

Source information limits lower-risk claims from synthetic validation cohorts

Tianyi Yu

`tiangyi.1.yu@kcl.ac.uk`

King's College London

Research Article

Keywords: synthetic data, predictive performance, subgroup validation, statistical inference, simulation study

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11196055/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Enlarging a synthetic validation cohort can improve conditional precision without adding original subgroup information. We examined how release size and changes in majority source data affect lower-risk claims for a fixed minority population.

Methods

We combined a Bernoulli baseline with 50,000 independent source cohorts in each of 45 settings, a matched parametric experiment with 1,000 cohorts, and source-refit uncertainty estimation using 399 resamples in 200 cohorts. Each matched cohort contained 200 fixed minority records. An independent 1,000-source oracle diagnostic examined frozen error-distribution adjustments. Neural feasibility experiments included a completed 24-source, 120-fit public finite-record benchmark. Comparisons used paired Brier loss, retaining harmful, beneficial and equal-risk orientations.

Results

In a specified adverse baseline setting, conditional Wald false certification increased from 6.2% to 35.2% with synthetic expansion. An unpenalized fully interacted generator was invariant to majority size, yet boundary rejection rose from 11.1% to 44.5% as releases grew from 200 to 20,000 records. In a correctly specified shared control, source refitting reduced boundary rejection from 79/200 to 4/200; useful certification also decreased. Fixed shrinkage and deliberate misspecification retained signed errors, while frozen marginal centering reduced major distortions on fresh sources. Bootstrap availability ranged from 161 to 200 cohorts. In the public benchmark, minority-only certification of the beneficial natural model comparison rose from 4/24 to 17/24, while false certification of its reverse rose from 0/24 to 4/24. Natural-comparison majority-size intervals included zero. Neural fidelity limitations and integration-precision misses remained.

Conclusions

Conditional-release precision, source uncertainty and fitted-target distortion require separate assessment. Source refitting substantially reduced excess certification in correctly specified controls, with a power cost; variance correction did not remove every error under fixed bias. The public experiment illustrates these distinctions without establishing adequate-generator inference, an isolated majority-sharing effect or clinical validity.

Background

The research question is whether an enlarged synthetic validation cohort can justify a lower-risk claim that its original subgroup observations do not support. The procedure uses a zero noninferiority margin and requires evidence of strictly lower expected paired Brier loss; mathematically, it is a superiority test. Here, certification denotes this statistical decision. False substantive certification is assessed for genuinely inferior candidates, while rejection at equal risk is reported separately as boundary-null error. A second, more discriminating question is whether a shared generator changes that subgroup comparison when the subgroup's own records remain fixed but the majority source data change. These questions concern evidence for expected predictive loss. They do not equate a favourable statistical comparison with clinical safety, treatment effectiveness or benefit to patients.

Synthetic health data are being developed for several legitimate purposes, including accessible research resources, predictive modelling and longitudinal disease analysis. Multifaceted electronic-health-record benchmarks, longitudinal generators and recent multicountry cohort synthesis illustrate the diversity of these uses [1, 2, 3]. Methods designed for mixed clinical tables, complex missingness and longitudinal structure also evaluate preservation of clinically relevant relationships [4, 5, 6]. A comparison of obfuscation and generative techniques found utility to depend on the analysis [7]. Broader evaluations likewise distinguish fidelity, utility and privacy [8, 9, 10, 11]. A limitation of a particular inferential use should therefore not be mistaken for evidence that synthetic data have no value. Minority augmentation, for example, can improve some prediction objectives under specified data conditions [12].

The inferential distinction is older than contemporary generative modelling. Synthetic-data combining rules depend on how records were generated and on what the analyst is trying to estimate [13, 14, 15]. General distributional utility and analysis-specific utility need not agree [16], and established synthesis software makes this distinction explicit [17]. Recent experiments demonstrate inflated false-positive rates when artificial records are analysed as though they were observed independently from the target population [18]. Debiasing methods address additional discrepancies that variance correction alone cannot remove [19]. Bayesian consistency results require a compatible synthesis and analysis regime [20], while differentially private synthesis introduces further inferential constraints [21]. These results preclude presenting the existence of source-uncertainty inflation as a new finding.

The question is also distinct from whether synthetic data can help choose or evaluate a model. Synthetic model selection and ranking calibration have direct precedents [22, 23]. AutoEval combines synthetic and observed labels for efficient evaluation [24], and prediction-powered inference uses observed outcomes to correct machine predictions [25]. Risk-control procedures, synthetic-assisted inference and task-level calibration offer explicit guarantees under their respective information structures [25, 26, 27, 28, 29, 30, 31]. In particular, conformal risk control bounds marginal expected monotone loss [30]; that guarantee differs from a high-probability certificate for one fitted source-population comparison. Generating complete records from a small source dataset is not automatically equivalent to obtaining independently sampled covariates or additional observed labels. An analysis must identify which information is genuinely new and which uncertainty is conditional on an already fitted generator.

Subgroup evidence deserves particular care. A large total sample can coexist with few observations in the subgroup of interest. External-validation sample-size guidance accordingly concerns precision in the relevant performance measures, not only the headline number of records [32, 33]. Multicalibration and fairness-auditing research likewise distinguish aggregate properties from subgroup claims [34, 35]. Reporting and bias-assessment frameworks ask investigators to make their evaluation population, analysis choices and applicability explicit [36, 37]. Merely oversampling a subgroup synthetically cannot be assumed to satisfy these evidential requirements.

We use expected paired Brier loss because its target is transparent and it admits an exact verification example. Proper scoring rules assess probabilistic predictions in expectation [38, 39]; the Brier score reflects more than calibration alone [40, 41]. It is not, however, a complete measure of clinical utility [42]. Decision-curve analysis addresses a different, threshold-dependent question [43]. The present study therefore evaluates false certification of strictly lower expected loss. The baseline validates the measurement and uncertainty controls; the matched parametric experiment then separates release-size effects from changes in the fitted minority comparison. Completed neural experiments assess the behavior and limitations of the fitted procedures; their fidelity diagnostics determine the scope of the resulting claims. Figure 1 summarizes the design and evidence flow.

Methods

Simulation design and estimand

The simulation follows the aims, data-generating mechanisms, estimands, methods and performance-measures structure recommended for statistical-method evaluation [44]. Total source size n took values 200, 1,000 and 5,000. The subgroup probability q took values 0.02, 0.10 and 0.50. For each independent cohort, K was drawn from $\text{Binomial}(n, q)$, and subgroup events X from $\text{Binomial}(K, p)$, with p in $\{0.10, 0.14, 0.15, 0.16, 0.20\}$. Majority events were drawn independently conditional on K with probability 0.30. A source cohort did not represent any real person.

We adopted the binary Brier convention (prediction–outcome) squared, without doubling for the two outcome categories [38, 42]. Both comparison orientations were retained. The null for certification was expected paired loss difference greater than or equal to zero; certification required a one-sided upper bound strictly below zero. Boundary rejection at exactly zero and false certification under strict inferiority were reported separately. An empty source subgroup always returned inconclusive.

Synthesis and uncertainty procedures

The empirical generator used X/K as the subgroup event probability. The pooled negative control used all source events divided by n . The oracle used the known subgroup probability p . Synthetic subgroup size was K times an expansion ratio in $\{1, 5, 20, 100, 1000\}$; empty groups used an internal numerical

placeholder but remained ineligible for certification. Binomial counts represented each complete synthetic release without materializing redundant binary rows.

All tests used one-sided alpha = 0.05. Conditional Wald tests used the synthetic plug-in draw variance $\frac{\tilde{p}(1-\tilde{p})}{m}$, where $\tilde{p}$ is the generated event proportion. The variance-corrected calculation instead used the fitted generator probability $\frac{p_g(1-p_g)}{m}$ for conditional draw variance, plus its estimated fitting variance. Conditional exact tests used the corresponding one-sided binomial tail at probability 0.15. The variance correction added a generator-specific fitting term to conditional draw variance. For the empirical generator with $K > 1$, this term was $\frac{\hat{p}(1-\hat{p})}{K-1}$. For the pooled source, it was $\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n-1}$, interpreted unconditionally over the iid mixture design. For the oracle it was zero. Empirical correction returned inconclusive at $K = 1$ because the required source variance was not estimable. Exact source tests used the original X and K only [45].

Writing z for the standard-normal 0.95 quantile, the higher-prediction comparison certified when $\tilde{p}-z \cdot \text{SE} > 0.15$; the reverse certified when $\tilde{p}+z \cdot \text{SE} < 0.15$. Exact procedures required the corresponding binomial-tail p -value to be strictly below 0.05. The absolute Brier coefficient 0.20 converts event-rate bounds into loss bounds.

At fixed $K = k$, the variance of the empirical-generator synthetic event rate is $p(1-p)\left[\frac{1}{k} + \frac{1-\frac{1}{k}}{m}\right]$. Its nonvanishing first term explains the source-information limit. The supplementary mathematical checks include exact enumeration and the discrete large-release boundary term. The correction is an established statistical strategy, not a proposed new method [13, 18].

Monte Carlo reporting and verification

The larger baseline used seed 9276301 with separate SeedSequence streams for the 45 source configurations and 50,000 source cohorts per configuration. All methods and release sizes shared the relevant source cohort; generated release draws were saved. Each probability estimate uses all 50,000 source cohorts as its denominator, including inconclusive cohorts, and has a Monte Carlo standard error and a pointwise 95% Wilson interval. These intervals are not simultaneous bands across the reported grid. Expansion contrasts use the standard error of paired replicate differences. Repeated exact-source results are not counted as independent observations.

Checks compared exact binomial-test size with enumeration, verified the paired Brier algebra and variance decomposition, and reconstructed decisions from stored source and synthetic counts. A secondary internal agent audit, conducted separately from implementation but within the same model family, additionally reconstructed selected counts and paired standard errors without importing the decision function. Such audits are computational cross-checks, not external journal peer review or guarantees of scientific novelty.

Matched-source parametric experiment

To separate conditional simulation noise from changes in a fitted generator, we used a mixed-covariate population with a known logistic outcome mechanism. Each of 1,000 independently sampled source cohorts supplied 200 minority records. Within a cohort, the minority records and target population remained fixed while the majority sample size increased from 800 to 3,200 under nested sampling. Majority outcome shifts were $\kappa = -1, 0$ and $+1$. Prespecified model comparisons included two harmful candidates with population excess Brier risk 0.002, their reversed beneficial comparisons, and a nondegenerate equal-risk comparison. These five comparisons shared source cohorts and were not treated as independent replications.

We fitted minority-only, fully interacted and shared main-effects logistic generators. The fully interacted model could represent the group-specific outcome relationship; the main-effects model deliberately omitted the group-by-age effect and was therefore misspecified for nonzero κ . Ridge penalties 0, 1 and 10 were fixed on the summed negative log likelihood, with the common intercept unpenalized. Generated minority covariates followed their empirical source distribution, and outcomes followed the fitted conditional probabilities. Fitted-generator mean paired losses integrated the binary outcome analytically, avoiding synthetic Monte Carlo error in this estimand. Nested releases contained 200, 4,000 and 20,000 records. Conditional normal and bounded-loss decisions were evaluated against the population truth, with original-source decisions as benchmarks. Rate intervals were pointwise exact binomial 95% Monte Carlo confidence intervals over independent source cohorts; paired-drift intervals were pointwise t intervals. The reported analyses did not adjust for multiplicity.

Full source-refit uncertainty

For source IDs 0–199, we refitted each of the 39 logistic generator configurations on 399 resampled source datasets. Minority rows were resampled with replacement. Majority resampling preserved the nested 800-record prefix and additional 2,400-record block; matched resampling indices were reused across outcome shifts. This was the declared block-stratified procedure, rather than an ordinary bootstrap of the combined majority sample. Each refit's minority covariate distribution also followed its resampled minority records. The sample variance of the refitted exact generator means estimated source variation. For a finite release, the one-sided normal bound added this variance to the fitted conditional paired-loss variance divided by release size; for an analytic-generator mean, the latter term was zero. The original point estimate was retained, so the procedure did not directly correct its bias.

A source/configuration bound required all 399 refits to pass the recorded validity criteria. We reported planned and evaluated denominators separately. Paired correction effects were evaluated on identical source cohorts with both decisions available, with source-level standard errors and pointwise t intervals; logical ranges retained uncertainty from unavailable pairs. Rate intervals were pointwise exact binomial intervals. These Monte Carlo intervals describe performance in the specified simulation and are not simultaneous guarantees over the configuration grid.

Independent oracle diagnostic

This prospectively frozen local diagnostic used original source IDs 0–999 as an inspected reference and fresh source IDs 1000–1999 for evaluation. For each fixed generator/comparison/release cell and each source-only comparison, we froze the reference mean error μ , sample standard deviation s and linearly interpolated empirical fifth-percentile error q . Errors were estimates minus known population truth. The three upper bounds for a fresh estimate T were $T + z_{0.95}S$, $T - \mu + z_{0.95}S$ and $T - q$. Synthetic-cell reference errors already included finite-release noise; no second conditional-variance term was added. All 785 cells, all orientations and all release sizes were retained without holdout-based tuning. Rates and paired differences used the 1,000 independent source cohorts as units. Reference parameters and original evidence hashes remained unchanged.

Bounded neural training and weighting diagnostic

The amendment fixed source cohort 0, fitting-seed offset 0, 6,000 updates per optimizer and 13 configurations: one minority-only, six shared and six balanced fits. The seven minority-only/shared cells were paired with their existing 600-update counterparts. Balanced fits repeated minority rows four or sixteen times to match majority row counts at 800 or 3,200. The original source arrays, predictors, architecture and sampling rules were retained. Optimizer traces verified actual update budgets, and saved sampler counters reported minority exposures. Raw retained rows were used to reconstruct paired losses, feature marginals and support violations. All shifts, predictor orientations and precision failures were reported. No across-source confidence interval or calibration rate was estimated from this single-source experiment.

Neural feasibility experiments

We used a conditional tabular generative adversarial network (CTGAN) [46] to model the joint mixed-covariate and binary-outcome distribution. The known-truth feasibility experiment comprised 24 independent source cohorts, with the same 200 minority records held fixed across one minority-only and six shared fits per cohort. The shared fits crossed majority sizes 800 and 3,200 with shifts $\kappa = -1, 0$ and $+1$. All 168 fits used 600 optimizer updates and one fitting-seed offset. Actual generated group membership was checked before retaining minority records. Numerical integration samples and nested releases served different purposes: the former estimated fitted-generator paired-loss means, while the latter supplied release-based decisions. Absolute and relative numerical-precision diagnostics were retained alongside feature support and outcome fidelity. Additional file 1 (Supplementary methods) specifies architecture, preprocessing, sampling and the mixed-covariate population.

Public finite-record benchmark

The public benchmark used the CDC Diabetes Health Indicators derivative distributed through UCI and its linked author export [47]. The derivative contains 253,680 records. The linked author export is licensed CC0; input and metadata fingerprints are retained with the provenance record. The supplied binary outcome was retained unchanged. Comparison with the author's three-class export showed that the positive label represented diabetes, while both the normal and prediabetes classes mapped to zero. We did not rely on UCI's inconsistent combined-outcome description. The author's stated 2015 provenance and UCI's additional 2014 link could not be reconciled through a raw BRFSS reconstruction, which was not performed.

The minority group comprised recorded household-income categories 1 and 2, representing annual income below USD15,000. The original income variable was replaced by this group indicator. A fixed split, stratified jointly on outcome and group, allocated 76,104 records to training, 25,368 to tuning and 152,208 to evaluation. The evaluation minority comprised 12,957 records, including 3,282 positive outcomes. Logistic regression and histogram gradient boosting were selected using development data and frozen before evaluation. Logistic regression used $C = 0.01$; the selected boosting configuration used 15 leaves and L2 regularization 1. Models were not refitted using evaluation records. The natural model comparison was gradient boosting minus logistic regression. Fixed logit perturbations of -0.25 and $+0.25$ to the development-selected boosting model supplied stress comparisons and their reversals; evaluation outcomes did not choose the base model, perturbations or reported directions.

The target for each comparison was the exact mean paired Brier loss over the fixed minority evaluation records, up to floating-point precision. It was not an estimate of a nationally representative population quantity. The natural gradient-boosting-minus-logistic-regression target was -0.00226287 . The two perturbed candidates had excess paired losses of $+0.00165468$ and $+0.00196615$ relative to the base model. All six orientations, including the natural reverse, were retained.

We sampled 24 source cohorts with replacement from the fixed evaluation population. Each supplied 200 minority records, held identical within that cohort across generator interventions. Majority sample sizes were 800 and 3,200 with nested prefixes. Separate age-stratum streams and deterministic interleaving produced majority proportions aged at least 55 years of 0.25 and 0.75. This fixed-composition stratified source law differs from unrestricted sampling of the original majority population. Covariates and labels remained attached to their sampled records. Each source supplied one minority-only fit and four shared fits, giving 120 fits with one fitting-seed offset and 6,000 updates per optimizer.

A preceding raw-day representation produced severe support violations and was retained as failed feasibility evidence. A dated amendment represented mental- and physical-health questionnaire day counts as categorical values from 0 to 30; Body mass index (BMI) remained continuous. The revised encoding was applied to every arm without changing predictors, source IDs, population targets or intervention definitions. It changes model representation and conditional training, rather than implementing an invisible clipping correction. Actual generated group membership was checked before retaining minority rows; subgroup labels were not overwritten. Nested releases contained 200, 4,000 and

20,000 minority records. Separate adaptive generated samples estimated fitted-generator paired-loss means, with numerical precision recorded independently of fitting completion.

We reported source-only decisions, conditional synthetic normal and bounded-loss decisions, paired majority-size and composition contrasts, and within-fit release-expansion contrasts. Independent source cohorts, rather than fits, generated rows or reversed comparisons, supplied the units of source-level uncertainty. No public source-refit correction or empirical multi-seed stability assessment was performed. Support, subgroup fidelity and absolute integration-precision misses were retained. Profile overlap between development and evaluation records involved coarse feature/outcome patterns and did not establish repeated-person identity. These analyses concern an unweighted cross-sectional finite-record benchmark, not electronic health records, incident-disease prognosis, external clinical validation or national representation.

Numerical precision and public interval reporting

Public fitted-generator means used adaptive integration starting at 20,000 retained minority rows, increasing in blocks of 20,000 to at most 200,000. The stopping target was an estimated absolute Monte Carlo standard error (MCSE) no larger than 0.0002 for every paired-loss comparison. Hitting the cap with a larger estimated MCSE was a retained precision miss. These estimates describe numerical noise after adaptive stopping; they are not confidence sequences. For intervention contrasts, we compared the aggregate conditional mean-MCSE bound with 10% of the across-source standard error. A stricter constituent check required every contributing mean's estimated MCSE to be no larger than the minimum of 0.0002 and 10% of the source-effect standard error. The aggregate and constituent checks answer different questions and were both retained. Additional file 1 defines their calculations and the distinct within-fit expansion diagnostics.

Public certification-rate intervals were two-sided pointwise 95% Wilson intervals over 24 independent source cohorts. Paired intervention and equal-weight marginal-mean intervals used source-level t intervals with 23 degrees of freedom. Their uncertainty includes source sampling and fitting variability, whereas conditional integration MCSE concerns drawing from a fixed fitted generator. All orientations and both normal and Hoeffding procedures are retained in the public rate, contrast and diagnostic tables (Additional file 5). None of these intervals adjusts for the many dependent comparisons.

Protocol evolution

Dated local protocols and amendments preserve the baseline's pilot-informed design, representation changes, bounded neural follow-up, reference freezing and fresh-source evaluation. These records are local research documentation, not registration with an external registry. Completed grids were retained with negative results, failed fits and numerical-precision misses; no new source or seed grid was used to obtain a preferred conclusion.

Computational assistance and accountability

Codex was used for literature retrieval, code debugging and manuscript grammar checking. Responsibility for the scientific claims and final submission remains with the named author. The author has reviewed and approved the final manuscript and accepts responsibility for its content.

Results

Transparent comparison and independently varied source cohorts

The completed baseline (Fig. 2) included 45 source settings: three total cohort sizes, three subgroup probabilities and five subgroup event probabilities. Each setting contained 50,000 independently simulated source cohorts, for 2.25 million source cohorts overall. These were simulated units, not observed patients. Every source cohort was reused across the three generator mechanisms, five expansion ratios and paired testing procedures. Those repeated analyses are paired comparisons, not additional independent source replications. The complete summary contains 5,400 method-condition rows, with raw source and release counts sufficient to reconstruct every decision.

The higher-probability candidate predicted 0.20 and the reference predicted 0.10. Its expected excess Brier loss was $0.03 - 0.20p$, where p is the target subgroup event probability. At $p = 0.14$, the candidate was inferior by 0.002 expected Brier units. Reversing the two predictions supplied the opposite comparison; $p = 0.15$ supplied the boundary null. Beneficial alternatives were retained rather than evaluating error control alone.

Synthetic expansion increased false certification in a pilot-informed adverse example

With total source size 1,000 and subgroup probability 0.10, conditional Wald certification of the inferior higher-probability candidate increased from 6.216% at one-fold expansion to 35.248% at 1000-fold expansion. The corresponding Monte Carlo 95% intervals were 6.008–6.431% and 34.830–35.668%. Intermediate expansion ratios of 5, 20 and 100 produced rates of 15.654%, 24.602% and 31.082%, respectively. The exact source-data benchmark certified 1.894% throughout because its evidence did not change when additional synthetic records were drawn.

Replacing the synthetic Wald test by an exact conditional binomial test did not solve this problem. Its certification rate was 6.688% at one-fold and 35.250% at 1000-fold expansion. Exactness for the fitted generator's conditional sampling distribution therefore did not imply validity for the original population target. The generator-variance correction reduced the rates in this example to 2.782% and 1.574%, respectively. This rate reduction is consistent with established synthesis-variance results and is specific to the evaluated setting.

Sparse source data exposed a separate failure of plug-in variance correction

At total size 200, subgroup probability 0.02 and event probability 0.16, the lower-probability candidate was inferior by 0.002 expected Brier units. At 1000-fold expansion, the synthetic Wald and conditional exact procedures certified it in 54.136% and 54.112% of source cohorts. The plug-in correction certified 45.270%. Independent reconstruction showed that this last count exactly matched the 22,635 source cohorts with more than one subgroup observation and no subgroup events: the estimated variance collapsed to zero.

The exact source test certified none of these inferior comparisons, but had no useful power at this sample size. The largest observed subgroup count was 15. Even with no events, at least 19 original observations would be needed for the relevant one-sided exact upper bound to fall below 0.15. The same test also had zero observed certification power for the beneficial lower-prediction comparison at $p = 0.10$ in this sparse setting. It correctly expressed inadequate information rather than extracting a favourable decision from it.

Expansion did not universally increase error. In the sparse adverse example, synthetic Wald certification decreased from 62.034% at one-fold expansion to 54.136% at 1000-fold expansion. Small-sample approximation failure and source uncertainty are distinct mechanisms, and their combination need not be monotonic. The effect of expansion therefore depends on the source setting and inference procedure.

Correct variance did not eliminate approximation error or generator bias

The truth-knowing oracle removed generator-fitting uncertainty. For the higher-prediction comparison at the boundary null with source size 1,000 and subgroup probability 0.10, its variance-corrected normal approximation still rejected 5.586% at one-fold expansion, with a Monte Carlo 95% interval of 5.388–5.791%. Correct variance accounting did not guarantee exact finite-sample calibration. Conditional exact tests and larger generated samples helped distinguish this approximation issue from omitted source uncertainty.

The pooled generator deliberately ignored subgroup-specific event rates. Its variance estimate described the full mixture law, whose mean could differ materially from the minority target. Increasing the synthetic sample made this biased comparison precise; changing its variance did not restore the subgroup estimand. This is an intentionally misspecified negative control. It provides no estimate of how frequently a well-trained modern generator behaves this way.

Fixed-minority parametric controls separate two mechanisms

All 1,000 planned source cohorts passed artifact checks (Fig. 3). No main-analysis decision cell was unavailable. For the unpenalized fully interacted model, changing majority size produced a maximum absolute mean paired drift of 8.57×10^{-10} across the five comparisons and three outcome shifts. This

numerical invariance is the expected negative control: without penalization, the group-specific outcome fit separates, while the empirical minority covariate distribution is unchanged.

Nevertheless, conditional synthetic certification did not inherit that invariance to release size. With 3,200 majority records and $\kappa = 0$, the fully interacted model falsely certified the positive-offset harmful candidate in 42/1,000 cohorts at one-fold expansion and 172/1,000 at 100-fold expansion. Boundary rejection rose from 111/1,000 to 445/1,000. The corresponding original-source normal decisions were 5/1,000 false certifications and 31/1,000 boundary rejections. The original-source procedure also certified the reversed beneficial candidate in only 194/1,000 cohorts. Its lower error rate therefore needs to be read alongside its limited power; the comparison does not establish an exact or generally efficient source-only procedure.

For the unpenalized shared main-effects generator, increasing majority size changed the exact fitted-generator mean loss for the equal-risk comparison. The paired mean changes were + 0.000662 (95% Monte Carlo interval, + 0.000582 to + 0.000742) at $\kappa = -1$, + 0.0000614 (− 0.000000426 to + 0.000123) at $\kappa = 0$, and − 0.000235 (− 0.000286 to − 0.000184) at $\kappa = +1$. Thus, majority changes could move the fitted comparison in either direction. The correctly specified $\kappa = 0$ result did not clearly distinguish the average drift from zero. The nonzero- κ changes arose under deliberately omitted group interactions and cannot establish a new general failure of competent shared generators.

Source refitting reduced error in correctly specified controls

Figure 4 summarizes correction, useful power and unavailable bounds on matched sources.

The source-refit analysis completed all 200 prespecified source cohorts, with 399 refits for each generator configuration. Before-and-after comparisons used these same 200 cohorts, rather than comparing the corrected subset with the larger main experiment. Before-and-after counts below compare the conditional normal bound with the source-refit normal bound. Of 3,112,200 requested refits, 922 failed the declared numerical validity checks: 909 triggered possible separation or ill-conditioning diagnostics and 13 ended with an abnormal optimizer exit. These failures affected 284 source/configuration pairs across 47 sources. Under the requirement that all 399 refits be valid, evaluated denominators ranged from 161 to 200. An unavailable bound remained inconclusive; it was not counted as an evaluated noncertification or successful coverage event.

At majority size 3,200, zero outcome shift and zero penalty, the correctly specified shared main-effects generator supplied an available bound for every source. For releases of 20,000 minority records, source refitting reduced harmful certification from 32/200 to 0/200 in each of the two harmful orientations and boundary rejection from 79/200 to 4/200. The corrected boundary rate was 2.0% (pointwise exact 95% Monte Carlo interval, 0.55–5.04%). Useful certification also fell: the two beneficial reversals changed from 149/200 and 147/200 to 32/200 and 40/200. Original-source normal bounds certified these beneficial comparisons in 30/200 and 41/200 cohorts. Thus, correction substantially reduced the

conditional-release excess while returning useful certification to a scale similar to the source-only benchmark in this setting.

The unpenalized separate and fully interacted controls exposed an additional availability cost. For the fully interacted model at zero shift and majority size 3,200, 39 source/configuration procedures were unavailable; the separate model had the same availability but no majority-size or shift setting. Within the same 161 available sources, boundary rejection fell from 67/161 to 7/161, a paired difference of -0.373 (95% Monte Carlo interval, -0.448 to -0.297). The corrected conditional rejection rate, 7/161, differs from the affirmative issuance frequency over all planned sources, 7/200. Boundary truth coverage was 154/161 among available procedures; allowing the 39 unknown outcomes to be either uncovered or covered gives a planned-source coverage range of 0.770–0.965. This logical range is not a confidence interval. Neither low harmful certification nor withholding unavailable decisions demonstrates general nominal coverage.

Fixed shrinkage and structural misspecification retained signed errors

Source refitting improved the deliberately biased configurations as well, but did not remove all incorrect decisions. With penalty 10, majority size 3,200 and shift -1 , the shared main-effects generator's harmful-negative certifications fell from 159/200 to 59/200; all 200 bounds were available. At shift $+1$, harmful-positive certifications fell from 144/200 to 36/200, while boundary rejection fell from 196/200 to 127/200. This latter configuration had 199 available bounds, so its conditional corrected frequencies were 36/199 and 127/199. The planned harmful issuance rates were 29.5% (95% interval, 23.28–36.34%) at shift -1 and 18.0% (12.94–24.04%) at shift $+1$. Both orientations are required to describe the signed effect.

Residual decisions were not entirely due to synthetic-draw noise. For the exact fitted-generator mean at shift $+1$, the same main-effects configuration produced 37 harmful-positive certifications and 130 boundary rejections per 200 planned sources, with 199 evaluated procedures. These analytic-mean results are distinct from the finite-release counts of 36 and 127. Strong penalization also affected the correctly specified interacted model: at shift $+1$ its finite-release boundary rejection was 33/200, although harmful-positive certification fell from 68/200 before correction to 3/200 afterward. Weaker interacted penalization gave much smaller residual counts. The strongest findings therefore concern specified combinations of shrinkage and misspecification, rather than a general failure of source-aware inference.

Frozen marginal-error diagnostics on independent sources

The independent diagnostic completed 1,000 fresh source cohorts, with no unavailable estimates across the 785 fixed cells. Reference error distributions had been frozen using the earlier 1,000 sources before the fresh sources were generated. In the strongly penalized main-effects model ($\lambda = 10$) with majority size 3,200 and shift $+1$, the boundary rejection count for a 20,000-record release was 646/1,000 under

the uncentered fixed-reference normal bound. Subtracting the frozen mean error while retaining the same reference scale reduced it to 63/1,000; the empirical error-quantile bound gave 64/1,000. Pointwise 95% exact binomial intervals were 61.55–67.57%, 4.87–7.99% and 4.96–8.10%, respectively. These are paired evaluations in fresh sources, not additional bootstrap replicates.

The harmful comparisons showed corresponding signed effects. In the same shift + 1 configuration, harmful-positive certification fell from 237/1,000 to 6/1,000 after centering and remained 6/1,000 with the empirical quantile. At shift – 1, harmful-negative certification changed from 288/1,000 to 0/1,000 under both centered and quantile bounds. Centering also changed useful certification: the shift + 1 negative-offset and positive-offset beneficial reversals changed from 941/1,000 and 5/1,000 to 312/1,000 and 325/1,000, respectively. Thus, the frozen-reference location adjustment reduced one directional excess while restoring useful decisions in the opposite orientation; it was not a uniform reduction in every certification frequency.

The correctly specified unpenalized main-effects control at shift 0 and majority size 3,200 behaved differently. Boundary counts were 54/1,000 before centering, 54/1,000 after centering and 51/1,000 with the empirical quantile. Across all fixed boundary cells, centered-normal rejection counts ranged from 40 to 69 per 1,000 sources, compared with 38 to 71 for empirical quantiles. These dependent, pointwise summaries do not establish uniform 5% control. The intervals condition on the frozen reference and do not include uncertainty from estimating its mean, scale or quantile.

Known-truth neural findings remain feasibility evidence

The known-truth CTGAN pilot completed 168 fits across 24 source cohorts at 600 optimizer updates per fit. All 15 pointwise intervals for mean majority-size contrasts included zero. Composition contrasts differed, with 25 of 30 pointwise intervals excluding zero; these included dependent reversed comparisons and no multiplicity adjustment. Substantial distortion also occurred in the minority-only control. Its negative-offset harmful certifications rose from 17/24 to 21/24 as release size increased from one-fold to 100-fold, whereas the opposite harmful counts were 0/24 and 1/24. At 100-fold expansion the beneficial reversals were certified in 1/24 and 23/24. These asymmetric findings cannot be attributed solely to majority sharing.

All 168 fits recorded generated values outside the simulated Age/BMI support; 87 missed the absolute integration-MCSE target. Neither a small relative numerical error nor a completed fit establishes fidelity. The bounded training and balancing diagnostic therefore preceded any additional source grid.

A bounded diagnostic reused one prespecified source cohort and fitting seed, increasing training from 600 to 6,000 updates and evaluating balanced shared fits. All 13 new fits completed. Training reduced the groupwise retained-release support violation from 3,427/20,000 (17.135%) to 217/20,000 (1.085%) and substantially reduced its paired-loss deviations. These retained-row measurements differ from the earlier pilot's saved support counts before prefix truncation. Improvements were not uniform across majority shifts: absolute deviations for the two harmful comparisons and the boundary improved in four

of seven budget-matched cells and worsened in three. The reversed comparisons are dependent views of these same comparisons.

Balancing duplicated the existing minority records rather than adding new observations. It raised recorded minority exposure to 48.83–50.71% of training exposures, but changed empirical weighting, preprocessing and conditional sampling as well. It therefore did not isolate exposure alone. Balancing did not consistently improve minority outcome fidelity or the fitted comparison. At $\kappa = 0$, balanced fits generated minority outcome rates of 0.20351 and 0.04714 at majority sizes 800 and 3,200, respectively, from the same minority source with observed outcome rate 0.12. The corresponding fitted boundary differences were -0.01654 and $+0.02331$, despite a fixed equal-risk population comparison. These individual fitted quantities are not population error-rate estimates.

Three of the thirteen new fits missed the absolute integration-precision target; the seven matched older fits contained three further misses. All were retained. Because the diagnostic included one source and one fitting seed and did not establish generator adequacy, the known-truth neural results remain negative or inconclusive feasibility evidence. No additional source grid was launched to obtain more precise estimates for this configuration.

Completed public benchmark: decisions, fidelity and uncertainty

Figure 5 presents both comparison orientations and source-level fidelity diagnostics.

All 120 planned fits completed across 24 original source cohorts, with no fitting failure or unavailable matched-source cell. Twenty-three fits did not meet the absolute numerical integration-precision target; none was excluded. No invalid or fractional questionnaire-day values were observed in the returned streams of the categorical-day, 6,000-update run, but this did not establish distributional fidelity. Among 20.2 million returned release and integration rows, 45 had nonpositive BMI and 6,403 fell outside the BMI range of the fixed evaluation population. These are separate, potentially overlapping diagnostics. The original continuous-day pilot, including its severe support failure, remains part of the evidence history.

Across shared settings, equal-weight mean generated mental-health days ranged from 13.13 to 16.66, compared with 7.389 in the finite minority population; physical-health days ranged from 18.63 to 20.77, compared with 9.969. Minority-only means were 7.220 and 10.406. The matched real sources averaged 7.684 and 10.354, respectively. Across individual fits, generated-minus-own-source diabetes prevalence ranged from -0.23050 to $+0.39509$, with mean absolute difference 0.11034. These marginal checks do not fully characterize joint fidelity, but they show why valid day categories cannot establish adequate synthesis.

The natural model comparison favoured gradient boosting over logistic regression in the fixed minority evaluation population, with mean paired Brier difference -0.00226287 . In the minority-only generator, conditional normal certification of that beneficial comparison rose from 4/24 sources for 200 generated

rows to 17/24 for 20,000. Certification of its genuinely inferior reverse rose from 0/24 to 4/24. Across the four shared-generator settings, the beneficial natural comparison changed from 8–13 certificates at the smaller release size to 16–22 at the larger size; the harmful reverse changed from 0–4 to 2–7. These are ranges across specified settings, not independent replications or simultaneous confidence statements.

The constructed stress comparisons showed marked directional differences under conditional normal bounds. At 20,000 generated rows, the negative-logit harmful candidate was certified in 11/24 minority-only sources and 14–19/24 shared sources; its beneficial reverse was certified in 12/24 and 3–10/24, respectively. The positive-logit harmful candidate was certified in 8/24 minority-only sources and 2–7/24 shared sources, while its beneficial reverse was certified in 16/24 and 16–22/24. Expansion was not uniformly harmful: the negative-logit harmful count in the shared setting with 3,200 majority records and older share 0.25 decreased from 20/24 to 19/24. Original-source normal bounds certified none of the three harmful directions, but only 1/24, 2/24 and 1/24 of the corresponding beneficial directions. Low observed false certification therefore coexisted with low useful power; zero events in 24 sources cannot demonstrate nominal 5% control.

The bounded-loss comparator was more conservative but did not eliminate every false decision against the finite target. At majority size 3,200 and older share 0.75, with 20,000 generated records, the negative-logit harmful comparison received 10/24 conditional Hoeffding certificates, compared with 19/24 conditional normal certificates. Original-source Hoeffding bounds issued no certificates in any of the six directions, including the beneficial comparisons. Thus, conditional concentration and normal approximation behaved differently, while neither supplied new observed-source information. Complete method-specific rates, all six orientations, source denominators and paired contrasts are retained in Additional file 5.

All twelve pointwise intervals for composition contrasts in fitted mean loss included zero. Increasing majority size from 800 to 3,200 also gave intervals including zero for the natural gradient-boosting-minus-logistic-regression comparison: -0.003364 (95% Monte Carlo interval, -0.010335 to $+0.003607$) at older share 0.25 and -0.005494 (-0.012758 to $+0.001769$) at share 0.75. This is absence of clear evidence in 24 sources, not an equivalence result. At share 0.75, the negative-logit harmful stress contrast was -0.009768 (-0.017603 to -0.001933), whereas the positive-logit harmful contrast was $+0.010756$ ($+0.002352$ to $+0.019161$). Their exact reversals also excluded zero. These four nonzero intervals represent two dependent stress-family findings under exploratory pointwise inference, with no multiplicity adjustment.

All 24 fitted-estimand intervention summaries had aggregate conditional-MCSE bounds below 10% of their source-level standard errors, but eight failed the stricter per-constituent requirement. The 23 fit-level absolute precision misses remained relevant. Within-fit continuous expansion and expansion-interaction means had intervals including zero and failed the relative numerical-precision check. Under a fixed fitted distribution these nested mean contrasts have zero conditional expectation; they are numerical-draw

diagnostics, not evidence that release size changes the fitted target. Binary certification decisions have a different sampling distribution, and one nested release does not identify their conditional Monte Carlo uncertainty.

Discussion

The experiments distinguish uncertainty in the original source sample, finite-sample failure of estimated uncertainty and distortion of the fitted-generator target. A larger synthetic release reduces conditional draw noise. Its inferential value depends on whether the analysis also addresses the source and model errors relevant to the target. Conditional exactness alone does not supply that assurance.

The parametric controls provide the clearest separation. The unpenalized fully interacted generator was insensitive to majority size, yet its release-based decisions changed markedly with synthetic expansion. Under shared restrictions, changing majority data could separately move the fitted minority comparison in either direction. Of 90 mean-drift cells, only the main-effects configuration with $\lambda = 10$, $\kappa = +1$ and the equal-risk pair exceeded the absolute 0.001 screening scale. That scale was an experimental criterion, not a clinically important difference. The largest shifts combined deliberate misspecification and fixed strong shrinkage; they cannot establish a general failure of appropriately tuned shared generators.

Source refitting is a successful control in this study. In correctly specified settings it removed much of the excess certification, while reducing useful certification to a scale similar to the source-only comparison. In some configurations, numerical failures made bounds unavailable; these unknown outcomes cannot be counted as successful coverage. Residual errors under fixed shrinkage or omitted group structure are compatible with established limitations of variance-only correction. At a common frozen reference scale, subtracting the reference mean error removed large directional distortions on fresh sources. This supports a role for marginal systematic error in the specified population. Because the diagnostic replaced source-specific bootstrap scales, it does not identify bias as the sole cause of the earlier residual errors. It also uses known truth and cannot be deployed as written when that truth is unavailable.

The sparse baseline identifies a separate problem: an estimated source-variance component may collapse to zero when the subgroup contains no observed events. An exact source bound [45] avoids this particular false certainty, but can have no useful power at the available sample size. An unresolved comparison can accurately reflect inadequate information. Bounded-loss concentration and modern confidence procedures provide other source-level options [48, 49, 50], whose assumptions and power must be assessed for the intended estimand.

The public experiment extends the analysis to frozen, development-trained predictors on observed records, but its target remains a finite unweighted table. The natural comparison favored gradient boosting. Synthetic decisions nevertheless included both beneficial certifications and false certifications of the reversed comparison. Its majority-size intervals included zero, while two constructed stress families showed opposite signed changes under pointwise inference. These findings neither establish

invariance nor identify a universal harmful majority effect. Reversed comparisons are dependent, and the 24 source cohorts provide limited precision.

Neural fidelity changes the interpretation materially. Longer training improved support in the known-truth minority-only control, but improvements across cells were uneven. In the completed public experiment, allowed questionnaire-day values coexisted with distorted day distributions and outcome prevalence. Equal optimizer updates did not equalize minority exposure, conditional-category frequencies or optimization difficulty. No prespecified adequacy gate was passed, and no public source-refit or multi-initialization analysis was performed. Consequently, these results describe the fitted CTGAN procedures; they do not show that a generator of established adequate quality fails because of majority sharing. Additional replication of an unqualified configuration would narrow Monte Carlo intervals without resolving that substantive limitation.

The contribution is a controlled comparison of mechanisms and their limits. The constant-predictor example is an affine transformation of a binomial problem. Multiple-synthesis rules, Bayesian compatibility, generative-bootstrap theory, synthetic-assisted guardrails and task exchangeability already address related inferential questions under specified information structures [13, 14, 15, 20, 21, 26, 27, 31]. Their guarantees cannot be refuted outside their conditions. GESPI and AutoEval were not benchmarked. Comparisons involving an extra error allowance would require a source-only benchmark with the same allowance. The successful controls and bounded failures are useful precisely because they distinguish what synthetic expansion changes from what it leaves unresolved.

Several further limitations restrict applicability. The baseline uses constant predictors and a binary event mechanism; the parametric populations are manufactured and the ridge penalties are fixed. A 0.002 Brier difference has no established clinical importance here. Although beneficial alternatives and both harmful directions were retained, some source procedures had little power. The public label and provenance limitations, coarse-profile overlap, single derivative dataset and unweighted sampling prevent claims about incident disease, national representation or external clinical validity. Pointwise intervals are not simultaneous guarantees over the many dependent comparisons. No clinical intervention, deployed model or privacy mechanism was evaluated. Privacy assessment requires its own threat models and controls [51], and source-size and generator-quality evaluations have direct precedents [52]. Synthetic data can still serve other justified purposes; fidelity, utility and privacy require separate evidence.

Conclusions

Synthetic validation should report the original subgroup information, the fitted target and uncertainty from both source fitting and conditional generation. In the executed controls, source refitting substantially reduced conditional-release overconfidence, with an accompanying power cost. Fixed bias and sparse plug-in failures required additional scrutiny. The public benchmark showed both useful and false conditional decisions, but unresolved fidelity prevented a claim about adequately trained

generators or an isolated majority-sharing mechanism. These limits are part of the result: the number of generated records alone does not establish evidence for lower population risk.

Declarations

Ethics approval and consent to participate

The simulation experiments involved no participants. The secondary analysis used a freely available CC0 survey derivative, without participant recruitment, access to identified records or re-identification. Ethical clearance was not required for this public-domain secondary analysis under section 4.1.4 of the King's College London Research Ethics Policy. No project-specific ethics approval or exemption certificate was obtained.

Consent for publication

Not applicable to the simulated results; no identifiable individual information is reported in the public benchmark.

Competing interests

The author declares no competing interests.

Clinical trial number

not applicable.

Funding

This work received no funding.

Author Contribution

TY conceived the study, wrote the initial manuscript and code, prepared the figures and document layout, and developed and revised the manuscript. The author read and approved the final manuscript. Computational assistance is described in Methods.

Acknowledgements

Not applicable.

Data Availability

The public input is the CC0 CDC Diabetes Health Indicators derivative [47]. Additional file 1 describes the executed methods. Additional files 2–5 contain code, protocol and environment records, completed result tables, figure source maps and the complete public benchmark tables. Additional file 6 contains the frozen-oracle holdout arrays; Additional file 7 contains the public-data setup, frozen records, predictors and provenance; Additional files 8–11 contain lossless baseline source and release counts. Archive manifests record file paths, sizes and SHA-256 hashes. These accompanying files reproduce the aggregate comparisons and figures with the stated dependencies. They do not include every parametric source draw, bootstrap refit, neural checkpoint or generated row. Those complete raw artifacts are retained in the project archive; no public repository deposit or persistent identifier is claimed. The original download locations, supplied-label mapping and unresolved source-year discrepancy are recorded with the public data.

Abbreviations

BMI, body mass index; CTGAN, conditional tabular generative adversarial network; HGB, histogram gradient boosting; LR, logistic regression; MCSE, Monte Carlo standard error; UCI, University of California, Irvine.

References

1. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, Mooney SD, Malin BA. A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun.* 2022;13(1):7609. <https://doi.org/10.1038/s41467-022-35295-1>.
2. Theodorou B, Xiao C, Sun J. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nat Commun.* 2023;14(1):5305. <https://doi.org/10.1038/s41467-023-41093-0>.
3. Liang ZJ, Li Z, Jackson NJ, Guo KH, Caro-Vega Y, Moreira RI, Paredes F, Bernadin J, Varela D, Cesar C, Blasimme A, Perkins JM, Asiaee A, Duda SN, Malin BA, Shepherd BE, Yan C. Generating synthetic multi-national longitudinal cohorts for clinically grounded HIV research. *Nat Commun.* 2026;17(1):7857. <https://doi.org/10.1038/s41467-026-74492-0>.
4. Sella N, Guinot F, Lagrange N, Albou L-P, Desponds J, Isambert H. Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. *npj Digit Med.* 2025;8(1):49. <https://doi.org/10.1038/s41746-025-01431-6>.
5. Yoon J, Mizrahi M, Ghalaty NF, Jarvinen T, Ravi AS, Brune P, Kong F, Anderson D, Lee G, Meir A, Bandukwala F, Kanal E, Arık SÖ, Pfister T. EHR-Safe: generating high-fidelity and privacy-preserving

- synthetic electronic health records. *npj Digit Med*. 2023;6(1):141. <https://doi.org/10.1038/s41746-023-00888-7>.
6. Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digit Med*. 2023;6(1):98. <https://doi.org/10.1038/s41746-023-00834-7>.
 7. Marino S, Cassidy R, Nanni J, Wang Y, Liu Y, Tang M, Yuan Y, Chen T, Sinha A, Pandian B, Dinov ID, Burns ML. *npj Digit Med*. 2025;8(1):526. <https://doi.org/10.1038/s41746-025-01935-1>. Medical data sharing and synthetic clinical data generation – maximizing biomedical resource utilization and minimizing participant re-identification risks.
 8. Hernandez M, Osorio-Marulanda PA, Catalina M, Loinaz L, Epelde G, Aginako N. Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Front Digit Health*. 2025;7:1576290. <https://doi.org/10.3389/fdgth.2025.1576290>.
 9. Kurakova A, Homayouni H. A comprehensive evaluation framework for synthetic medical tabular data generation. *J Biomed Inform*. 2025;171:104939. <https://doi.org/10.1016/j.jbi.2025.104939>.
 10. Adams T, Birkenbihl C, Otte K, Ng HG, Rieling JA, Näher A-F, Sax U, Prasser F, Fröhlich H. Alzheimer's Disease Neuroimaging Initiative. On the fidelity versus privacy and utility trade-off of synthetic patient data. *iScience*. 2025;28(5):112382. <https://doi.org/10.1016/j.isci.2025.112382>.
 11. Qian Z, Callender T, Cebere B, Janes SM, van der Navani N. Synthetic data for privacy-preserving clinical risk prediction. *Sci Rep*. 2024;14(1):25676. <https://doi.org/10.1038/s41598-024-72894-y>.
 12. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*. 2024;5(4):100946. <https://doi.org/10.1016/j.patter.2024.100946>.
 13. Raab GM, Nowok B, Dibben C. Practical Data Synthesis for Large Samples. *J Priv Confidentiality*. 2018;7(3):67–97. <https://doi.org/10.29012/jpc.v7i3.407>.
 14. Reiter JP. Inference for Partially Synthetic, Public Use Microdata Sets. *Surv Methodol*. 2003;29(2):181–8. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6785-eng.pdf>.
 15. Reiter JP, Kinney SK. Inferentially Valid, Partially Synthetic Data. Generating from Posterior Predictive Distributions Not Necessary. *J Official Stat*. 2012;28(4):583–90. <https://www.scb.se/contentassets/ca21efb41fee47d293bb5bf7be7fb3/inferentially-valid-partially-synthetic-data-generating-from-posterior-predictive-distributions-not-necessary.pdf>.
 16. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and Specific Utility Measures for Synthetic Data. *J Royal Stat Society: Ser (Statistics Society)*. 2018;181(3):663–88. <https://doi.org/10.1111/rssa.12358>.
 17. Nowok B, Raab GM, Dibben C, synthpop. Bespoke Creation of Synthetic Data in R. *J Stat Softw*. 2016;74(11). <https://doi.org/10.18637/jss.v074.i11>.
 18. Decruyenaere A, Dehaene H, Rabaey P, Polet C, Decruyenaere J, Vansteelandt S, Demeester T. The Real Deal Behind the Artificial Appeal: Inferential Utility of Tabular Synthetic Data. Uncertainty in

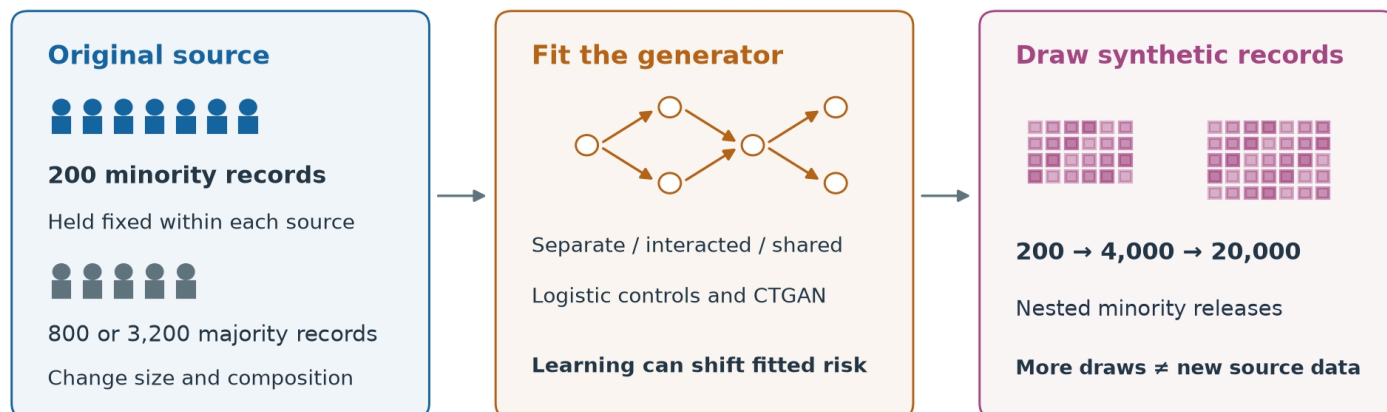
- Artificial Intelligence. 2024;244:966–996.
<https://proceedings.mlr.press/v244/decruyenaere24a.html>
19. Decruyenaere A, Dehaene H, Rabaey P, Polet C, Decruyenaere J, Demeester T, Vansteelandt S. Adv Neural Inf Process Syst. 2024;37:41539–76. <https://doi.org/10.52202/079017-1315>. Debiasing Synthetic Data Generated by Deep Generative Models.
 20. Räisä O, Jälkö J, Honkela A. On Consistent Bayesian Inference from Synthetic Data. J Mach Learn Res. 2025;26(74):1–65. <https://www.jmlr.org/beta/papers/v26/23-1428.html>.
 21. Nombo L, Charest A-S. Inference with Combining Rules from Multiple Differentially Private Synthetic Datasets. Statistica Sinica. 2025;585–612. <https://doi.org/10.5705/ss.202022.0299>. 35(Online Special Issue I).
 22. Filatov G, Deeva I. When Synthetic Data Is Enough: Calibration for Tabular Model Ranking. Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI (AISTATS 2026 workshop). 2026. <https://openreview.net/forum?id=YECegW8nBY>
 23. Shoshan A, Bhonker N, Kviatkovsky I, Fintz M, Medioni G. Synthetic data for model selection. International Conference on Machine Learning. 2023;202:31633–31656.
<https://proceedings.mlr.press/v202/shoshan23a.html>
 24. Boyeau P, Angelopoulos AN, Li T, Yosef N, Malik J, Jordan MI. AutoEval Done Right: Using Synthetic Data for Model Evaluation. International Conference on Machine Learning. 2025;267:5276–5290.
<https://proceedings.mlr.press/v267/boyeau25a.html>
 25. Angelopoulos AN, Bates S, Fannjiang C, Jordan MI, Zrnic T. Prediction-powered inference. Science. 2023;382(6671):669–74. <https://doi.org/10.1126/science.adi6000>.
 26. Tran L, Ye T, Ding P, Han F. Generative modeling for the bootstrap. arXiv. 2026.
<https://arxiv.org/abs/2602.17052>
 27. Tan L, Zrnic T. Valid Inference with Synthetic Data via Task Exchangeability. arXiv. 2026.
<https://arxiv.org/abs/2606.13629>
 28. Bates S, Angelopoulos A, Lei L, Malik J, Jordan M. Distribution-free, Risk-controlling Prediction Sets. J ACM. 2021;68(6):1–34. <https://doi.org/10.1145/3478535>.
 29. Angelopoulos AN, Bates S, Candès EJ, Jordan MI, Lei L. Learn then test: Calibrating predictive algorithms to achieve risk control. Annals Appl Stat. 2025;19(2):1641–62.
<https://doi.org/10.1214/24-AOAS1998>.
 30. Angelopoulos A, Bates S, Fisch A, Lei L, Schuster T. Conformal Risk Control. International Conference on Learning Representations. 2024;2024:55198–55218.
https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html
 31. Bashari M, Lee Y, Lotan RM, Dobriban E, Romano Y. General Synthetic-Powered Inference. arXiv:2509.20345v3; ICML 2026 conference work. 2026. <https://arxiv.org/abs/2509.20345v3>
 32. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, Snell KIE. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med.

- 2021;40(19):4230–51. <https://doi.org/10.1002/sim.9025>.
33. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, Bonnett LJ, Collins GS. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820. <https://doi.org/10.1136/bmj-2023-074820>.
 34. Hebert-Johnson U, Kim M, Reingold O, Rothblum G, Multicalibration. Calibration for the (Computationally-Identifiable) Masses. *International Conference on Machine Learning*. 2018;80:1939–1948. <https://proceedings.mlr.press/v80/hebert-johnson18a>
 35. Cherian JJ, Candès EJ. Statistical Inference for Fairness Auditing. *J Mach Learn Res*. 2024;25(149):1–49. <https://jmlr.org/papers/v25/23-0739.html>.
 36. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix A-L, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.
 37. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, Beam AL, Van Calster B, Celi LA, Denaxas S, Denniston AK, Ghassemi M, Heinze G, Kengne AP, Maier-Hein L, Liu X, Logullo P, McCradden MD, Liu N, Oakden-Rayner L, Singh K, Ting DS, Wynants L, Yang B, Reitsma JB, Riley RD, Collins GS, van Smeden M. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. <https://doi.org/10.1136/bmj-2024-082505>.
 38. Brier GW, VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Mon Weather Rev*. 1950;78(1):1–3. [https://doi.org/10.1175/1520-0493\(1950\)078%3C0001:vofeit%3E2.0.co;2](https://doi.org/10.1175/1520-0493(1950)078%3C0001:vofeit%3E2.0.co;2).
 39. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *J Am Stat Assoc*. 2007;102(477):359–78. <https://doi.org/10.1198/016214506000001437>.
 40. Murphy AH. A New Vector Partition of the Probability Score. *J Appl Meteorol*. 1973;12(4):595–600. [https://doi.org/10.1175/1520-0450\(1973\)012%3C0595:anvpot%3E2.0.co;2](https://doi.org/10.1175/1520-0450(1973)012%3C0595:anvpot%3E2.0.co;2).
 41. Dimitriadis T, Gneiting T, Jordan AI. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*. 2021;118(8):e2016191118. <https://doi.org/10.1073/pnas.2016191118>
 42. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagn Prognostic Res*. 2017;1:19. <https://doi.org/10.1186/s41512-017-0020-3>.
 43. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006;26(6):565–74. <https://doi.org/10.1177/0272989x06295361>.
 44. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–102. <https://doi.org/10.1002/sim.8086>.

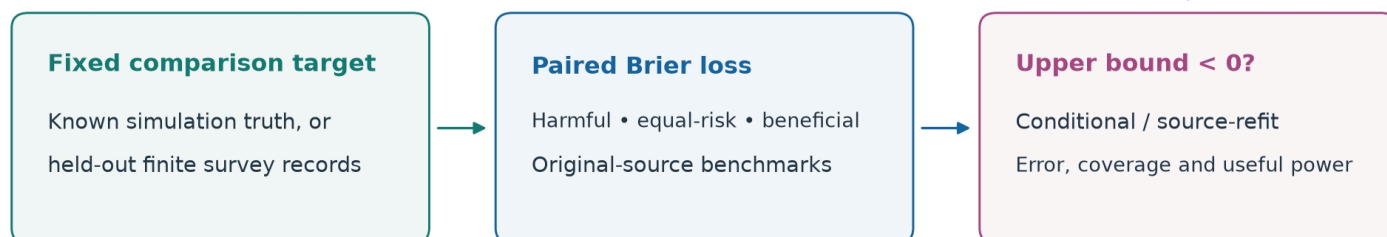
45. Clopper CJ, Pearson ES, THE USE OF CONFIDENCE OR FIDUCIAL LIMITS, ILLUSTRATED IN THE CASE OF THE BINOMIAL. *Biometrika*. 1934;26(4):404–13. <https://doi.org/10.1093/biomet/26.4.404>.
46. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems*. 2019;32. <https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html>
47. CDC Diabetes Health Indicators [Dataset]. UCI Machine Learning Repository. 2017. <https://doi.org/10.24432/C53919>
48. Hoeffding W. Probability Inequalities for Sums of Bounded Random Variables. *J Am Stat Assoc*. 1963;58(301):13–30. <https://doi.org/10.1080/01621459.1963.10500830>.
49. Maurer A, Pontil M. Empirical Bernstein Bounds and Sample Variance Penalization. *Proceedings of the 22nd Annual Conference on Learning Theory*. 2009. <https://www.learningtheory.org/colt2009/papers/012.pdf>
50. Waudby-Smith I, Ramdas A. Estimating means of bounded random variables by betting. *J Royal Stat Soc Ser B: Stat Methodol*. 2024;86(1):1–27. <https://doi.org/10.1093/jrsssb/qkad009>.
51. Giomi M, Boenisch F, Wehmeyer C, Tasnádi B. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *Proceedings on Privacy Enhancing Technologies*. 2023;2023(2):312–328. <https://doi.org/10.56553/popets-2023-0055>
52. Vu MH, Edler D, Wibom C, Rosvall M, Melin B. Anonymization and visualization of health data and biomarkers. *npj Digit Med*. 2026;9(1):347. <https://doi.org/10.1038/s41746-026-02662-x>.

Figures

a What supplies evidence for a minority comparison?

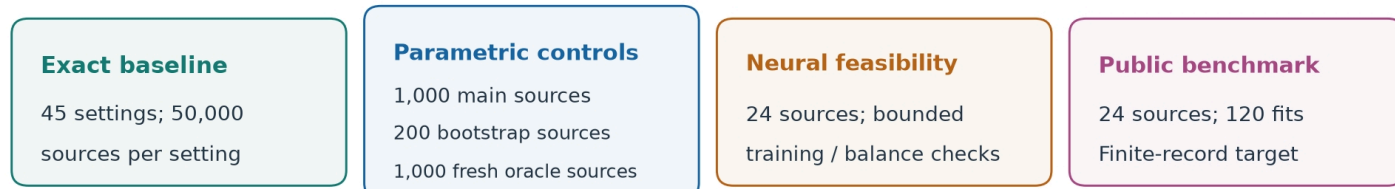


b Separate the statistical decision from its population target



Unit of uncertainty: independent original source cohort

c Linked evidence layers and limits



Source refitting can help; bias and fidelity require separate checks.

Oracle bounds use known truth. Neural and survey findings do not establish clinical validity.

Figure 1

Evidence flow for synthetic minority validation

Matched experiments hold minority records fixed while varying majority sources and release size. Icons and networks are schematic. The Bernoulli baseline separately varies subgroup size and includes a boundary null. Known-truth branches include equal-risk comparisons; public comparisons retain the natural model pair and fixed perturbations. Source-refit bounds are limited to parametric experiments. Bootstrap uses 200 original sources; oracle evaluation uses 1,000 fresh sources and known truth. The bounded neural follow-up uses one original source and one fitting-seed offset. Public sources sample an unweighted cross-sectional finite population with replacement. Independent source cohorts determine

uncertainty; fits, reversed comparisons and synthetic rows are not independent source replications. Fidelity and statistical validity require separate checks.

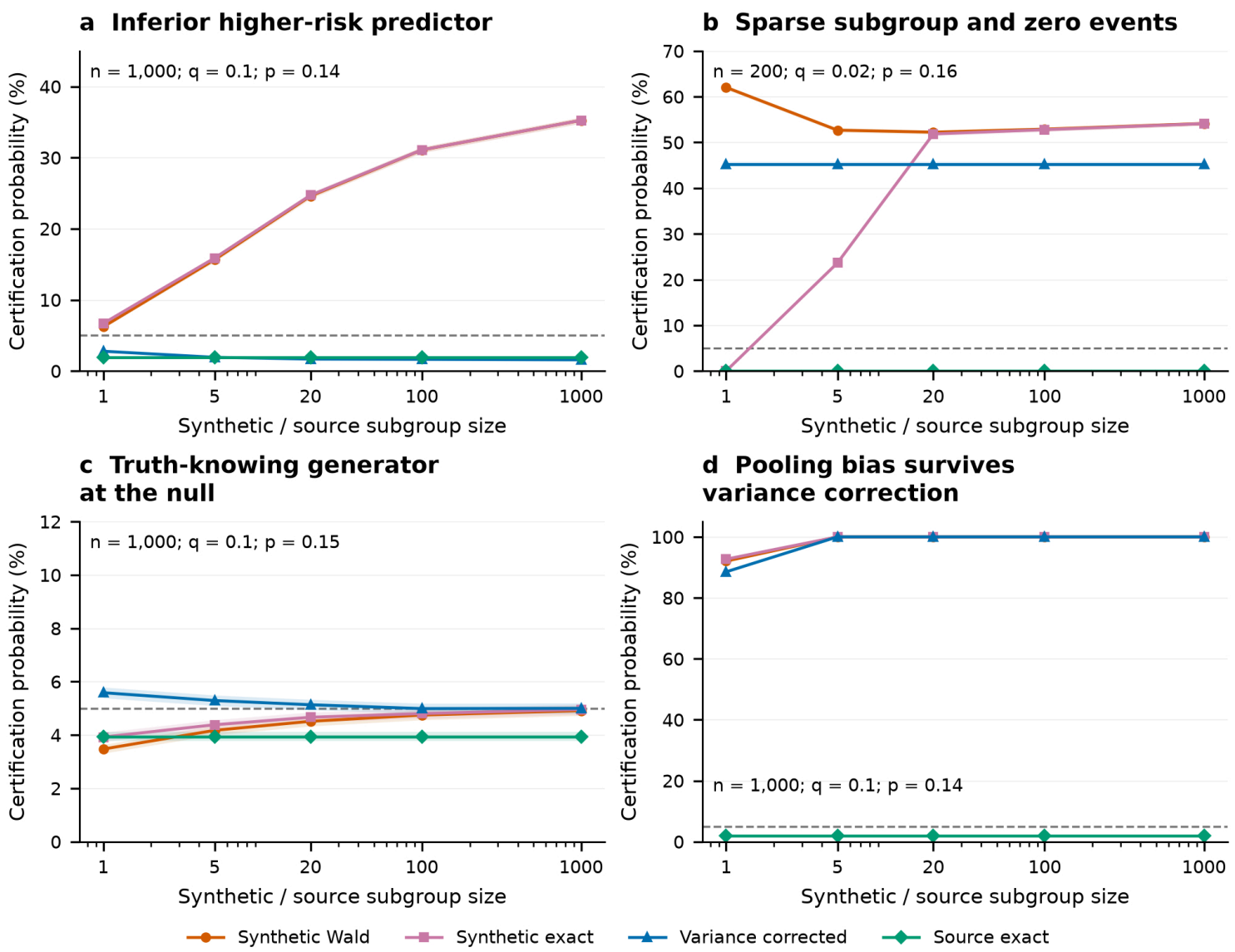


Figure 2
Mechanistic controls for certification from synthetic data

Across 50,000 source cohorts per scenario, curves show certification rates; shading gives pointwise 95% Wilson Monte Carlo intervals. Panels show an inferior higher-risk predictor with empirical generation (a), a sparse subgroup with the lower-prediction comparison and empirical generation (b), a truth-knowing generator at the boundary null (c), and an intentionally pooled generator (d). Here n is total source cohort size, q is subgroup prevalence, and p is the subgroup outcome probability. The horizontal axis is synthetic release size divided by source subgroup size; source-exact results are repeated as a constant benchmark, not independent evidence. Dashed lines mark the 5% reference. Colors and marker shapes identify the four inference procedures. Panel-specific vertical scales should be read separately. These known-truth Bernoulli controls distinguish omitted source uncertainty, sparse-sample Wald failure, and

deliberately imposed pooling bias; variance correction does not remove estimand bias. Intervals describe Monte Carlo uncertainty across source cohorts, not uncertainty within one fitted generator. No multiplicity adjustment is applied. Results do not establish clinical safety or general performance of modern generators.

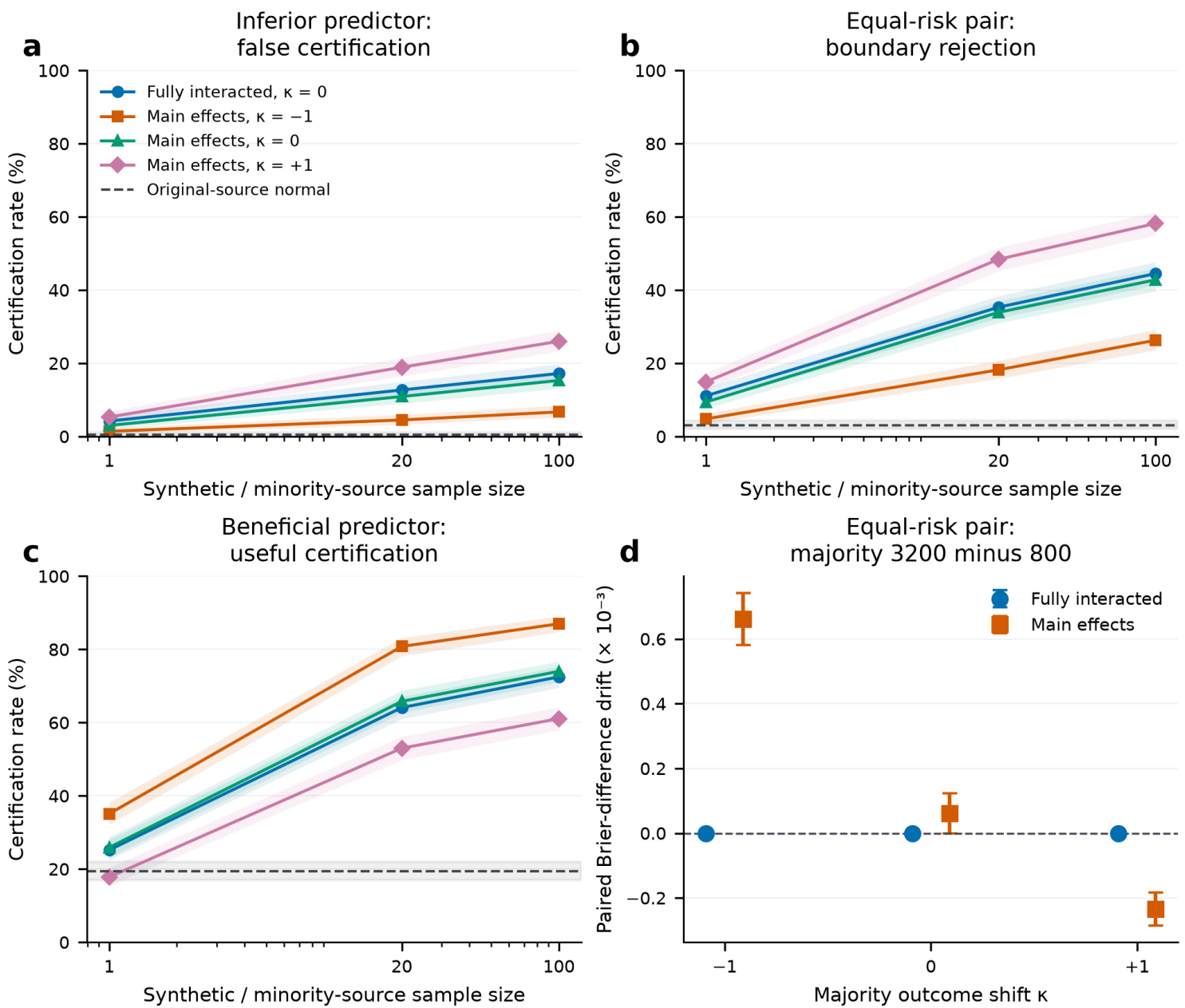


Figure 3

Conditional synthetic precision and majority-source dependence

The completed main parametric experiment uses 1,000 independent source cohorts and minority $n = 200$. All displayed fitted models use zero ridge penalty. Panels a–c hold majority $n = 3,200$; curves use conditional normal bounds on nested synthetic releases. Shading shows pointwise exact binomial 95% Monte Carlo confidence intervals across source cohorts, not confidence in a single fitted generator. The

dashed source-only benchmark uses the original minority paired losses. Panel a uses the positive-offset harmful predictor, panel c its reversed beneficial comparison; these are dependent views, not independent replications. Panel b uses the equal-risk pair. Panel d shows within-source changes in exact fitted-generator mean paired Brier loss when majority n increases from 800 to 3,200, for the equal-risk pair; intervals are pointwise t intervals over 1,000 paired cohorts. Fully interacted models are correctly specified; shared main-effects models omit the group-by-age effect and are deliberately misspecified for nonzero κ . No failed fits occur in plotted cells. No multiplicity adjustment is applied. Penalized sensitivities remain in the complete result tables. Bootstrap calibration and neural-generator findings are not represented here.

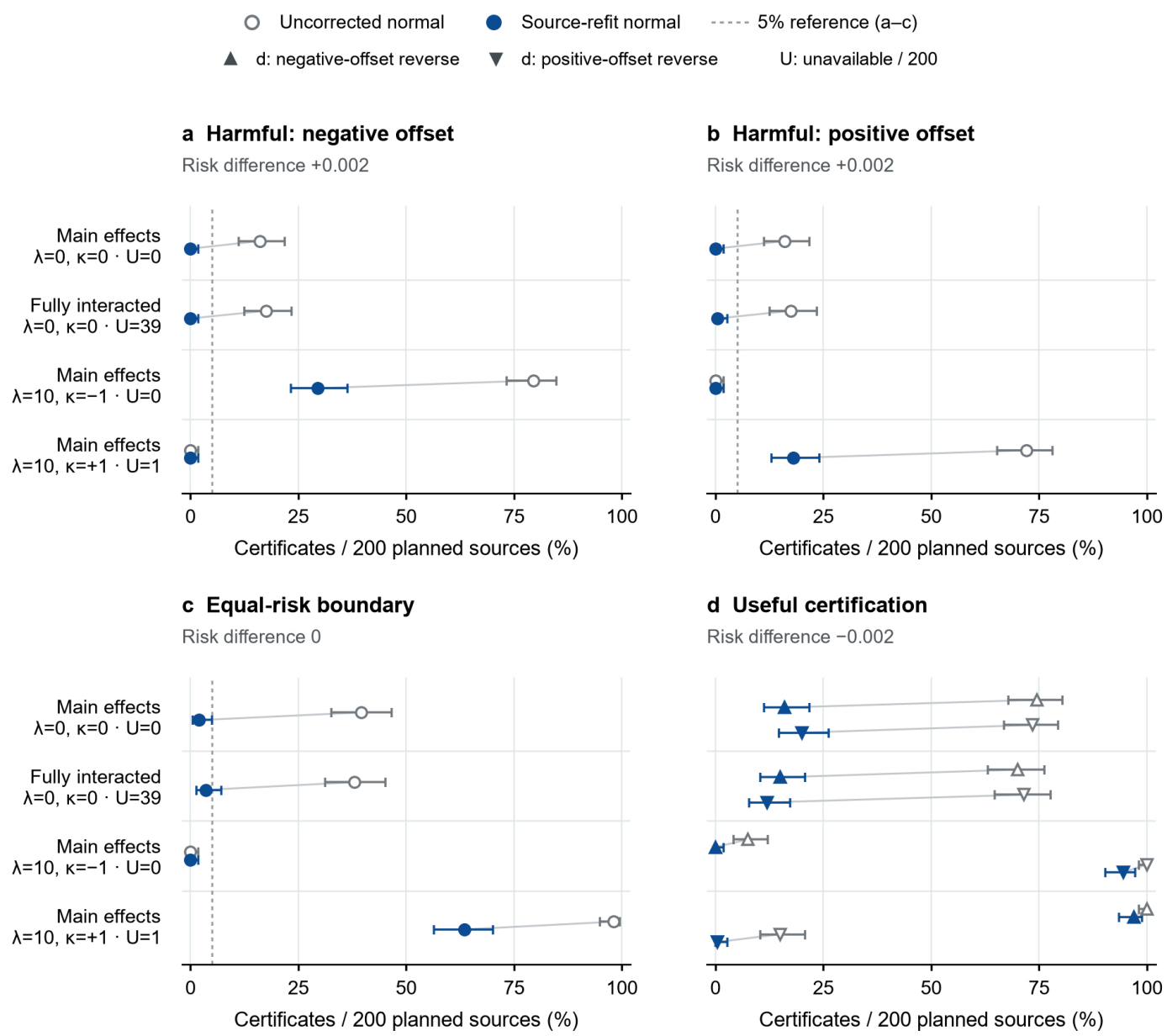


Figure 4

Source-refit correction on matched source cohorts

Points show certificates among 200 independent source cohorts (IDs 0–199), with 200 minority and 3,200 majority records per cohort and 20,000-record minority releases. Fitted estimates and releases are matched across stages. Whiskers are pointwise two-sided 95% exact binomial intervals for procedural issuance; connecting segments are not paired-change intervals.

Panels a–b show harmful candidates (excess Brier risk +0.002); c shows equal-risk boundary rejection; d shows beneficial reversals (–0.002). Comparisons share sources. Displayed configurations comprise unpenalized controls at $\kappa=0$ and fixed strong-penalty main-effects models at both nonzero shifts. Main-effects models omit a group-by-age term and are deliberately misspecified at $\kappa \neq 0$; $\lambda=10$ was not tuned.

Unavailable bounds remain unknown: interacted boundary issuance is 7/200 planned, versus 7/161 evaluated. Source refitting uses 399 resamples with nested majority blocks and adds finite-release variance without subtracting model bias. The 5% reference does not establish calibration. Complete tables retain other configurations, Hoeffding bounds and analytic-generator results. Neither universal improvement nor clinical validity is established.

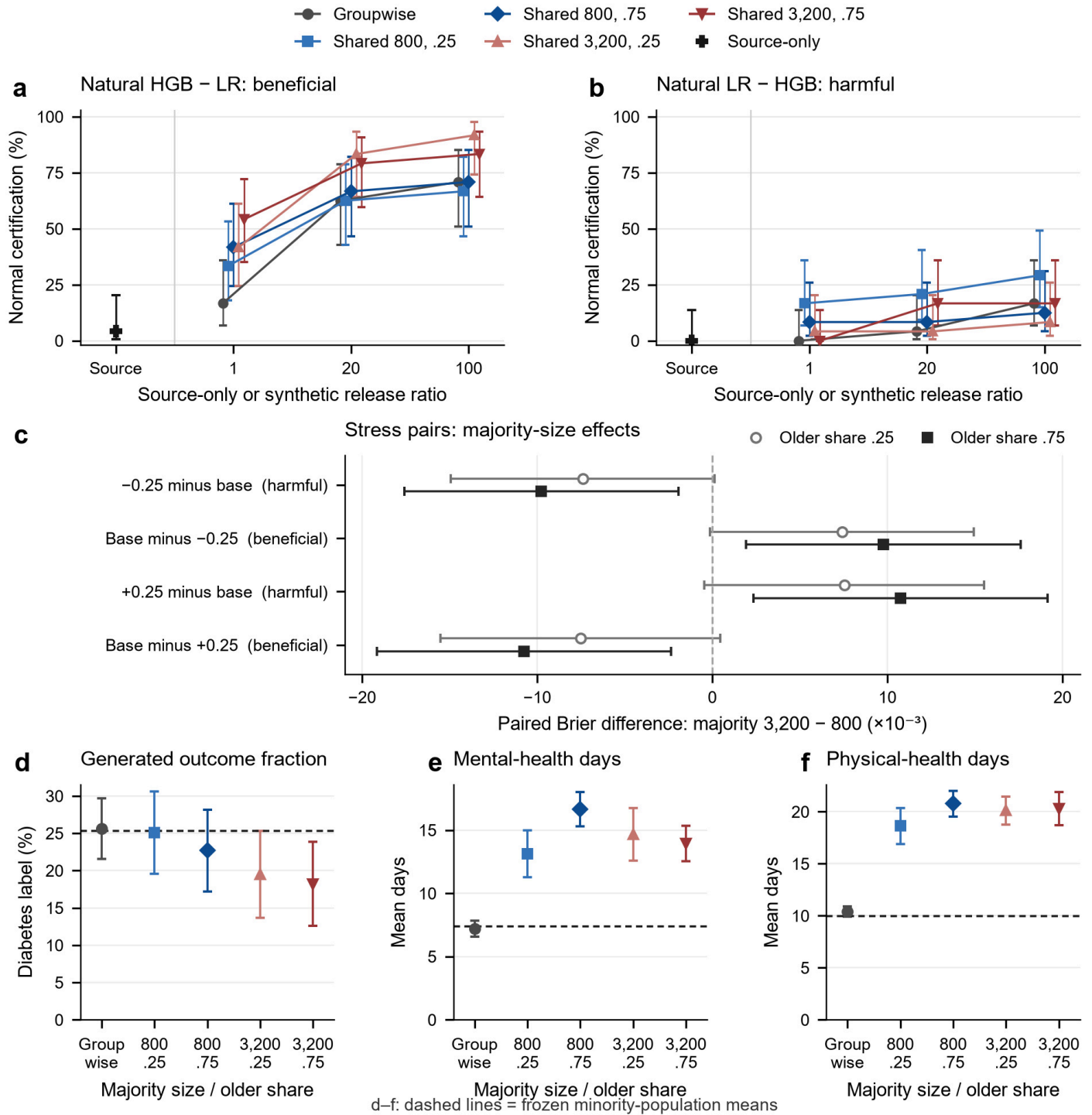


Figure 5

Completed public CTGAN study. The 120 fits span 24 source cohorts; all 23 integration-precision misses remain. (a,b) One-sided 95% normal-bound certification for the natural comparison and its reverse, with source-only decisions and release ratios 1, 20 and 100. Bars are pointwise 95% Wilson intervals across sources; harmful-direction certificates are false relative to the frozen target. (c) Within-source changes in fitted-generator paired Brier loss for majority 3,200 minus 800, showing all four logit-perturbation orientations. Bars are pointwise 95% paired-source t intervals; reversed orientations are dependent. (d-f) Generated minority outcome fractions and day-count means, averaged equally across 24 fits, with

pointwise 95% source-level t intervals. Dashed lines show frozen minority-table means. Source-level intervals include fitting and sampling variation. The target is the unweighted 12,957-record minority evaluation table. No multiplicity adjustment is applied; generator adequacy and source-error calibration are not established.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile01methods.docx](#)