rather than trying to rescue it after the audit. On the 2,000-record selection partition, Brier losses were 0.0788 for the independent marginal, 0.0795 for the target-conditional marginal, 0.0764 for the Gaussian copula, 0.2478 for compact CTGAN, and 0.2127 for compact TVAE, compared with 0.0746 for the real-development reference. The Gaussian copula was selected before audit. At the final 8,000-record look, simultaneous Brier upper gaps were 0.0475, 0.0517, 0.0453, 0.2093, and 0.1782 for the same candidates, against the locked tolerance of 0.080 (Figure 15). CTGAN and TVAE both fail under this matched compact budget. For the benchmark, what matters more is that the independent-marginal negative control also passes. The threshold therefore does not discriminate the declared dependence-destroying control in this case. We retain the complete locked record as `EXCLUDED\_NEGATIVE\_CONTROL\_PASSED`, do not retune the tolerance, and do not count the case as favorable health-domain or clinical evidence.

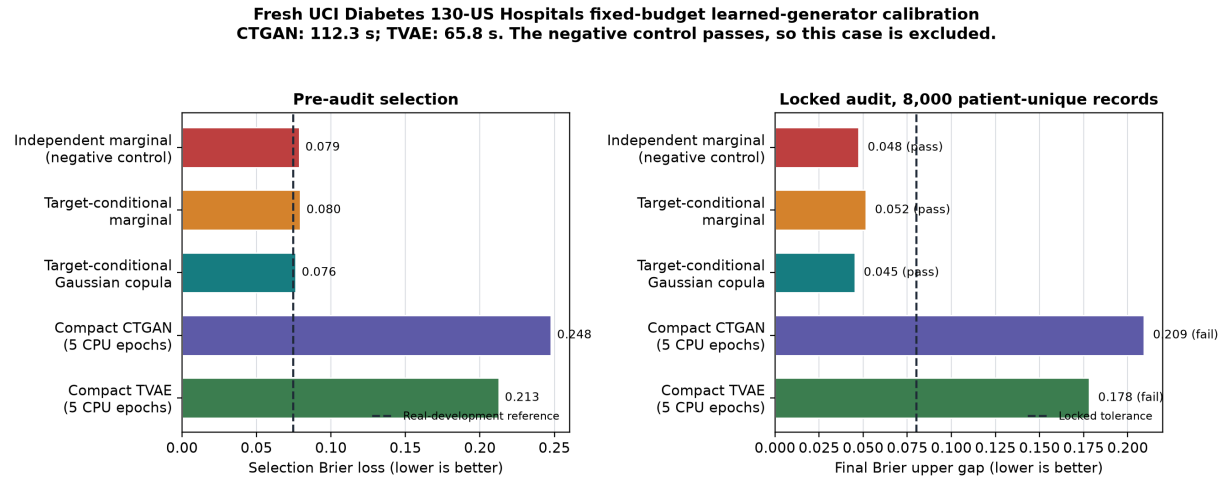


Figure 15: UCI Diabetes 130-US Hospitals patient-unique health calibration. Every candidate was fit on the same 1,000-record development subsample under the locked five-epoch CPU budget. CTGAN and TVAE fail the final Brier check, but so does the case itself as a discriminator: the independent-marginal negative control remains below the locked tolerance. The entire record is consequently retained as an excluded threshold-calibration failure, not as a health, clinical, privacy, or deployment result.

7.8 New York: external state, scale, and seed sensitivity

The New York study is where the full-grid design earns its keep. Across the 33 locked candidates, 26 passed and 7 failed. All three independent-marginal controls failed. At the smallest 2,500-row training size, only one of three $\varepsilon = 0.5$ runs and one of three $\varepsilon = 1$ runs passed; all three $\varepsilon = 4$ runs passed. At 10,000 and 50,000 training rows, all DP replicates at all three budgets passed the declared grid. The pre-specified rare-group analysis had 3,771 audit rows. Figure 16 shows the scale and rare-group results, while Figure 17 displays every pass and hold in the 33-candidate grid. These results show that the observed behavior depends on data scale and seed in this public configuration. They do not establish a universal sample-size threshold for DP synthesis.

7.9 Florida: a risk diagnostic must detect a known leak

The Florida RiskLab separates two questions: whether an attack battery reacts to an intentional leak, and what its result says about the remaining candidates under that named attack. The row-resampling positive control reached membership AUC 0.782 (95% CI 0.777–0.787) and evaluation advantage 0.564. The other five candidates had AUCs from 0.502 to 0.504, with intervals that crossed 0.5. The observed attribute-prediction accuracy gaps ranged from -0.0087 to -0.0005. Table 8 reports the numerical diagnostics; Figures 18 and 19 show the membership and attribute results. These findings validate the sensitivity of this particular attack battery to the positive control. They do not demonstrate attack completeness, protection against a stronger attacker, or a general privacy guarantee.

7.10 One attack family is a weak instrument

A single attack cannot distinguish a release that leaks little from an attack that is badly matched to the release. We therefore re-scored the same locked Florida artifacts, verified by hash, under

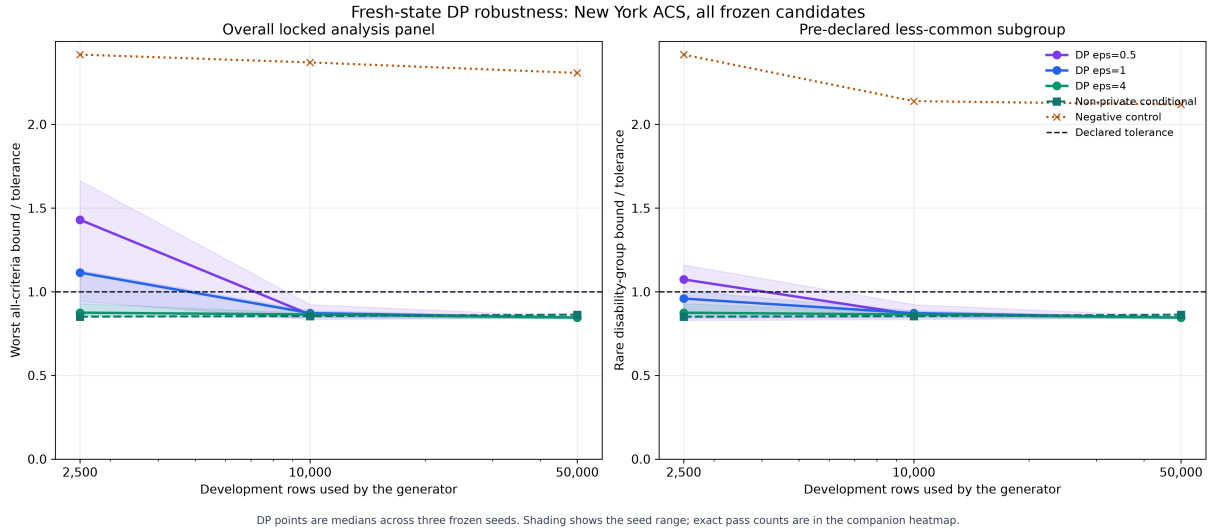


Figure 16: New York robustness results across fixed training scales, privacy budgets, and seeds. The low-scale failures are reported as part of the main result, not excluded after model selection.

Table 8: Florida held-out risk diagnostics. The attack is a specific full-record nearest-neighbor membership test with held-out calibration and evaluation.

Candidate	Membership AUC (95% CI)	Evaluation advantage	Attribute accuracy gap
Row-resample positive control	0.782 (0.777, 0.787)	0.5640	−0.0087
Independent marginal	0.504 (0.497, 0.512)	0.0057	−0.0005
Target conditional	0.502 (0.494, 0.510)	−0.0009	−0.0039
DP conditional, $\varepsilon = 0.5$	0.504 (0.497, 0.512)	0.0020	−0.0053
DP conditional, $\varepsilon = 1$	0.503 (0.496, 0.510)	0.0009	−0.0077
DP conditional, $\varepsilon = 4$	0.502 (0.495, 0.508)	0.0023	−0.0037

three structurally different instruments. Two are membership attacks proper, in that they predict whether a candidate real record was in the generator's training set: the locked nearest-neighbour distance, and a distance-ratio attack scoring isolation, that is how much closer the nearest synthetic record is than the next, which is invariant to the global rescaling that would flatter a raw-distance attack. The third is deliberately not a membership attack. It is a similarity discriminator trained to separate synthetic records from held-out reference records, and we label it honestly: distinguishing synthetic from real rows is principally a two-sample fidelity test, and it bears on membership only insofar as a generator memorises its training rows. All three use the locked calibration discipline, choosing thresholds on the calibration partitions and scoring the evaluation partitions once. This is a post-audit diagnostic: the locked membership numbers are unchanged and no release decision is revised. Reported AUCs are point estimates on the fixed evaluation partitions; the locked Florida record carries the bootstrap interval for the headline attack, and we do not attach intervals to the added families here.

The nearest-neighbour family reproduces the locked positive-control AUC of 0.7824 exactly, which confirms the arm scores the same released artifacts. The distance-ratio family independently detects the intentional row-resampling control at 0.7710, against at most 0.5023 for every other candidate. The similarity discriminator does not: it reaches only 0.5381 on the deliberately leaky control, against at most 0.5056 elsewhere (figure 20). It orders the control first, but it is close

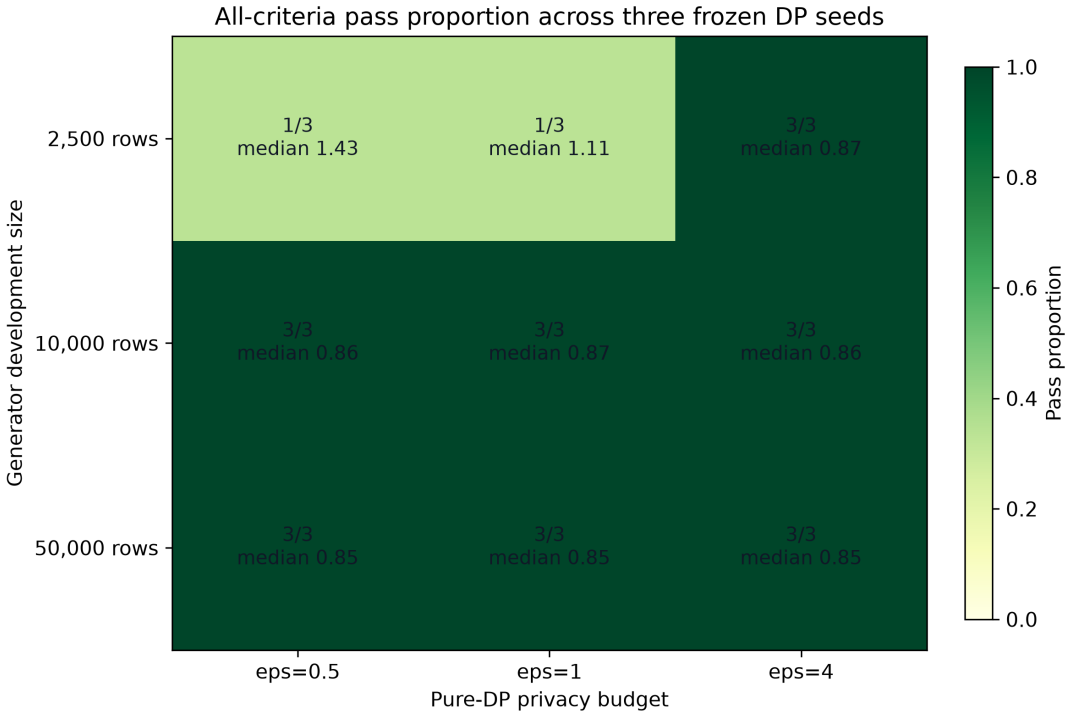


Figure 17: Pass/hold heatmap for the complete 33-candidate New York grid. A pass means all declared technical checks passed in this study; a hold is not a claim that the method cannot be useful elsewhere.

enough to chance that we do not treat it as a detector for this release.

We report that failure rather than dropping the family, and it is the most useful observation in this subsection. The result is what the labelling predicts: a two-sample test asks whether synthetic rows look like real rows in aggregate, which is a different question from whether a particular record trained the generator, and it answers the membership question only weakly even against a memorising control. Had it been the only instrument in the battery, every candidate including that control would have scored near chance and the release would have looked uniformly safe. The positive control is what exposes the instrument as uninformative. A low score under a named attack is therefore evidence about that attack, not about the release, unless the same attack is shown to detect a leak that is known to exist.

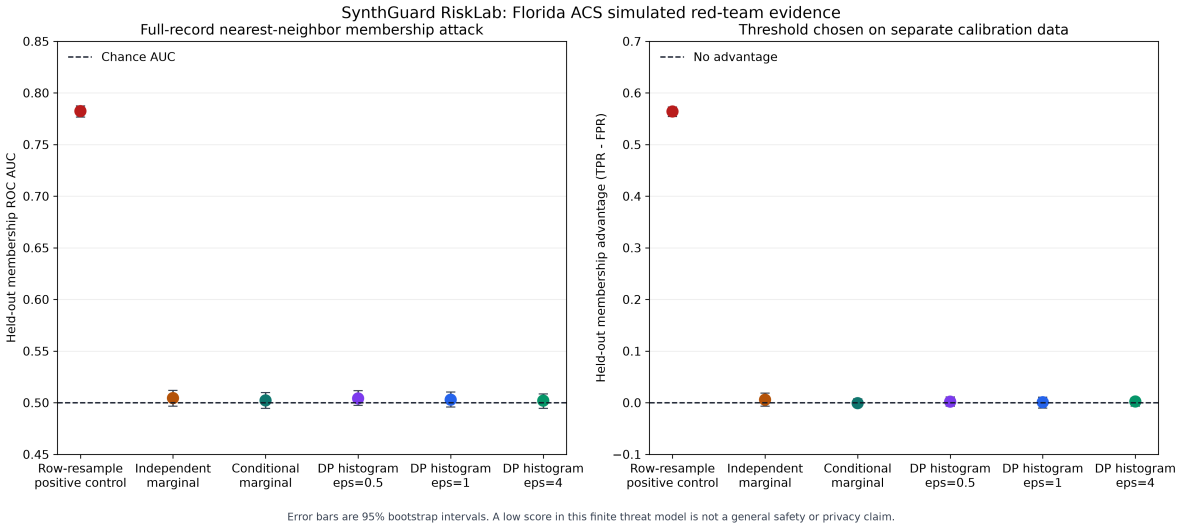


Figure 18: Florida RiskLab membership diagnostic. The intentional row-resampling control is visibly separated from the remaining candidates under this exact attacker and encoding. This is a calibration check for the attack battery, not a general privacy result.

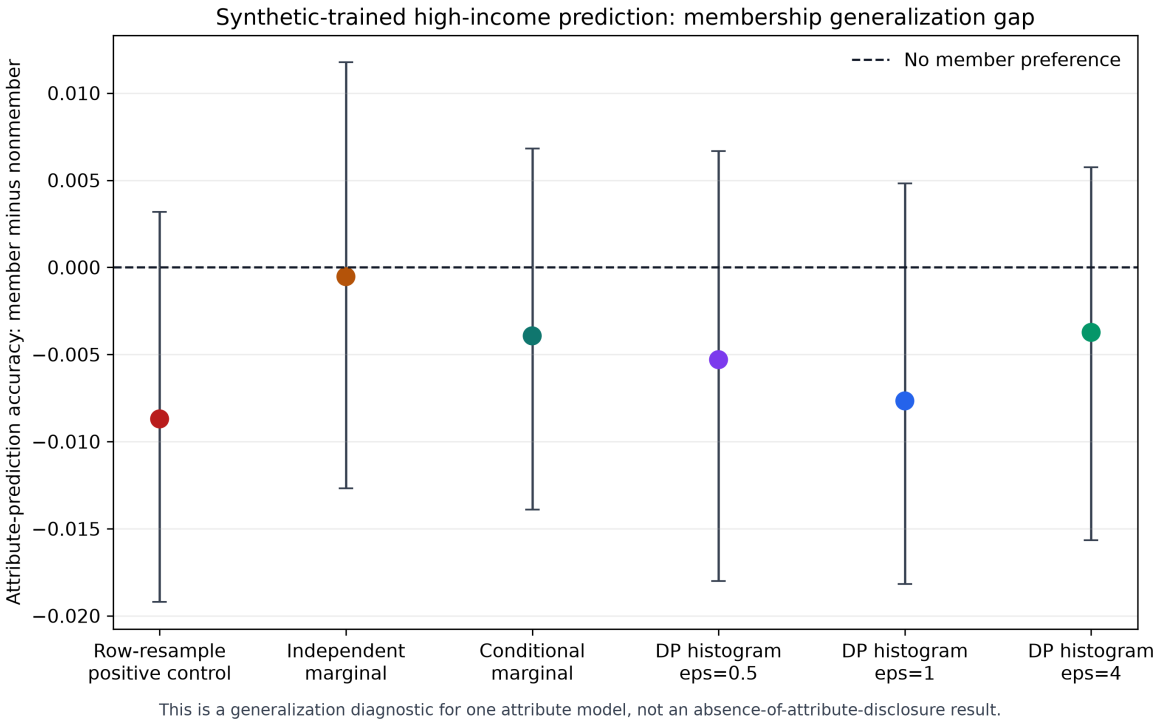


Figure 19: Florida attribute-prediction generalization-gap diagnostic. The plot is reported alongside membership evidence; neither metric alone is treated as a universal privacy score.

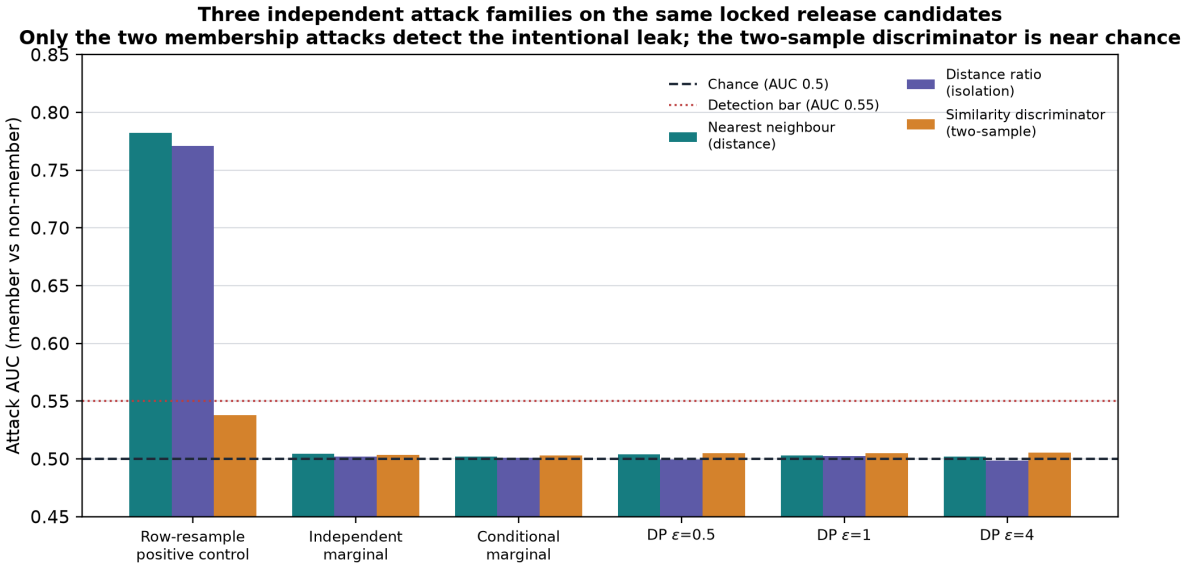


Figure 20: Three independent membership-attack families applied to the same locked Florida candidates. The two distance-based families separate the intentional row-resampling control from every other candidate; the similarity discriminator, which is a two-sample fidelity test rather than a membership attack, stays near chance even on that control and is therefore reported as uninformative for this release rather than as evidence of safety.

7.11 A sealed evaluator can block a release despite passing utility checks

The sealed-audit prototype accepts a local real training table, audit table, and synthetic table, and emits only aggregate JSON outputs. Its public demonstration uses the California data solely to make the implementation inspectable. All declared technical checks pass in that demo, but the release gate returns `BLOCKED`. The reasons are explicit: the required attestations for audit independence, candidate-panel locking, analysis-panel locking, and audit sampling documentation are false, and the manifest remains in draft status. No raw data, row identifiers, paths, model coefficients, or synthetic records are exported.

This is not an administrative afterthought. It is the mechanism by which a technical evaluator refuses to transform a reproducible public experiment into a fictional release approval. In a real deployment, the same code can produce a decision only after the organization supplies truthful evidence about the data lifecycle and audit separation.

8 Supplementary Multi-Domain Temporal Robustness Tracks

The primary certificate in proposition 1 is not applied to the following public tracks. They were added to test whether the public use visibly distinguishes a structure-preserving baseline from a dependence-destroying negative control when the data have time, location, or operational dependence. Both tracks use strictly earlier development and selection windows and a later audit window. Table 9 reports both chronological comparisons. They are descriptive robustness evidence, not official v1 certificate cases.

The BRFSS track uses public state-year prevalence records from 2019–2024 [2]. It predicts a reported prevalence value using year, reporting area, question, and breakout fields. The selected row-bootstrap candidate was close to the real-development reference on the held-out 2024 window, while the independent-marginal negative control had much higher normalized absolute error. The rows share state, question, subgroup, and time structure. Therefore, the record is marked `NOT_ELIGIBLE_FOR_V1_IID_CERTIFICATE`.

The Chicago track uses public 2024 taxi-trip records [3]. It predicts whether a recorded tip is at least 20 percent of fare from post-trip operational fields, after excluding trip identifiers, taxi identifiers, and coordinates. The bootstrap baseline was again close to the real-development reference on the March audit window, while the independent-marginal control failed visibly. The records retain time, company, spatial, and unobserved vehicle dependence, so this result is also marked `NOT_ELIGIBLE_FOR_V1_IID_CERTIFICATE`.

These tracks broaden the public engineering testbed to public health and operations, but they do not broaden the primary scientific claim. They compare only an intentionally unsafe row bootstrap with a negative control, not a panel of model-based synthesizers. Their purpose is to keep a temporal or clustered case visible in the benchmark while the required cluster-aware or time-series certificate is developed and independently checked.

Multi-domain challenge status. The public challenge registry now names seven non-ACS records across financial services, consumer finance, e-commerce, population health, transport operations, and census income microdata. Their evidence tiers are deliberately unequal: the Default Credit, Online Shoppers, and Adult records are reproducible L1 row-level technical audits; the Bank Marketing and Diabetes 130-US Hospitals records are kept visible as excluded calibration failures; and the BRFSS and Chicago records are time-based diagnostics, not row-IID certificate cases. The pre-specified learned-model panel rule asks for CTGAN and TVAE under the same CPU-only five-epoch, batch-500 budget, next to the negative control, the conditional marginal baseline, and

Table 9: Supplementary public temporal robustness tracks. Losses are shown as reference / row bootstrap / independent marginal; lower is better. The row-bootstrap candidate reuses original records and makes no privacy claim. Neither track receives the row-independent certificate.

Track and audit unit	Split sizes (development / selection / audit)	Audit losses: reference / bootstrap / independent marginal
CDC BRFSS prevalence, calendar year	20,222 / 5,134 / 4,641	0.03781 / 0.03788 / 0.18651
Chicago taxi operations, calendar window	19,284 / 15,399 / 14,641	0.10192 / 0.10266 / 0.23703

the statistical baseline. It has now been run once, on the newly locked Adult case (Section 7.5). The separate Diabetes run uses the same matched learned slots but is excluded because its negative control passed. The 300-epoch Adult sensitivity is a post-audit robustness record and does not change the locked panel or its evidence tier. These are the authors' own L1 records and do not count toward the external milestone. No after-the-fact TVAE comparison is added to the already-locked Online Shoppers study.

9 What the Benchmark Does Not Show

The paper's strongest claims are deliberately narrow. First, the completed row-level technical studies cover one public survey family and five public UCI records; the retained Bank Marketing and Diabetes 130-US Hospitals cases remain excluded calibration failures. The supplementary health and transport records make the challenge registry genuinely multi-domain, but they are not pooled with the row-IID certificate. Second, the public ACS and UCI demonstrations use a row- or session-level iid technical diagnostic. They do not validate survey-weighted inference, household-aware splitting, temporal transfer, a confidential release setting, or real-world credit, advertising, clinical, or consumer-targeting decisions. Third, the reported generators are simple marginals, a regularized Gaussian copula, exact-DP histograms, and fixed-budget learned-model tests (CTGAN on Online Shoppers; CTGAN and TVAE together on Adult and Diabetes). This is not a claim to cover all modern synthesizers. The 300-epoch Adult sensitivity narrows the learned-model gaps, and the post-audit data-scaling arm shows both learned models passing the same locked tolerance by 16,000 fit records. Both arms reuse a previously inspected audit, carry nominal-only bounds, and make no certificate claim. A three-seed repetition (section 7.6) confirms the pass-fail crossover and the control's failure at every size, but withdraws the single-seed reading that both learned models overtake the copula: CTGAN does so in only two of three seeds. They establish that the compact-budget failures are largely a small-sample effect on this table; they do not rank either architecture in general, and the crossover is one measurement on one public data set, not a transferable sample-size threshold for synthesis. More broadly, learned candidates elsewhere in this paper are reported from a single generation seed, so they are evidence about the specific locked artifact, not about how the generator behaves on average. No after-the-fact TVAE comparison is added to any earlier locked study. Fourth, the Florida attacks are named empirical diagnostics. A low score under those diagnostics is not proof of non-identifiability or protection against an unmodeled attacker. The battery now spans three instruments instead of one, but section 7.10 shows the limit of that defence directly: the third is a two-sample fidelity test rather than a membership attack, and it sits near chance even on a deliberately leaky control, so breadth helps only when each instrument is separately shown to detect a known leak. The two genuine

membership attacks both work by proximity in a fixed encoding, and none of the three addresses linkage against external auxiliary data. Fifth, the finite-sample certificate requires a locked panel and a valid audit split; it does not repair model selection or p-hacking performed after audit access. Sixth, an independent-block certificate is available only when the sampling design justifies independent blocks. The temporal certificate of proposition 10 extends the framework to time-ordered audits, but it inherits a declared dependence horizon that no algorithm can verify from data, it changes the estimand to an equally weighted block mean while discarding buffered rows, and it is validated here only on a simulated regime process. It therefore does not retroactively certify the descriptive BRFSS and Chicago tracks, which were locked without a declared horizon. Seventh, the variance-adaptive and anytime instruments of section 4.3 are reported as a measurement on already-locked evidence; they revise no recorded outcome, and selecting a bound family after seeing which one passes is exactly the adaptation proposition 5 forbids. Finally, table 2 is a documented-scope comparison and the executable scenarios test workflow coverage; they are not an empirical ranking of the numerical accuracy, speed, or completeness of SDMetrics, SynthEval, or Anonymeter.

The repository therefore retains failures and invalidations. The New York low-scale failures are included in the main results. The prior CTGAN exploratory run is marked invalid because its claimed replication was not record-disjoint. The sealed demonstration is blocked despite its passing technical grid. These records are part of the contribution: an evaluation benchmark should make it harder, not easier, to hide unfavorable or ineligible evidence.

10 Reproducibility and Adoption

The public package contains locked configurations, source provenance, candidate metadata, synthetic artifacts, checksums, figures, implementation code, unit tests, and run guides for the two California studies, New York, Florida, five UCI technical records, the Adult compute sensitivity, and the two supplementary chronological tracks. The software is archived under CC BY 4.0 at Zenodo [42]. Version 0.1.14 accompanies this manuscript and carries DOI 10.5281/zenodo.21909680; the concept DOI 10.5281/zenodo.21503396 always resolves to the latest version. Anyone reproducing the results below should use the version DOI, because every number reported here is checked against the locked records in that exact version. It installs a `synthguard` command with audit planning, bounded-gap auditing, and fail-closed registry validation. A container, example configuration, Python 3.11/3.13 continuous-integration workflow, generator-adapter contract, and machine-readable public release registry are included. The test suite checks split integrity, scheduled row- and block-bound calculations, the variance-adaptive and anytime-valid bounds together with Monte Carlo coverage of each, the buffered-block temporal certificate, the reconstruction attack and a regression guard against the seeding collision that would trivially defeat it, the certification-margin planner, the audit-reuse boundary, generator manifests, the external challenge freeze, DP histogram composition, learned-baseline artifacts, attack calibration, release-gate behavior, and independent-rerun labels. The Online Shoppers example remains blocked because reproducibility does not replace accountable human release approval.

Two non-author rerun records for the immutable v0.1.11 archive have been received. They differ in how independently they can be verified, and we describe that openly instead of averaging over it. One record comes from a reviewer at a separate institution. It was produced on a different machine and a different Python version (3.12.13 rather than the locked 3.13); it selected the same Gaussian-copula candidate, its release packet stayed blocked, and it matched the locked aggregate result. It also honestly discloses one failed automated test caused by a numerical difference of about one part in a billion (0.284226616625272 against the locked 0.28422661477241873), which is the

Table 10: External evidence status at manuscript freeze. "Complete" refers only to the named artifact set, not to release approval.

Record	Status	Claim boundary
Non-author rerun A	Complete private attestation plus matched machine report	Same-institution relationship disclosed; release packet remained blocked.
Non-author rerun B	Qualified private attestation; machine JSON pending	Different institution; version deviation and failed numerical test retained; not exact conformance.
Steel-energy challenge	Frozen, unassigned, outcome-free	No external-validation result and no iid certificate claim.
Software DOI	Published	Version 0.1.14 archived at 10.5281/zenodo.21909680.

kind of tiny, environment-dependent difference a genuine independent run produces, not a sign of a problem. We treat this as one completed independent rerun; supplying its original machine-readable report and logs would strengthen it from a signed-form record to a fully machine-verified one. The second record comes from a reviewer at the lead author's own institution, which we disclose rather than obscure. It supplies a complete machine report produced on a verifiably different computer: a different processor model, a different host name, half as many cores as the authors' machine, and a third distinct interpreter, Python 3.14.2. All thirteen packet hashes matched, the same candidate was selected, and the release packet stayed blocked. It too discloses deviations under the newer interpreter, namely one failed automated test and two locked result files that differ numerically. Its signed form, however, left the environment and outcome fields blank, so the form is not yet self-contained on its own.

An earlier submission from a third person was withdrawn from the evidence base entirely. Its machine report could not be distinguished from a run on the authors' own computer, showing the same operating-system build, the same Python version, and no separate hardware signature. Rather than count it weakly, we park it: the files are retained unaltered for provenance and support no claim. The pre-registered milestone in the frozen challenge specification asks for two independent reruns of the Online Shoppers case from two distinct institutions. On its numeric terms that bar is now met: two non-author reruns have been received, from two distinct institutions, each produced on a verifiably distinct machine and each on the required case. We state that plainly because it is what the specification asked for and it was fixed before any record arrived.

We nevertheless decline to declare the milestone complete, for two reasons, both of which we would rather state than bury. First, each record still carries a completeness gap: one lacks its original machine-readable report, and the other has a signed form whose environment and outcome fields were left blank. Second, only one of the two institutions is external to the authors; the other is the lead author's own, which we disclose rather than let a reader infer independence we do not have. A fail-closed registry validator in the release enforces this refusal automatically, so the public status cannot drift upward without the missing evidence arriving. Reviewer names, signatures, and forms are kept in a private evidence archive and are not part of the public release.

Table 10 makes the external boundary explicit. A signed narrative record and a machine-readable conformance record are separate evidence objects; neither is silently inferred from the other.

For an institution, adoption should follow a simple sequence: define the permitted use and tolerances; establish a training/selection/audit separation at the appropriate person, household, site, or time unit; lock candidate and analysis panels; run the audit within the organization's controlled

environment; review the complete evidence packet; and only then consider a release decision under the organization's legal, privacy, and governance requirements. SynthGuard supplies technical evidence; it does not replace the organization responsible for the release.

Intake is not universal model selection. To make this workflow easier to start, the repository includes SynthGuard Scout, a local schema-and-constraint intake tool. Given a declared use card, privacy requirement, audit unit, target, identifier columns, and aggregate table profile, it returns a shortlist of candidate *evaluations*: an independent-marginal negative control, transparent conditional baselines, a regularized Gaussian copula where applicable, a finite-domain DP option where a public binning rule is viable, and a clearly marked slot for a fixed-budget learned baseline. It does not assert that a model is best for arbitrary data. The fresh Online Shoppers stress test illustrates why: its compact CTGAN is a valid candidate to evaluate but does not pass this fixed audit. One result is not a neural-generator ranking. For time- or cluster-dependent records, Scout returns `ROUTING_REQUIRED` and refuses to route the case to the row-iid certificate. Its output is a planning artifact; the subsequent candidate panel, workflow, tolerance panel, and audit schedule still have to be locked before protected-audit outcomes are observed.

The remaining external steps are specific: obtain the second rerunner's original JSON and logs or repeat that rerun in the locked environment; have an unrelated third institution run the frozen steel-energy challenge without author audit access; and publish the returned favorable or unfavorable aggregate packet unchanged. The software archive is public, but the strict external milestone remains pending and the steel-energy challenge carries no result claim.

Data and Code Availability

The public software, benchmark configurations, aggregate result records, tests, container files, and reproduction instructions are archived as SynthGuard-ReleaseBench version 0.1.14 at https://doi.org/10.5281/zenodo.21909680. The public archive does not contain reviewer forms, signatures, email addresses, local file paths, credentials, confidential audit rows, or restricted row-level data. Public source data are obtained from the cited providers using the documented retrieval and integrity-check procedures.

11 Conclusion

Synthetic-data evaluation needs a defensible middle ground between "looks realistic" and "safe to release." SynthGuard-ReleaseBench provides one: declare the intended use and tolerance, lock the candidate and analysis panels, evaluate them on a protected audit with simultaneous uncertainty, calibrate the workflow with controls, report privacy evidence for a named threat model, and preserve a separate human release decision.

The empirical record shows why each part matters. Some transparent candidates certify for declared uses; compact learned models do not under the tested budgets; one entire case is excluded when its negative control certifies; a known leaky control is detected by the held-out attack; and the sealed prototype blocks an otherwise passing utility grid when lifecycle evidence is missing. These are use-specific outcomes, not a universal generator ranking.

The contribution is therefore a reproducible method for making release evidence specific, falsifiable, and honest about uncertainty and governance. Its strongest evidence includes the failures it refuses to hide. A benchmark earns trust not by making every candidate pass, but by making unsupported release claims difficult to produce.

A Protocol Checklist

Table 11 lists the minimum fields that make a SynthGuard evidence record auditable.

Table 11: Minimum evidence fields in a SynthGuard record.

Field	Required record
Declared use	Named analysis, target population, real reference workflow, bounded loss, and tolerance.
Candidate integrity	Generator family, version, data access, fixed seed, selection rule, and candidate hash.
Audit integrity	Unit of splitting, audit size, relationship to fitting, scheduled looks, and access controls.
Uncertainty	Per-analysis gap, bound, alpha allocation, sample size, decision, and all holds.
Risk evidence	Threat model, attacker information and access, calibration split, evaluation split, metric, and uncertainty interval.
Mechanism claim	Exact privacy definition, adjacency, public-domain policy, composition, and any exclusions.
Governance	Attestations, lifecycle status, release-gate status, and human decision record.

B The Anytime Confidence Sequence in Full

Section 4.3 reports interval widths from a predictable plug-in empirical Bernstein confidence sequence. We give the exact construction so the numbers can be reproduced without reading the implementation.

Fix a cell (g, a) and let $X_1, X_2, \ldots$ be its paired gaps, bounded in $[-B_a, B_a]$. Rescale to the unit interval by $U_t = (X_t + B_a)/(2B_a)$, so a bound on the mean of U maps back by $\mu_X = 2B_a\mu_U - B_a$. All quantities below are computed on the U scale at the cell level $\delta = \alpha/(|\mathcal{G}||\mathcal{A}|)$; no spending across looks is required, because the sequence is valid at every t simultaneously.

Define the predictable estimates, each using only the first $t-1$ observations,

$$\widehat{\mu}_t = \frac{\frac{1}{2} + \sum_{s<t} U_s}{t}, \qquad \widehat{\sigma}_t^2 = \frac{\frac{1}{4} + \sum_{s<t}(U_s - \widehat{\mu}_s)^2}{t},$$

with $\widehat{\mu}_1 = 1/2$ and $\widehat{\sigma}_1^2 = 1/4$; the constants $1/2$ and $1/4$ are a prior observation at the midpoint of the unit interval. Set the predictable betting fraction

$$\lambda_t = \min\left\{ \sqrt{\frac{2\log(2/\delta)}{\widehat{\sigma}_t^2\, t\, \log(1+t)}}\ ,\ \tfrac{1}{2} \right\},$$

where the cap $\lambda_t \le 1/2$ keeps the process well defined. With $\psi_E(\lambda) = \{-\log(1-\lambda) - \lambda\}/4$, the interval after t observations is centred and half-widened by

$$\widehat{c}_t = \frac{\sum_{s\le t} \lambda_s U_s}{\sum_{s\le t} \lambda_s}, \qquad w_t = \frac{\log(2/\delta) + \sum_{s\le t}(U_s - \widehat{\mu}_s)^2 \psi_E(\lambda_s)}{\sum_{s\le t} \lambda_s},$$

and the reported statement is $\mu_U \in [\widehat{c}_t - w_t,\ \widehat{c}_t + w_t]$ for all t simultaneously with probability at least $1-\delta$. Mapping back to the original scale multiplies both $\widehat{c}_t$ and w_t by $2B_a$ and shifts the centre by $-B_a$.

Validity is the predictable plug-in empirical Bernstein construction of Waudby-Smith and Ramdas [24]: the process $\prod_{s\le t}\exp\{\lambda_s(U_s-\mu_U)-(U_s-\widehat{\mu}_s)^2\psi_E(\lambda_s)\}$ is a non-negative supermartingale with initial value one under the true mean, so Ville's inequality bounds the probability that it ever exceeds $1/\delta$. Because λ_t and $\widehat{\mu}_t$ depend only on the past, no adaptivity correction is needed, and the sequence may be stopped at a data-dependent time.

Two properties explain the widths in figure 5. The half-width adapts to the realised variance through $\widehat{\sigma}_t^2$, so concentrated gap vectors give narrow intervals; and the $\log(1+t)$ factor in λ_t is what buys time-uniformity, which is why the sequence can be tighter than a scheduled bound that must spend α across looks in advance.

C Certificate Assumptions and Failure Modes

The proposition requires more than a held-out CSV file. The loss, tolerance, candidate set, scheduled looks, and audit unit must be fixed before audit evaluation. If a protected audit is repeatedly used to alter preprocessing, tune a generator, select a subgroup, or search for a favorable metric, the stated union-bound guarantee no longer accounts for that adaptation. A fresh audit or a more appropriate adaptive-inference design is then needed.

Bounded losses are a practical choice. Brier loss already lies in $[0,1]$; clipped log loss and normalized descriptive error are bounded by declaration. Bounding makes the uncertainty rule transparent but can be conservative. It is therefore appropriate to report the actual gap, the half-width, and the tolerance together, not only a binary label.

The iid assumption is conditional on the frozen pre-audit workflow. It must be replaced or augmented when the target sampling unit is a household, patient, site, time block, or spatial cluster. When blocks can be justified as independent before audit, proposition 9 operates on equally weighted block means and reports the effective number of blocks. It does not turn dependent time windows into independent observations. The public ACS examples use row-level diagnostics only. A sealed deployment should split at the unit that matches the risk and inference target.

D Completed Study Inventory

The complete study-by-study inventory, including every evidence boundary and the outcome-free steel-energy challenge, is supplied as Supplementary Tables S1–S2. Moving the inventory out of the main article keeps the result narrative focused without removing unfavorable or excluded records from the release.

E Exploratory CTGAN Correction

The repository preserves an earlier exploratory ACS CTGAN artifact together with a correction record. A subsequent overlap audit found that the alleged 60,000-row fresh replication was not record-disjoint from the original 100,000-row sample: 28,098 rows overlapped. That work is marked `INVALIDATED_NOT_RECORD_DISJOINT` and is excluded from the results and conclusions of this manuscript. Retaining the correction is intentional. It records an integrity failure that future benchmark submissions must be able to detect.

References

[1] U.S. Census Bureau. (2024). *American Community Survey Public Use Microdata Sample documentation.*

[2] Centers for Disease Control and Prevention. (2024). *BRFSS Prevalence Data*. Public data API, resource dttw-5yxu.

[3] City of Chicago. (2024). *Taxi Trips (2024-)*. Chicago Data Portal, resource ajtu-isnz.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 265–284.

[5] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248), 636–638.

[6] J. Jordon, J. Yoon, and M. van der Schaar. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

[7] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3), 224–232.

[8] R. McKenna, G. Miklau, and D. Sheldon. (2021). Winning the NIST contest: A scalable and general approach to differentially private synthetic data.

[9] S. Moro, P. Cortez, and P. Rita. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22–31.

[10] National Institute of Standards and Technology. (2022). *SDNist: Synthetic Data Report Tool*.

[11] B. Nowok, G. M. Raab, and C. Dibben. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74(11), 1–26.

[12] J. Opoku and D. Banahene. (2026). Scale-aware uncertainty-aware machine learning for gold prospectivity mapping: Evidence-range audits for exploration targeting. *EarthArXiv preprint*. DOI: 10.31223/X5DJ45.

[13] J. Opoku and D. Banahene. (2026). Spectral adaptive conformal prediction for structured non-exchangeable data. *arXiv preprint* arXiv:2606.15950.

[14] J. P. Reiter. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3), 441–462.

[15] D. Ridgeway, M. Theofanos, T. Manley, and C. Task. (2021). Challenge design and lessons learned from the 2018 differential privacy challenges. NIST Technical Note 2151.

[16] DataCebo. (2024). *SDGym documentation*.

[17] F. Houssiau, S. N. Cohen, L. Szpruch, O. Daniel, M. Lawrence, R. Mitra, H. Wilde, and C. Mole. (2022). A framework for auditable synthetic data generation. *arXiv preprint* arXiv:2211.11540.

[18] F. Houssiau, J. Jordon, S. N. Cohen, O. Daniel, A. Elliott, J. Geddes, C. Mole, C. Rangel-Smith, and L. Szpruch. (2022). TAPAS: a toolbox for adversarial privacy auditing of synthetic data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

[19] K. Tikhomirov. (2020). Singularity of random Bernoulli matrices. *Annals of Mathematics*, 191(2), 593–634.

[20] I. Dinur and K. Nissim. (2003). Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202–210.

[21] M. Hardt and J. Ullman. (2014). Preventing false discovery in interactive data analysis is hard. In *55th IEEE Annual Symposium on Foundations of Computer Science*, 454–463.

[22] T. Steinke and J. Ullman. (2015). Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Conference on Learning Theory*, 1588–1628.

[23] A. Maurer and M. Pontil. (2009). Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Conference on Learning Theory*.

[24] I. Waudby-Smith and A. Ramdas. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1), 1–27.

[25] Y. Du and N. Li. (2025). Systematic assessment of tabular data synthesis. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[26] DataCebo. (2026). *SDMetrics documentation: Quality and diagnostic reports*. `https://docs.sdv.dev/sdmetrics`.

[27] D. J. Schuirmann. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680. DOI: 10.1007/BF01068419.

[28] A. D. Lautrup, T. Hyrup, A. Zimek, and P. Schneider-Kamp. (2024). SynthEval: A framework for detailed utility and privacy evaluation of tabular synthetic data. arXiv:2404.15821.

[29] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi. (2022). A unified framework for quantifying privacy risk in synthetic data. arXiv:2211.10459.

[30] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, 3–18.

[31] J. Snoke, G. Raab, B. Nowok, C. Dibben, and A. Slavković. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A*, 181(3), 663–688.

[32] T. Stadler, B. Oprisanu, and C. Troncoso. (2022). Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium*, 1451–1468.

[33] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems 32*, 7333–7343.

[34] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. (2023). TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, 17564–17579.

[35] H. Zhang, J. Zhang, B. Srinivasan, Z. Shen, X. Qin, C. Faloutsos, H. Rangwala, and G. Karypis. (2024). Mixed-type tabular data synthesis with score-based diffusion in latent space. In *International Conference on Learning Representations*.

[36] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. (2017). PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems*, 42(4), Article 25.

[37] I.-C. Yeh. (2009). *Default of Credit Card Clients [Dataset]*. UCI Machine Learning Repository. DOI: 10.24432/C55S3H.

[38] C. Sakar and Y. Kastro. (2018). *Online Shoppers Purchasing Intention Dataset [Dataset]*. UCI Machine Learning Repository. DOI: 10.24432/C5F88Q.

[39] B. Becker and R. Kohavi. (1996). *Adult [Dataset]*. UCI Machine Learning Repository. DOI: 10.24432/C5XW20.

[40] J. Clore, K. Cios, J. DeShazo, and B. Strack. (2014). *Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]*. UCI Machine Learning Repository. DOI: 10.24432/C5230J.

[41] S. V. E., C. Shin, and Y. Cho. (2023). *Steel Industry Energy Consumption [Dataset]*. UCI Machine Learning Repository. DOI: 10.24432/C52G8C.

[42] J. Opoku and D. Banahene. (2026). *SynthGuard-ReleaseBench: Locked-audit evidence for synthetic tabular data releases*. Zenodo. Version 0.1.14, DOI: 10.5281/zenodo.21909680. Concept DOI 10.5281/zenodo.21503396 resolves to the latest version.