Calibration-Aware Conditional Synthetic Augmentation under Explicit Feasibility Constraints: Repeated Validation on Mobile and Sensor Miscarriage-Risk Records

Sadia Saba[a,*], Muhammad Amir Saeed[a], Antonio Candelieri[a]

[a]*Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy*

Abstract

Synthetic tabular data are increasingly used for augmentation and privacy-oriented data access, but their value for probabilistic risk prediction cannot be inferred from distributional resemblance or discrimination alone. We evaluated a calibration-aware conditional synthetic augmentation framework for miscarriage-risk prediction from repeatedly collected mobile and healthcare-sensor records. A conditional Wasserstein generative adversarial network with gradient penalty was conditioned on the binary outcome and prespecified age and body-mass-index strata. Its objective combined adversarial learning with moment matching, correlation preservation, raw-domain feasibility, and label-consistency penalties. Generative and predictive hyperparameters were selected on a dedicated tuning partition using constrained Bayesian optimization and then locked for 20 repeated held-out evaluations. In each repetition, elastic-net logistic models were fitted to real records alone or to real plus 4,000 synthetic records. Recalibration was estimated only from a separate real-data calibration partition, and all metrics were evaluated on held-out real records. Synthetic augmentation did not change discrimination: mean AUC was 0.641 for both training strategies, with a paired difference of 0.000 (0.000, 0.000). It produced small improvements in Brier

*Corresponding author

Email addresses: s.saba5@campus.unimib.it (Sadia Saba), m.saeed@campus.unimib.it (Muhammad Amir Saeed), antonio.candelieri@unimib.it (Antonio Candelieri)

score (-0.001), log loss (-0.002), absolute calibration-slope error (-0.248), and 10-bin equal-width ECE (-0.071). Tie-preserving 10-bin ACE decreased by -0.007, but its resampling interval included zero (-0.016 to 0.002; adjusted Wilcoxon $p = 0.751$). Differences largely disappeared after Platt or temperature calibration. No Bayesian-optimization candidate satisfied all feasibility constraints; mean classifier two-sample-test AUC was 0.612, mean pre-projection domain-violation rate was 87.5%, and the nearest-neighbour memorization proxy was 0.022%. Conditional augmentation therefore acted as a modest regularizer for the uncalibrated predictor, but it neither improved ranking nor yielded a generator that naturally respected the prespecified feature domain. Because the data comprise repeated records and were split at record level, the findings are methodological and do not establish clinical deployment readiness.

Keywords: synthetic tabular data, probability calibration, conditional WGAN-GP, constrained Bayesian optimization, mobile sensing, repeated validation

1. Introduction

Clinical and health-related risk prediction is not only a ranking problem. A model may separate observations with higher and lower outcome risk while assigning probabilities that are systematically too high, too low, or excessively dispersed. Discrimination and calibration consequently answer different questions: the area under the receiver operating characteristic curve (AUC) assesses ranking, whereas calibration assesses whether predicted probabilities agree with observed event frequencies. Proper scoring rules, including the Brier score and log loss, evaluate the complete probability forecast and are therefore essential when model outputs may be interpreted as risks rather than arbitrary scores [1, 2].

The distinction is particularly important for data collected through mobile phones and healthcare sensors. These records combine comparatively stable characteristics, such as age, body mass index, and reproductive history, with repeated measurements of physical activity, temperature, heart rate, stress, blood pressure, location, and alcohol-related behaviour. The public resource analysed in this study was developed for real-time miscarriage-risk research using mobile and Internet-of-Things measurements [3, 4, 5]. Its rows are measurement records rather than a documented set of independent par-

ticipants. This feature of the data affects the interpretation of sample size, resampling uncertainty, and leakage control.

Synthetic augmentation is often proposed when the observed predictor space is sparse, imbalanced, or unevenly represented. Generative adversarial networks learn an implicit distribution, and conditional variants enable generation according to outcomes or clinically meaningful strata [6, 7]. For health data, however, a generated record can be useful for a downstream classifier while remaining statistically distinguishable or clinically implausible; conversely, a highly realistic synthetic dataset may provide little predictive benefit. Modern evaluations therefore separate fidelity, utility, privacy risk, and task-specific performance rather than treating them as interchangeable properties [8, 9, 10, 11, 12].

Calibration adds a further ambiguity. An apparent calibration improvement after augmentation may reflect a useful regularization effect, but it may also disappear once a conventional calibration map is fitted to real data. A fair comparison must therefore evaluate real-only and augmented predictors under the same downstream regularization and the same post-hoc calibration procedure. It must also avoid relying on a single expected calibration error (ECE), because binned calibration estimators can be sensitive to the binning rule, the number of bins, finite-sample bias, and tied prediction scores [13, 14].

This study develops and evaluates a calibration-aware conditional WGAN-GP pipeline under explicit feasibility constraints. The generator is guided by moment, correlation, domain, and label-consistency penalties; hyperparameters are selected through constrained Bayesian optimization; and final assessment uses repeated outer splits with separate real-data training, calibration, and test partitions. The study follows the reporting principles of TRIPOD+AI and deliberately distinguishes methodological internal validation from prospective clinical evaluation [15, 16].

The central research question is whether conditional synthetic records add predictive information beyond a real-only elastic-net model and beyond standard real-data recalibration. The analysis therefore tests four related hypotheses: augmentation may alter discrimination; it may alter probability quality; any apparent benefit may be reproducible by conventional recalibration; and predictive utility may coexist with poor generative feasibility. The aim is not to presume that synthetic augmentation is beneficial, but to identify the conditions under which a measured benefit is scientifically defensible.

2. Related work

2.1. Synthetic generation for tabular and health data

GANs established adversarial learning as a flexible route to implicit density modelling [6]. Conditional GANs introduced label- or covariate-controlled generation [7], while WGAN-GP improved training stability through a Wasserstein critic and gradient penalty [17]. For structured health records, medGAN demonstrated adversarial synthesis of multi-label discrete records, and CTGAN introduced mode-specific normalization and conditional sampling for mixed-type tabular data [18, 19].

The tabular synthesis literature now extends beyond GANs. Reviews describe probabilistic, copula-based, variational, adversarial, autoregressive, and diffusion-based approaches, and emphasize that model suitability depends on the data type and intended use [20]. Diffusion models such as TabDDPM explicitly model mixed continuous and categorical variables and provide an important non-adversarial comparison for future work [21]. Health-specific reviews and empirical frameworks similarly conclude that there is no universal generator or single quality metric that dominates across datasets and use cases [8, 9, 11, 12].

2.2. Evaluation of fidelity, utility, and privacy risk

Synthetic-data evaluation is inherently multidimensional. Marginal statistics and correlation differences quantify selected aspects of resemblance; classifier two-sample tests evaluate whether a learned discriminator can separate real and synthetic samples [22]; and downstream train-on-synthetic or augmentation experiments assess task utility. Sample-level precision, recall, and authenticity metrics further distinguish fidelity, diversity, and copying behaviour [10].

Privacy must be assessed independently of resemblance. Nearest-neighbour distances may reveal obvious near-duplicates, but they do not establish protection against membership, linkage, or attribute-inference attacks. Empirical privacy audits have shown that synthetic data are not automatically anonymous and that privacy–utility trade-offs can be unpredictable [23, 24]. Accordingly, the present study describes its nearest-neighbour analysis as a memorization-oriented screen rather than a privacy guarantee.

2.3. Probability calibration and recalibration

Calibration is central to risk prediction because numerical probabilities may influence communication or downstream decisions. Calibration intercept and slope diagnose systematic bias and over- or under-dispersion, whereas Brier score and log loss are proper scoring rules [1, 2]. Platt scaling, isotonic regression, beta calibration, and temperature scaling provide post-hoc mappings with different flexibility–stability trade-offs [25, 26].

Binned calibration errors require particular care. Equal-width ECE can be small when predictions occupy a narrow range and many bins are empty, while equal-frequency estimators can become unstable if tied scores are split arbitrarily. Empirical studies have shown that conclusions can change with the estimator and binning design, motivating the use of multiple bin counts, tie-preserving bins, calibration curves, and non-binned measures [13, 14].

2.4. Bayesian optimization under feasibility constraints

Bayesian optimization is appropriate when each hyperparameter evaluation is expensive and stochastic [27, 28, 29]. Constrained variants model both the objective and unknown constraints, selecting candidates according to expected improvement weighted by the probability of feasibility [30]. This formulation is especially relevant for synthetic-data pipelines because a candidate with favourable predictive utility may still violate fidelity, memorization, or domain requirements.

A critical reporting distinction is whether a feasible configuration was actually observed. Selecting the lowest-objective infeasible candidate may allow a planned experiment to continue, but it does not constitute successful constrained optimization. The present study retains and reports this distinction explicitly.

2.5. Limitations of existing work

The preceding literature reveals five recurring limitations. First, many synthetic-data studies emphasize marginal resemblance or downstream AUC without jointly evaluating calibration, proper scoring rules, feasibility, and privacy-oriented diagnostics. Second, augmentation is rarely compared with simple post-hoc recalibration, making it difficult to determine whether synthetic observations provide incremental value. Third, mixed-type generators commonly rely on post-generation repair, but raw violation rates are not always reported. Fourth, privacy is often inferred from visual dissimilarity or nearest-neighbour distances rather than evaluated under an explicit threat

model. Fifth, prediction-model reports do not always separate tuning, training, calibration, and testing or describe the unit of analysis with sufficient precision.

2.6. Contributions of this study

The study makes the following contributions:

(C1) It evaluates conditional synthetic augmentation as a probabilistic intervention, with discrimination, proper scoring rules, calibration intercept and slope, ECE/ACE sensitivity, and decision curves assessed together.

(C2) It compares augmented and real-only predictors under identical elastic-net regularization and corresponding real-data calibration procedures, isolating the incremental contribution of synthetic observations.

(C3) It integrates predictive utility with explicit C2ST, memorization, and raw-domain feasibility constraints in Bayesian optimization and reports whether a feasible candidate was actually found.

(C4) It uses a leakage-conscious repeated design with a dedicated tuning partition and separate real training, calibration, and test partitions in every outer split.

(C5) It reports post-projection utility together with pre-projection violations, preventing deterministic repair from being mistaken for learned feature-domain validity.

(C6) It provides an empirical case study showing how a small predictive regularization effect can coexist with weak generative feasibility, thereby clarifying the distinction between utility and realism.

3. Materials and methods

3.1. Data source, predictors, and unit of analysis

The analysis used the Mendeley Data resource *HIBA ASRI_ Miscarriage Prediction Risk Factors*, version 1 [3]. The analysed file contained 15 candidate predictors representing demographic characteristics, previous miscarriage history, activity indicators, location, temperature, heart-rate-related measurements, stress, blood pressure, alcohol consumption, and intoxication-related status. The binary outcome indicated miscarriage status.

A total of 50,000 records were analysed. The term *record* is used deliberately because the source publications describe repeated acquisition from mobile phones and healthcare sensors [4, 5]. The record count therefore cannot be interpreted as the number of independent pregnant participants. No stable participant identifier was available in the analysed file, so the recorded split mode was row-level. Row-level partitioning prevents an identical row from appearing in more than one partition, but it cannot prevent repeated records from the same participant from crossing partitions.

3.2. Nested study design and leakage control

The complete dataset was first divided into a dedicated tuning partition and a final evaluation pool. The tuning partition contained 9,999 records and was used for all generative and predictive hyperparameter selection. The evaluation pool contained 40,001 records and was not used until the selected configuration had been locked.

Within the tuning partition, each Bayesian-optimization evaluation used an internal training and validation split. Preprocessing, auxiliary-classifier training, cWGAN-GP fitting, synthetic generation, and elastic-net fitting were performed using only the internal training data; the predictive objective and feasibility constraints were evaluated on the corresponding validation data.

The locked configuration was assessed over 20 repeated outer splits of the evaluation pool. Each split contained 60% real training records, 20% real calibration records, and 20% held-out real test records. All preprocessing parameters, condition encodings, auxiliary classifiers, generators, synthetic records, and predictive models were re-estimated within the current outer training partition. Calibration maps were fitted only on the real calibration partition. All reported predictive metrics were computed on held-out real test records, and the same test records were used for real-versus-augmented paired comparisons within each split.

3.3. Preprocessing and conditioning representation

Numeric variables were converted robustly, median-imputed from the current training partition, centred, and scaled. Variables that could not be reliably parsed as numeric were encoded categorically with explicit levels for missing and previously unseen values. The transformation estimated from the training partition was applied unchanged to calibration and test data.

For conditional generation, age and body mass index were grouped using prespecified cut points. Let $y \in \{0, 1\}$ denote the binary outcome, and let $\mathbf{s}$ denote the one-hot subgroup representation. The conditioning vector was

$$\mathbf{c} = [y, \mathbf{s}]^\top. \tag{1}$$

This design enabled outcome-specific synthesis while preserving broad age and body-mass-index strata.

3.4. Conditional WGAN-GP architecture

Let $\mathbf{x} \in \mathbb{R}^d$ denote a processed record. The generator maps latent noise and a conditioning vector to a synthetic record,

$$\widetilde{\mathbf{x}} = G_\theta(\mathbf{z}, \mathbf{c}), \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_k), \tag{2}$$

where k is the latent dimension. The critic $D_\omega(\mathbf{x}, \mathbf{c})$ returns a real-valued score.

3.4.1. Critic objective

For real and generated observations with the same conditioning structure, the critic minimized

$$\begin{aligned} \mathcal{L}_D = \mathbb{E}\left[D_\omega(\widetilde{\mathbf{x}}, \mathbf{c})\right] - \mathbb{E}\left[D_\omega(\mathbf{x}, \mathbf{c})\right] \\ + \lambda_{\text{gp}} \mathbb{E}\left[\left(\left\|\nabla_{\widehat{\mathbf{x}}} D_\omega(\widehat{\mathbf{x}}, \mathbf{c})\right\|_2 - 1\right)^2\right], \end{aligned} \tag{3}$$

where $\widehat{\mathbf{x}}$ is sampled on line segments between real and generated records. The critic was updated n_{critic} times for each generator update.

3.4.2. Generator objective and auxiliary penalties

The generator minimized

$$\begin{aligned} \mathcal{L}_G = -\mathbb{E}\left[D_\omega(\widetilde{\mathbf{x}}, \mathbf{c})\right] + \lambda_{\text{mom}} \mathcal{L}_{\text{mom}} + \lambda_{\text{corr}} \mathcal{L}_{\text{corr}} \\ + \lambda_{\text{dom}} \mathcal{L}_{\text{dom}} + \lambda_{\text{label}} \mathcal{L}_{\text{label}}. \end{aligned} \tag{4}$$

The moment term aligned feature-wise means and variances,

$$\mathcal{L}_{\text{mom}} = \frac{1}{d} \sum_{j=1}^{d} \left[(\widetilde{\mu}_j - \mu_j)^2 + (\widetilde{\sigma}_j^2 - \sigma_j^2)^2\right]. \tag{5}$$

The correlation term compared batch correlation matrices,

$$\mathcal{L}_{\text{corr}} = \frac{1}{d^2} \left\| \widetilde{R} - R \right\|_F^2 . \tag{6}$$

The raw-domain term penalized numeric values outside expanded training ranges and invalid one-hot blocks,

$$\begin{aligned} \mathcal{L}_{\text{dom}} = \frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} \left[\max(0, l_j - \widetilde{x}_{ij})^2 + \max(0, \widetilde{x}_{ij} - u_j)^2 \right] \\ + \sum_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j \in g} \widetilde{x}_{ij} - 1 \right)^2 + \mathcal{P}_{[0,1]}(\widetilde{\mathbf{x}}_g) \right] , \end{aligned} \tag{7}$$

where l_j and u_j are training-derived bounds, $\mathcal{G}$ is the set of categorical one-hot blocks, and $\mathcal{P}_{[0,1]}$ penalizes values outside $[0, 1]$.

The label-consistency loss used a frozen auxiliary classifier C_ψ trained only on the current real training partition,

$$\mathcal{L}_{\text{label}} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log C_\psi(\widetilde{\mathbf{x}}_i) + (1 - y_i) \log\{1 - C_\psi(\widetilde{\mathbf{x}}_i)\} \right] . \tag{8}$$

The classifier parameters were fixed during generator updates, while gradients were retained with respect to the generated features.

3.5. Synthetic generation, projection, and raw feasibility

For each outer split, the trained generator produced 4,000 records with outcome proportions matching the real training partition. Numeric outputs were projected to the observed training range. Each categorical block was projected to a valid one-hot vector by retaining its maximum component.

Projection ensured that the records used for predictor training were syntactically valid, but it could conceal unstable raw outputs. The pre-projection matrix was therefore retained. A raw record was considered domain-violating when any numeric component exceeded its permitted range or any categorical block violated its range or sum-to-one condition. The raw domain-violation rate was evaluated before repair and used as a Bayesian-optimization constraint.

3.6. Predictive model and probability calibration

The downstream predictor was elastic-net logistic regression. For a training dataset $\mathcal{D}$, coefficients were estimated by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}; \mathcal{D}) + \lambda_\phi \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right] \right\}. \tag{9}$$

Two base models were fitted in every outer split: one using real training records only and one using the union of real and synthetic training records. Both used the same locked α and λ_ϕ.

Five calibration strategies were evaluated for each training source: no recalibration, Platt scaling, isotonic regression, beta calibration, and temperature scaling [25, 26]. Every calibration map was fitted to base-model predictions and outcomes from the real calibration partition and then applied to the real test partition. Synthetic records were never used to estimate a calibration map.

3.7. Constrained Bayesian optimization

The hyperparameter vector was

$$\mathbf{u} = (\lambda_{\text{gp}}, \lambda_{\text{mom}}, \lambda_{\text{corr}}, \lambda_{\text{dom}}, \lambda_{\text{label}}, k, n_{\text{critic}}, \alpha, \log_{10} \lambda_\phi). \tag{10}$$

The scalar validation objective was

$$f(\mathbf{u}) = w_{\text{auc}}(1 - \text{AUC}) + w_{\text{Brier}}\text{Brier} + w_{\text{NLL}}\text{NLL} + w_{\text{ECE}}\text{ECE}_{10}. \tag{11}$$

The constraints were written as $c_q(\mathbf{u}) \leq 0$:

$$c_1(\mathbf{u}) = \text{C2ST AUC} - 0.90, \tag{12}$$

$$c_2(\mathbf{u}) = \text{memorization fraction} - 0.05, \tag{13}$$

$$c_3(\mathbf{u}) = \text{raw domain violation rate} - 0.02. \tag{14}$$

Independent Gaussian-process surrogates modelled the objective and constraints. Candidate selection used expected improvement multiplied by the posterior probability of joint feasibility,

$$a(\mathbf{u}) = \text{EI}(\mathbf{u}) \prod_{q=1}^{3} \Phi\left(\frac{-\mu_{c_q}(\mathbf{u})}{\sigma_{c_q}(\mathbf{u})} \right), \tag{15}$$

where μ_{c_q} and σ_{c_q} are the posterior mean and standard deviation for constraint q [30]. The run comprised 12 initial Latin-hypercube candidates followed by 8 sequential BO evaluations. The selected hyperparameters were $\lambda_{\text{gp}} = 14.922$, $\lambda_{\text{mom}} = 1.755$, $\lambda_{\text{corr}} = 2.025$, $\lambda_{\text{dom}} = 2.232$, $\lambda_{\text{label}} = 0.803$, $k = 25$, $n_{\text{critic}} = 2$, $\alpha = 0.231$, and $\log_{10} \lambda_\phi = -1.202$. If no feasible candidate was observed, the lowest-objective candidate was retained for evaluation and explicitly labelled infeasible. Complete objective weights, search bounds, neural-network architecture, optimizer settings, training budgets, and seed construction are reported in Supplementary Table A.13.

3.8. Predictive evaluation measures

AUC quantified discrimination. For outcomes y_i and clipped probabilities p_i, the Brier score and log loss were

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2, \tag{16}$$

$$\text{LogLoss} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right]. \tag{17}$$

Calibration intercept was estimated from a logistic model with the prediction logit as an offset, and calibration slope from a logistic regression of the outcome on the prediction logit. Ideal values are 0 and 1, respectively [2].

For B bins, equal-width ECE was

$$\text{ECE}_B = \sum_{b=1}^{B} \frac{n_b}{n} \left| \bar{p}_b - \bar{y}_b \right|. \tag{18}$$

Tie-preserving adaptive calibration error (ACE) used equal-frequency score groups but never divided identical probabilities between groups. ECE and ACE were evaluated with 5, 10, 15, and 20 nominal bins because conclusions can depend on binning design [13, 14].

Decision-curve net benefit at threshold t was

$$\text{NB}(t) = \frac{\text{TP}(t)}{n} - \frac{\text{FP}(t)}{n} \frac{t}{1 - t}, \tag{19}$$

with exploratory thresholds from 0.05 to 0.50 [31]. Because the source data do not define a clinical action linked to these thresholds, decision curves were treated as descriptive rather than deployment evidence.

3.9. Synthetic-data fidelity and memorization-oriented diagnostics

Feature-level fidelity was assessed with the Kolmogorov–Smirnov statistic, standardized mean difference, and real–synthetic correlation differences. Principal-component overlays summarized global geometry. A held-out classifier two-sample test attempted to distinguish real from synthetic records; AUC near 0.5 indicates low distinguishability, whereas larger values indicate systematic distributional differences [22].

For the memorization screen, nearest-neighbour distances were computed in the processed feature space. The reference threshold was the first percentile of real-to-real nearest-neighbour distances. The memorization fraction was the proportion of synthetic records whose nearest real neighbour was closer than this threshold. This statistic detects unusually close records under the selected metric but does not quantify membership-inference, linkability, or attribute-disclosure risk [23, 24].

3.10. Subgroups and statistical analysis

Prespecified subgroups were defined by age, body mass index, alcohol consumption, and previous miscarriage history. A subgroup estimate was retained only when minimum requirements for total observations, events, and non-events were met.

Performance was summarized across outer splits using means, standard deviations, and bootstrap resampling intervals. Real-versus-augmented comparisons were paired by split. Paired mean differences and bootstrap intervals were supplemented by paired t tests and Wilcoxon signed-rank tests, with Benjamini–Hochberg adjustment across related comparisons. Because outer splits reuse records from the same finite evaluation pool, the intervals quantify resampling stability under this design rather than independent-study population uncertainty.

3.11. Complete analysis algorithm

Algorithm 1 Calibration-aware conditional augmentation with constrained tuning and repeated held-out evaluation

Require: Real records $\mathcal{D}$; tuning fraction ρ; BO budget B; outer repetitions R; synthetic size M; calibration methods $\mathcal{C}$

Ensure: Locked hyperparameters, held-out predictions, performance summaries, fidelity diagnostics, and paired comparisons

Stratify $\mathcal{D}$ into tuning data $\mathcal{D}_{\text{tune}}$ and evaluation pool $\mathcal{D}_{\text{eval}}$
for $b = 1, \ldots, B$ **do**
    Select candidate $\mathbf{u}_b$ from the initial design or constrained acquisition function
    Split $\mathcal{D}_{\text{tune}}$ into BO-training and BO-validation records
    Fit preprocessing and condition encoders on BO-training records only
    Train the auxiliary classifier and cWGAN-GP under $\mathbf{u}_b$
    Generate and project synthetic BO-training records; retain raw outputs
    Fit the augmented elastic-net predictor and evaluate $f(\mathbf{u}_b)$ on BO-validation records
    Compute C2ST, memorization, and raw-domain constraint residuals
    Update the Gaussian-process models for the objective and constraints
end for
Select the best feasible $\mathbf{u}^\star$; if none is feasible, select the lowest-objective candidate and flag infeasibility
for $r = 1, \ldots, R$ **do**
    Split $\mathcal{D}_{\text{eval}}$ into real training, real calibration, and real test partitions
    Fit preprocessing, condition encoder, auxiliary classifier, and cWGAN-GP on real training records
    Generate M synthetic records; retain raw and projected versions
    Fit real-only and real-plus-synthetic elastic-net predictors using the same locked regularization
    for all $c \in \mathcal{C}$ **do**
        Fit calibration map c on the real calibration partition only
        Apply c to held-out real test probabilities from both training sources
        Store AUC, Brier, log loss, calibration intercept/slope, ECE, ACE, and decision-curve values
    end for
    Compute subgroup, feature-fidelity, C2ST, domain-violation, and memorization diagnostics
end for
Pair real-only and augmented results by outer split and corresponding calibration method
Produce bootstrap summaries, multiplicity-adjusted tests, figures, tables, and reproducibility metadata

3.12. Workflow

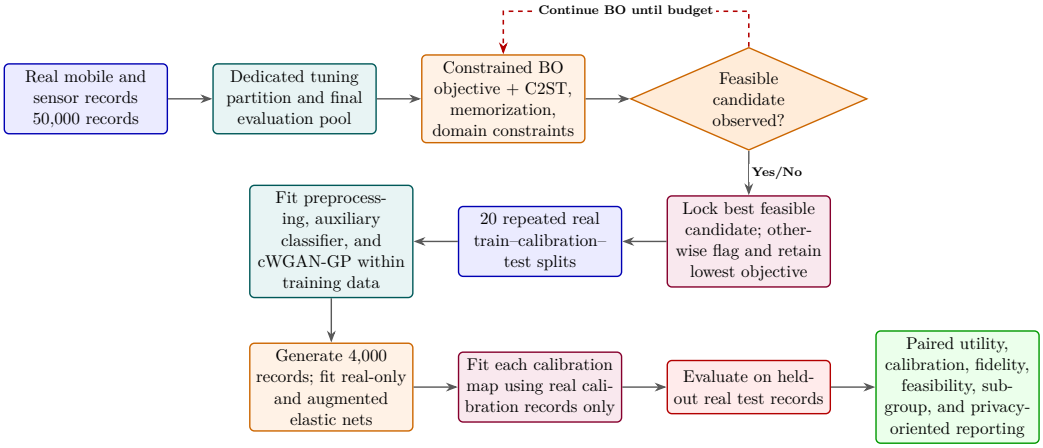

Figure 1: Complete analysis workflow. Hyperparameter selection was isolated from final evaluation. Within each outer split, synthetic generation used only the real training partition, recalibration used only real calibration records, and all predictive estimates were obtained from held-out real test records. Raw generator outputs were retained for feasibility assessment before deterministic projection.

4. Empirical results

4.1. Analysis flow and partition integrity

The analysis included 50,000 records. The dedicated tuning partition contained 9,999 records, and the final evaluation pool contained 40,001 records. All 20 outer repetitions completed successfully, with 4,000 synthetic records generated per split. Table 1 summarizes the design.

Table 1: Data partitions and validation design.

Item	Value
Raw records	50,000
Dedicated tuning records	9,999
Final evaluation-pool records	40,001
Completed outer splits	20
Training/calibration/test fractions	0.60 / 0.20 / 0.20
Synthetic records per outer split	4,000
BO evaluations	20
Split mode	row-level
Calibration methods	none, Platt, isotonic, beta, temperature
ECE/ACE bin counts	5, 10, 15, 20

The row-overlap audit was zero for the real training, calibration, and test partitions in every outer split. This confirms separation of record indices and prevents direct reuse of the same row across partitions. It does not establish participant separation, because the source file did not include a valid identifier and the split mode was row-level. The performance estimates must therefore be interpreted as record-level internal validation.

4.2. Constrained Bayesian optimization and feasibility

The constrained search completed 20 evaluations. None satisfied all three constraints (0 feasible candidates). Figure 2 shows that the predictive objective improved during the search, while the memorization constraint was consistently satisfied and the C2ST residual was usually non-positive. The domain residual, however, remained substantially above zero throughout the search. The selected candidate was therefore the lowest-objective infeasible configuration rather than a feasible constrained optimum.

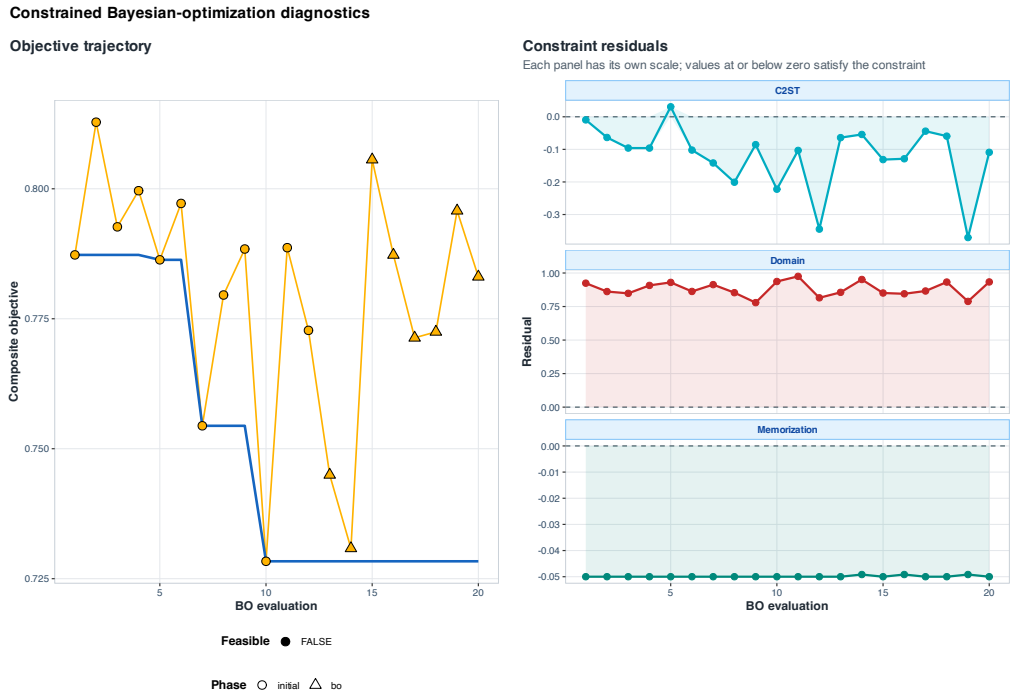

Figure 2: Constrained Bayesian-optimization diagnostics. The left panel shows the evaluated objective and cumulative best value. The right panels show residuals for the C2ST, raw-domain, and memorization constraints; values at or below zero satisfy the corresponding requirement. The persistently positive domain residual explains why no jointly feasible candidate was identified.

The failure to identify a feasible configuration is a substantive result. It indicates that the chosen architecture, training budget, search space, and 2% raw-domain threshold were not jointly compatible in the explored region. The outer-split mean raw domain-violation rate was 87.5%, confirming that this was not a tuning-partition anomaly. The subsequent evaluation therefore quantifies the predictive behaviour of a repaired, infeasible generator and should not be presented as evidence that constrained BO produced a naturally valid synthetic-data model.

4.3. Primary predictive comparison

Table 2 reports held-out performance for the main training and calibration strategies. For the uncalibrated predictors, mean AUC was 0.641 (bootstrap resampling interval 0.639–0.642) with real-only training and 0.641

(0.639–0.643) after augmentation. The paired difference was 0.000 (0.000–0.000), and the adjusted Wilcoxon result provided no evidence of a ranking change. The near-zero difference across all 20 paired splits is also visible in Fig. 4.

Table 2: Performance on held-out real test data. Values are mean (95 percent bootstrap resampling interval) across outer splits.

method	ACE (10 bins)	AUC	Brier score	Calibration intercept	Calibration slope	ECE (10 bins)	Log loss
Real + Synthetic	0.221 (0.207, 0.235)	0.641 (0.639, 0.643)	0.236 (0.235, 0.236)	-0.001 (-0.004, 0.002)	1.139 (1.120, 1.159)	0.052 (0.047, 0.056)	0.663 (0.663, 0.664)
Real + Synthetic + Isotonic	0.008 (0.007, 0.010)	0.719 (0.717, 0.721)	0.198 (0.198, 0.199)	0.003 (-0.006, 0.011)	1.002 (0.956, 1.048)	0.008 (0.006, 0.010)	0.559 (0.557, 0.560)
Real + Synthetic + Platt	0.222 (0.208, 0.237)	0.641 (0.639, 0.643)	0.236 (0.235, 0.236)	0.002 (-0.003, 0.006)	0.994 (0.966, 1.022)	0.095 (0.072, 0.120)	0.663 (0.662, 0.664)
Real only	0.228 (0.214, 0.242)	0.641 (0.639, 0.642)	0.236 (0.236, 0.237)	-0.000 (-0.002, 0.002)	1.387 (1.364, 1.410)	0.123 (0.115, 0.131)	0.665 (0.665, 0.666)
Real only + Isotonic	0.008 (0.007, 0.010)	0.719 (0.717, 0.720)	0.198 (0.198, 0.199)	0.003 (-0.007, 0.011)	1.002 (0.954, 1.049)	0.008 (0.006, 0.010)	0.559 (0.557, 0.560)
Real only + Platt	0.231 (0.218, 0.244)	0.641 (0.638, 0.642)	0.236 (0.235, 0.236)	0.002 (-0.003, 0.006)	0.994 (0.967, 1.022)	0.098 (0.074, 0.125)	0.663 (0.662, 0.664)

Table 3: Primary paired comparison of uncalibrated augmented versus real-only models. Differences are augmented minus real-only.

method	reference	metric	n_pairs	difference_ci	p_wilcoxon_bh
Real + Synthetic	Real only	AUC	20.000	approximately 0.000	1.000
Real + Synthetic	Real only	Brier score	20.000	-0.001 (-0.001, -0.001)	<0.001
Real + Synthetic	Real only	Log loss	20.000	-0.002 (-0.002, -0.002)	<0.001
Real + Synthetic	Real only	|Calibration intercept|	20.000	0.000 (-0.000, 0.001)	0.756
Real + Synthetic	Real only	|Calibration slope - 1|	20.000	-0.248 (-0.259, -0.237)	<0.001
Real + Synthetic	Real only	ECE (10 bins)	20.000	-0.071 (-0.081, -0.061)	<0.001
Real + Synthetic	Real only	ACE (10 bins)	20.000	-0.007 (-0.016, 0.002)	0.751

The probability-quality measures showed a different pattern. Mean Brier score improved by -0.001, and mean log loss improved by -0.002. Although these absolute changes are small, their paired intervals excluded zero and the adjusted tests were below 0.001. The calibration slope moved from 1.387 to 1.139, reducing absolute slope error by -0.248. These results indicate that augmentation changed the scale of the uncalibrated probabilities without changing their ordering.

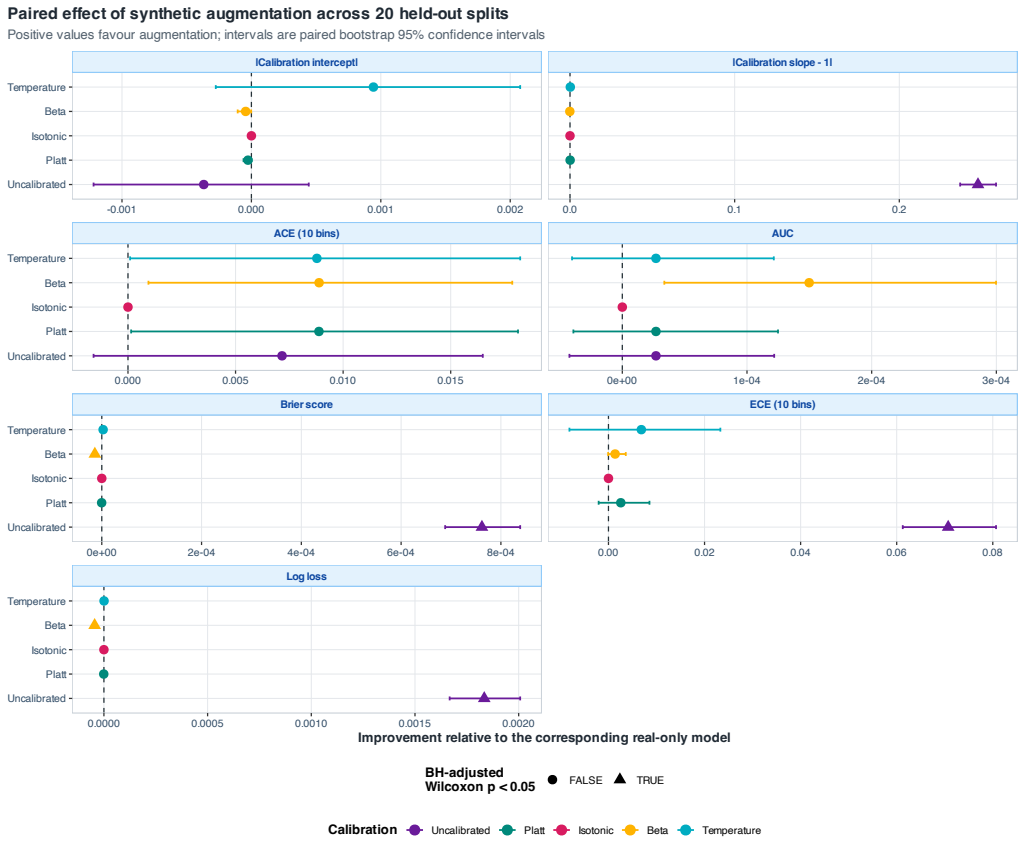

Figure 3: Paired effect of synthetic augmentation across the 20 outer splits. Effects are oriented so that positive values favour augmentation. The uncalibrated comparison shows no AUC change but small improvements in Brier score, log loss, calibration-slope error, and equal-width ECE. Corresponding calibrated pairs show whether any augmentation effect remains after applying the same calibration family.

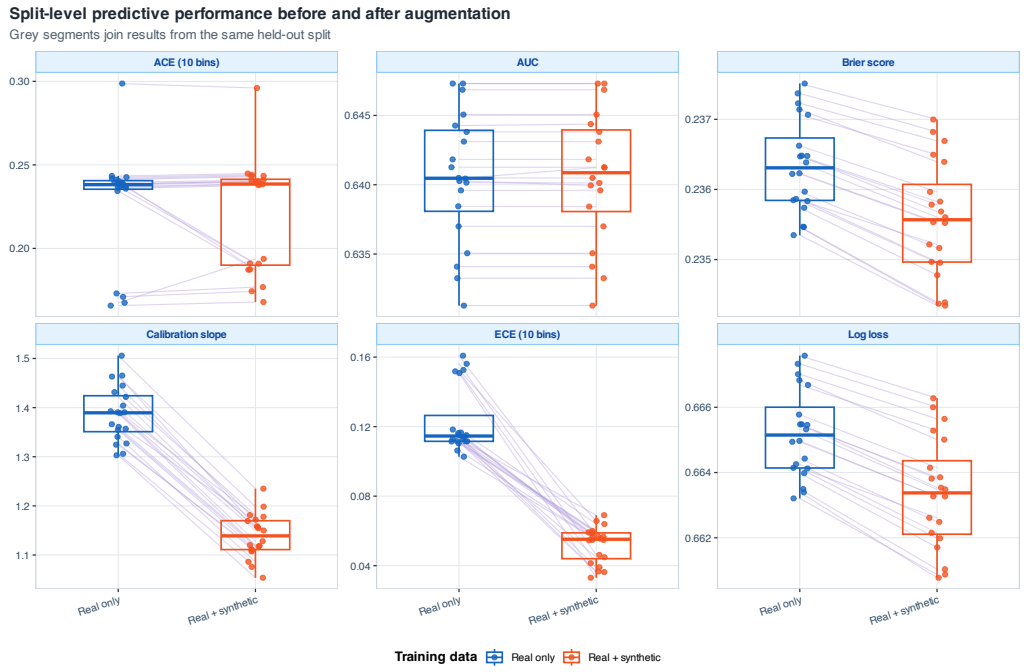

Figure 4: Split-level comparison of the uncalibrated real-only and augmented models. Each line connects estimates from the same outer split. AUC estimates almost completely overlap, whereas the Brier score, log loss, calibration slope, and equal-width ECE shift consistently after augmentation.

4.4. Calibration, estimator sensitivity, and standard recalibration

For uncalibrated predictions, 10-bin equal-width ECE decreased from 0.123 to 0.052, corresponding to a paired difference of -0.071. Tie-preserving 10-bin ACE also decreased numerically, from 0.228 to 0.221, with a paired difference of -0.007 (resampling interval -0.016 to 0.002; adjusted Wilcoxon $p = 0.751$). The ACE interval included zero, so it did not provide clear evidence of a difference between training strategies. The larger and more precise ECE reduction therefore depended on the binning estimator. Equal-width bins emphasize discrepancies across the unit interval, whereas equal-frequency bins emphasize the distribution of the observed score mass. Because the predictor produced many tied or closely clustered scores, calibration slope, Brier score, log loss, and the calibration curves receive greater interpretive weight than either binned estimator alone.

Figure 5 shows that augmentation moved the uncalibrated relationship toward the diagonal, particularly by reducing over-dispersion. Once Platt

or temperature calibration was fitted to the real calibration partition, corresponding real-only and augmented models became nearly indistinguishable. Table 3 confirms that their paired Brier, log-loss, slope-error, ECE, and ACE differences were close to zero and were generally not significant after multiplicity adjustment. Thus, most of the probability-quality benefit produced by augmentation could also be obtained through conventional real-data recalibration.

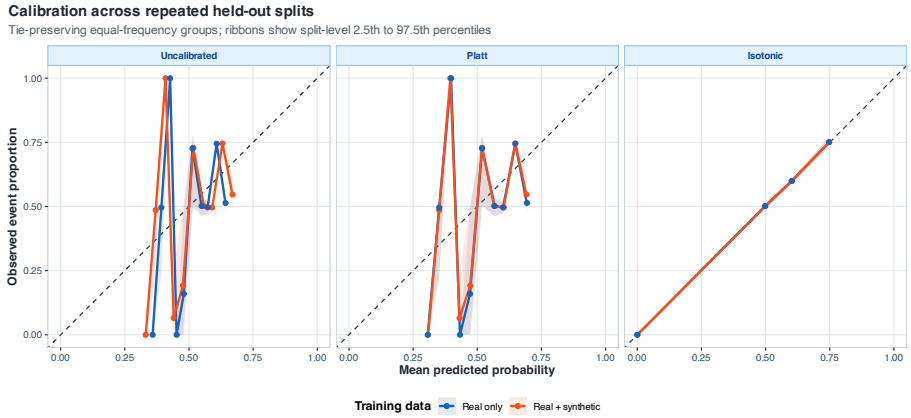

Figure 5: Calibration across repeated held-out splits using tie-preserving score groups. Real-only and augmented predictors are displayed within the same calibration family. The uncalibrated augmented curve is closer to the diagonal, the corresponding Platt curves are nearly superimposed, and the isotonic curves overlap because both training sources are mapped to essentially the same stepwise probability levels. Temperature-calibrated results are reported in the accompanying tables and supplementary comparisons.

Isotonic regression requires a separate caution. It yielded Brier scores near 0.198, log loss near 0.559, ECE near 0.008, and a reported AUC near 0.719 for both training sources. Isotonic calibration is a monotone step function and cannot create new predictor information; however, it can collapse observations into large tied groups and thereby alter empirical AUC through the treatment of ties. The identical real-only and augmented isotonic results suggest that the calibration mapping, rather than synthetic augmentation, drove these values. The supplementary score-resolution audit should therefore accompany any discussion of isotonic AUC.

4.5. Synthetic-data fidelity, feasibility, and memorization

Table 4 and Fig. 6 summarize the generative diagnostics. The mean held-out C2ST AUC was 0.612. This value is closer to chance than the BO constraint limit of 0.90 but remains clearly above 0.5, indicating residual multivariate differences between real and synthetic records. Mean feature-level KS was 0.222, and the largest feature-level KS was 0.628. Activity-related variables showed the largest marginal discrepancies, whereas some encoded variables were reproduced more closely.

Table 4: Synthetic-data fidelity and memorization-oriented diagnostics.

Diagnostic	Mean	SD
Held-out C2ST AUC	0.612	0.077
Raw domain-violation rate	0.875	0.012
Synthetic fraction below real–real 1st-percentile NN distance	0.022%	0.0%
Median synthetic-to-real NN distance	1.770	0.030
Median real-to-real NN distance	1.427	0.009
Mean feature KS statistic	0.222	0.145
Maximum feature KS statistic	0.628	
Mean absolute SMD	0.050	0.047

The dominant problem was raw feasibility. On average, 87.5% of pre-projection generated records violated at least one numeric-range or categorical-block rule. The final predictor was trained on clipped and projected records, so its modest utility benefit cannot be interpreted as evidence that the generator learned a plausible mixed-type distribution. Rather, deterministic repair was a material component of the augmentation pipeline.

The nearest-neighbour memorization fraction was only 0.022%. This argues against simple near-duplicate copying under the chosen processed-space distance and threshold. It does not rule out membership inference, attribute inference, or linkage attacks, and it should not be described as a formal privacy guarantee.

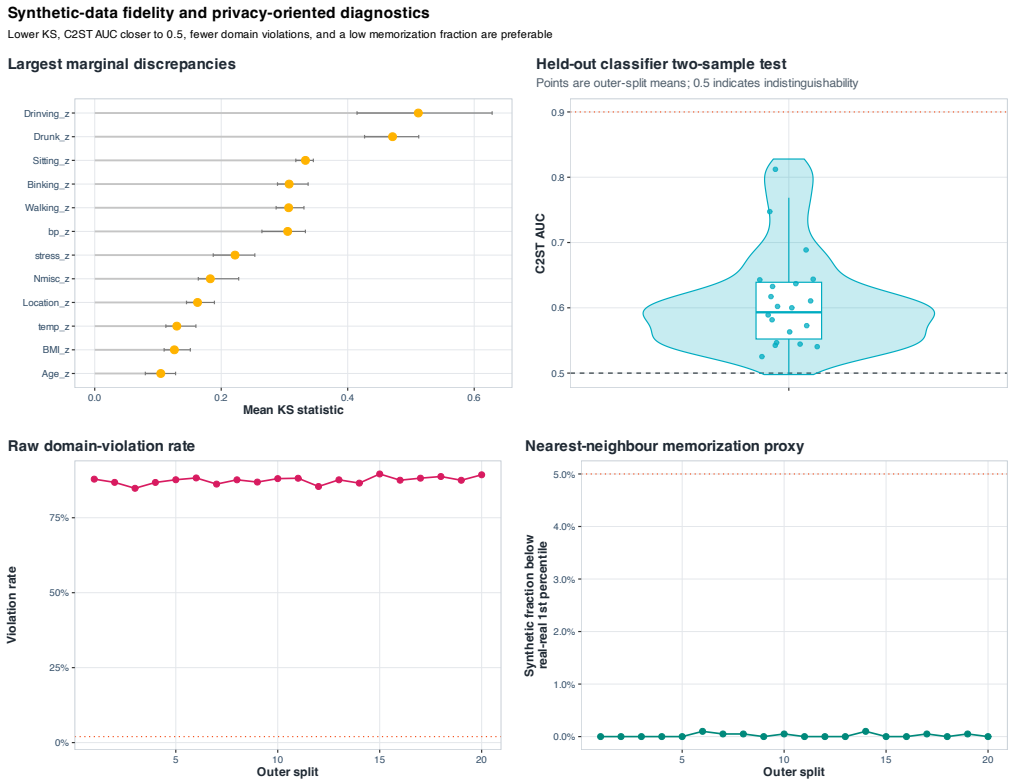

Figure 6: Synthetic-data fidelity, raw feasibility, and memorization-oriented diagnostics across outer splits. The feature panel reports the largest marginal discrepancies; the C2ST distribution measures real–synthetic distinguishability; the domain panel shows the pre-projection violation rate; and the nearest-neighbour panel screens for unusually close synthetic records.

4.6. Decision curves and subgroup analyses

Decision curves for corresponding real-only and augmented strategies were very similar over thresholds from 0.05 to 0.50. Because the source data do not define a specific intervention, harm ratio, or stakeholder-agreed threshold, these curves describe mathematical net benefit rather than a validated clinical policy. They are therefore reported in the supplement.

Exploratory subgroup comparisons showed negligible AUC changes and small improvements in Brier score or equal-width ECE in several age, body-mass-index, alcohol, and miscarriage-history groups. No subgroup finding overcame the broader limitations of row-level splitting, repeated use of the same evaluation pool, and multiple comparisons. The subgroup results are

best interpreted as a screen for pronounced heterogeneity rather than evidence of subgroup-specific clinical benefit.

5. Discussion

5.1. Principal findings

The study produced a trade-off rather than an unequivocal improvement. Conditional synthetic augmentation did not change AUC, but it modestly improved the uncalibrated elastic-net predictor's Brier score, log loss, calibration slope, and equal-width ECE. The paired design showed that these changes were stable across the repeated splits, although their absolute magnitude was small. Tie-preserving ACE also decreased numerically, but its interval included zero and the multiplicity-adjusted test was not significant. The strength of the calibration conclusion therefore depended on the estimator: equal-width ECE suggested a clear reduction, whereas ACE supported at most a smaller and uncertain change.

The most important comparative finding is that the augmentation effect largely disappeared after Platt or temperature calibration. Standard recalibration, fitted solely to real calibration records, captured most of the observed probability-scale improvement without requiring synthetic observations. The defensible interpretation is therefore narrow: augmentation acted as a mild regularizer for the uncalibrated predictor, not as a general route to superior risk prediction.

5.2. Augmentation versus recalibration

Synthetic augmentation and post-hoc calibration intervene at different stages. Augmentation alters the effective training distribution and can change coefficient estimates or shrinkage behaviour. Recalibration leaves the ranking model fixed and adjusts only the mapping from scores to probabilities. The similar Platt- and temperature-calibrated results indicate that the main augmentation effect occurred on the probability scale rather than in the learned ordering of records.

This comparison is necessary because an uncalibrated-only analysis would have overstated the contribution of the generator. Equal-width ECE fell substantially after augmentation, but proper scores changed only slightly and corresponding recalibrated models were nearly identical. Reporting all of these quantities prevents a metric-specific result from being presented as a universal calibration improvement.

5.3. Predictive utility does not establish generative fidelity

The predictive and generative findings point in different directions. The repaired synthetic records provided a small regularization benefit, but the raw generator violated the feature domain in approximately nine out of ten records. This discrepancy shows that downstream utility can arise even when generated observations do not form a faithful or naturally valid approximation to the source distribution.

Such utility may result from perturbing the training geometry, modifying effective class-conditional covariance, or changing the elastic-net solution. These mechanisms can be useful experimentally, but they do not justify releasing the synthetic records as realistic data or reusing them for arbitrary analyses. Fidelity, task utility, and privacy risk must remain separate claims [10, 11, 12].

5.4. What the constrained search established

No evaluated candidate satisfied all constraints. This does not make the optimization uninformative. The BO trajectory showed that the predictive objective could be improved while C2ST and memorization requirements were mostly satisfied, but the domain requirement remained binding. The search therefore exposed a persistent incompatibility between the selected generator architecture and the raw mixed-type feature constraints under the available budget.

The lowest-objective candidate was retained to complete the prespecified repeated evaluation, but it remained infeasible. The manuscript should consequently avoid statements that BO found an "optimal feasible generator." A more accurate conclusion is that constrained BO identified a configuration with favourable predictive objective but failed to find a configuration meeting the full feasibility specification.

5.5. Implications for mobile and sensor risk modelling

The data source is appropriate for a methodological study of repeated mobile and sensor records, but it is not equivalent to a conventional prospective patient cohort. The apparent stability of the estimates may partly reflect correlations among repeated observations. Moreover, the outcome definition, timing, sensor acquisition quality, and participant characteristics are inherited from the public resource and were not independently adjudicated.

The results therefore do not establish a clinically deployable miscarriage-risk model. A deployment claim would require participant-level temporal

validation, externally collected records, a clearly defined prediction time, and a decision pathway linked to risk thresholds. TRIPOD+AI and DECIDE-AI emphasize transparent reporting and staged clinical evaluation; the present work should be positioned before those deployment stages [15, 16].

5.6. Methodological lessons

Three broader lessons follow. First, synthetic augmentation should be compared with standard recalibration and not only with an uncalibrated baseline. Second, binned calibration statistics should preserve tied scores and be interpreted alongside calibration intercept, slope, and proper scoring rules. Third, post-generation repair should be disclosed together with the raw violation rate; otherwise, a syntactically valid synthetic dataset may conceal an unstable generator.

A fourth lesson concerns evaluation breadth. Recent health-data benchmarks increasingly evaluate fidelity, utility, and privacy jointly and find that no single generator dominates every dimension [9, 11, 12]. The present results support that multidimensional view and add calibration and explicit feasibility to the evaluation framework.

6. Limitations

The most important limitation is the unit of analysis. The source contains repeated mobile and sensor records, and no participant identifier was available for group-level splitting. Records from the same participant may therefore have crossed training, calibration, and test partitions, producing optimistic and overly stable performance estimates. Row-level overlap was excluded, but participant-level leakage could not be audited.

Second, the study used a single public dataset and did not perform external validation. Repeated internal splits measure variation under resampling of the same source; they do not establish transportability to another population, device, acquisition protocol, healthcare system, or time period.

Third, the generator operated in a processed feature space with a shared linear output and depended heavily on clipping and one-hot projection. The high raw domain-violation rate prevents claims that the model naturally generated plausible mixed-type records.

Fourth, the generative comparison was limited to the proposed cWGAN-GP. The study did not benchmark copula models, variational autoencoders,

CTAB-GAN variants, diffusion models such as TabDDPM, or simple perturbation and oversampling baselines [21]. Consequently, the results do not establish that adversarial generation is preferable to alternative augmentation methods.

Fifth, C2ST, KS, SMD, correlation differences, PCA, and nearest-neighbour distances are incomplete summaries of fidelity and privacy. The analysis did not include membership-inference, attribute-inference, or linkage attacks and did not provide differential-privacy guarantees [23, 24].

Sixth, the generator was conditioned on the observed outcome. The experiment evaluates supervised augmentation and does not support causal claims about activity, stress, alcohol-related variables, physiological measurements, or miscarriage.

Seventh, the outer splits overlap and are not independent studies. Bootstrap intervals across split-level estimates quantify resampling stability under the chosen procedure, not population-level uncertainty from independent cohorts.

Finally, the decision thresholds were illustrative. Clinical utility requires a specified action, stakeholder-informed consequences of false-positive and false-negative decisions, and prospective external validation.

7. Future work

The first priority is participant-level and temporal validation. Future data collection should retain stable participant identifiers and timestamps so that all records from one participant remain in one partition and models can be tested on both new participants and future time windows. External validation should include different devices, acquisition settings, and populations.

The generator should be redesigned for mixed-type data. Feature-specific output heads, explicit distributions for continuous, ordinal, binary, and categorical variables, and architectures that model temporal dependence may reduce reliance on deterministic projection. Diffusion-based tabular generators and non-neural probabilistic baselines should be evaluated under the same nested protocol [21].

Optimization should become feasibility-first. One option is a two-stage procedure in which only architectures achieving acceptable raw-domain validity enter the calibration-utility search. Alternative approaches include hard architectural constraints, constrained output transformations, repair-aware objectives, larger training budgets, and boundary-focused constrained BO.

Privacy evaluation should include explicit membership, linkage, and attribute-inference attacks, ideally with comparison to differentially private generators. The memorization screen can remain as a diagnostic but should not be the sole privacy-oriented analysis.

Finally, clinical evaluation should begin only after the intended use is specified. A future study should define the prediction time, candidate intervention, threshold range, and consequences of incorrect decisions in collaboration with clinicians and patients. Prospective evaluation should then follow appropriate prediction-model and clinical AI reporting guidance [15, 16].

8. Conclusion

Conditional synthetic augmentation produced a small but consistent improvement in several uncalibrated probability-quality measures without changing discrimination. The benefit was largely removed by standard real-data recalibration, indicating that augmentation acted mainly as a modest regularizer of the probability scale. At the same time, no Bayesian-optimization candidate met all feasibility constraints, real and synthetic records remained distinguishable, and most raw generator outputs required deterministic repair. Predictive utility therefore did not imply generative fidelity.

The study's main contribution is an evaluation framework that separates ranking, probability quality, recalibration, fidelity, raw feasibility, and memorization-oriented evidence. Under the current record-level internal validation, the results support a cautious methodological conclusion rather than a clinical claim. Participant-level external validation, a substantially more feasible mixed-type generator, stronger privacy auditing, and a defined clinical decision context are required before synthetic augmentation can be considered for practical miscarriage-risk modelling.

CRediT authorship contribution statement

Sadia Saba: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.

Muhammad Amir Saeed: Methodology, Software, Validation, Visualization, Writing – review and editing.

Antonio Candelieri: Supervision, Methodology, Validation, Writing – review and editing.

Data and code availability

The source dataset is available through Mendeley Data at https://doi.org/10.17632/5sbmhh6t3r.1 [3]. The final analysis code will be made available by the corresponding author upon reasonable request.

Ethics statement

This study is a secondary methodological analysis of a publicly available dataset and involved no new participant recruitment, intervention, or data collection. The source publication reports ethical approval and consent to participate as not applicable. The dataset was accessed and reused in accordance with its Creative Commons Attribution 4.0 licence. No claims regarding prospective clinical validation or direct patient management are made in this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the University of Milano-Bicocca and the scholarship support provided for this research.

References

[1] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Review 78 (1) (1950) 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[2] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, E. W. Steyerberg, Calibration: the achilles heel of predictive analytics, BMC Medicine 17 (2019) 230. doi:10.1186/s12916-019-1466-7.

[3] H. Asri, Hiba asri_ miscarriage prediction risk factors, Mendeley Data, version 1 (2021). doi:10.17632/5sbmhh6t3r.1.

[4] H. Asri, Z. Jarir, Real-time miscarriage prediction: A comprehensive real-world dataset and a new model, Procedia Computer Science 203 (2022) 763–768. doi:10.1016/j.procs.2022.07.114.

[5] H. Asri, Z. Jarir, Toward a smart health: big data analytics and iot for real-time miscarriage prediction, Journal of Big Data 10 (2023) 34. doi:10.1186/s40537-023-00704-9.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, Vol. 27, 2014, pp. 2672–2680.

[7] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[8] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, A. P. Sales, Generation and evaluation of synthetic patient data, BMC Medical Research Methodology 20 (2020) 108. doi:10.1186/s12874-020-00977-1.

[9] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, D. Rankin, Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions, Methods of Information in Medicine 62 (S 01) (2023) e19–e38. doi:10.1055/s-0042-1760247.

[10] A. Alaa, B. Van Breugel, E. S. Saveliev, M. van der Schaar, How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models, in: Proceedings of the 39th International Conference on Machine Learning, Vol. 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 290–306.

[11] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, N. Aginako, Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of

generative models with and without privacy guarantees, Frontiers in Digital Health 7 (2025) 1576290. doi:10.3389/fdgth.2025.1576290.

[12] I. Nanevski, M. Mohebi, S. Jager, K. Otte, F. Prasser, M. Schulte-Althoff, D. Furstenau, F. Biessmann, Evaluating the quality of tabular synthetic data in health care, PLOS Digital Health 5 (7) (2026) e0001522. doi:10.1371/journal.pdig.0001522.

[13] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, D. Tran, Measuring calibration in deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 38–41.

[14] R. Roelofs, N. Cain, J. Shlens, M. C. Mozer, Mitigating bias in calibration error estimation, in: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Vol. 151 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 4036–4054.

[15] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, A.-L. Boulesteix, et al., Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378. doi:10.1136/bmj-2023-078378.

[16] B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, Nature Medicine 28 (5) (2022) 924–933. doi:10.1038/s41591-022-01772-9.

[17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 5767–5777.

[18] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: Proceedings of the 2nd Machine Learning for Healthcare Conference, Vol. 68 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 286–305.

[19] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 7335–7345.

[20] J. Fonseca, F. Bacao, Tabular and latent space synthetic data generation: a literature review, Journal of Big Data 10 (2023) 115. doi:10.1186/s40537-023-00792-7.

[21] A. Kotelnikov, D. Baranchuk, I. Rubachev, A. Babenko, TabDDPM: Modelling tabular data with diffusion models, in: Proceedings of the 40th International Conference on Machine Learning, Vol. 202 of Proceedings of Machine Learning Research, PMLR, 2023, pp. 17564–17579.

[22] D. Lopez-Paz, M. Oquab, Revisiting classifier two-sample tests, International Conference on Learning Representations (2017).

[23] T. Stadler, B. Oprisanu, C. Troncoso, Synthetic data – anonymisation groundhog day, in: 31st USENIX Security Symposium (USENIX Security 22), USENIX Association, 2022, pp. 1451–1468.

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

[25] M. Kull, T. Silva Filho, P. Flach, Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Vol. 54 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 623–631.

[26] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 1321–1330.

[27] D. R. Jones, M. Schonlau, W. J. Welch, Efficient global optimization of expensive black-box functions, Journal of Global Optimization 13 (4) (1998) 455–492. doi:10.1023/A:1008306431147.

[28] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, Vol. 25, 2012, pp. 2951–2959.

[29] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, Taking the human out of the loop: A review of bayesian optimization, Proceedings of the IEEE 104 (1) (2016) 148–175. doi:10.1109/JPROC.2015.2494218.

[30] M. A. Gelbart, J. Snoek, R. P. Adams, Bayesian optimization with unknown constraints, in: Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2014, pp. 250–259.

[31] A. J. Vickers, E. B. Elkin, Decision curve analysis: A novel method for evaluating prediction models, Medical Decision Making 26 (6) (2006) 565–574. doi:10.1177/0272989X06295361.

Appendix A. Supplementary figures and tables

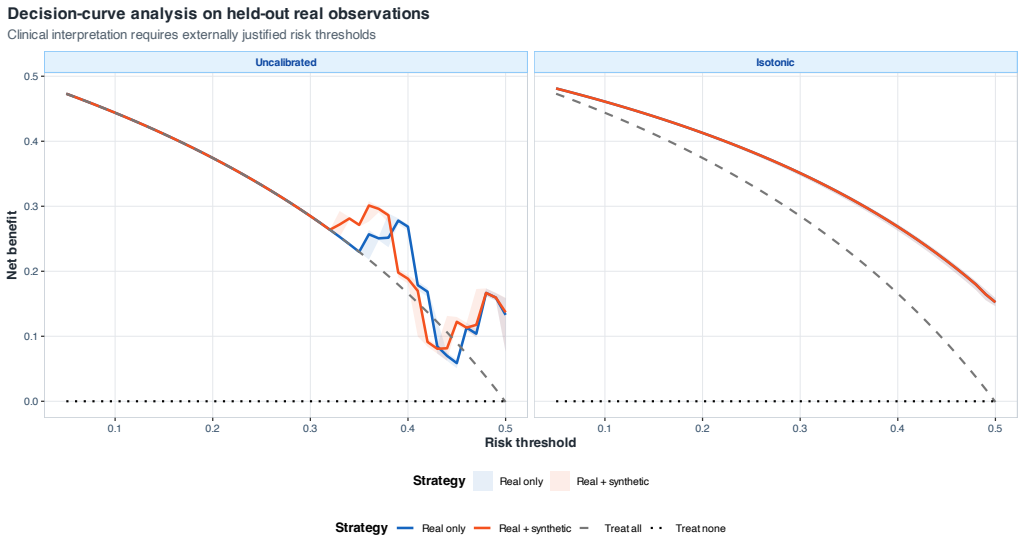

Figure A.7: Exploratory decision-curve analysis. Thresholds are not tied to a prespecified clinical action, so net benefit is descriptive.

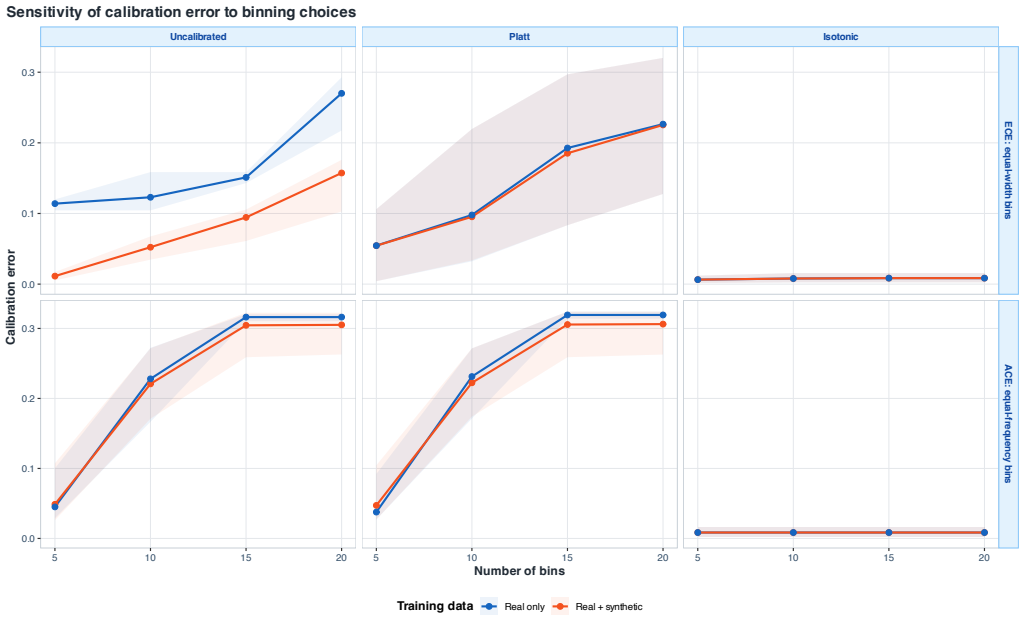

Figure A.8: Sensitivity of equal-width ECE and tie-preserving ACE to the number of nominal bins.

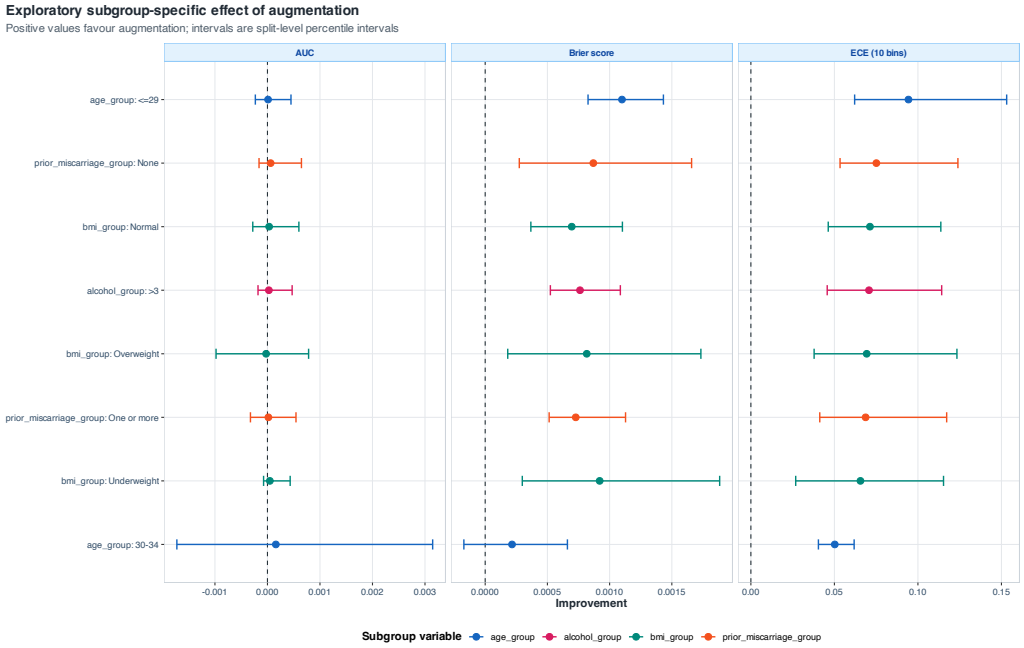

Figure A.9: Exploratory subgroup effects of augmentation. Positive values favour augmentation after orienting each metric so that higher values indicate benefit.

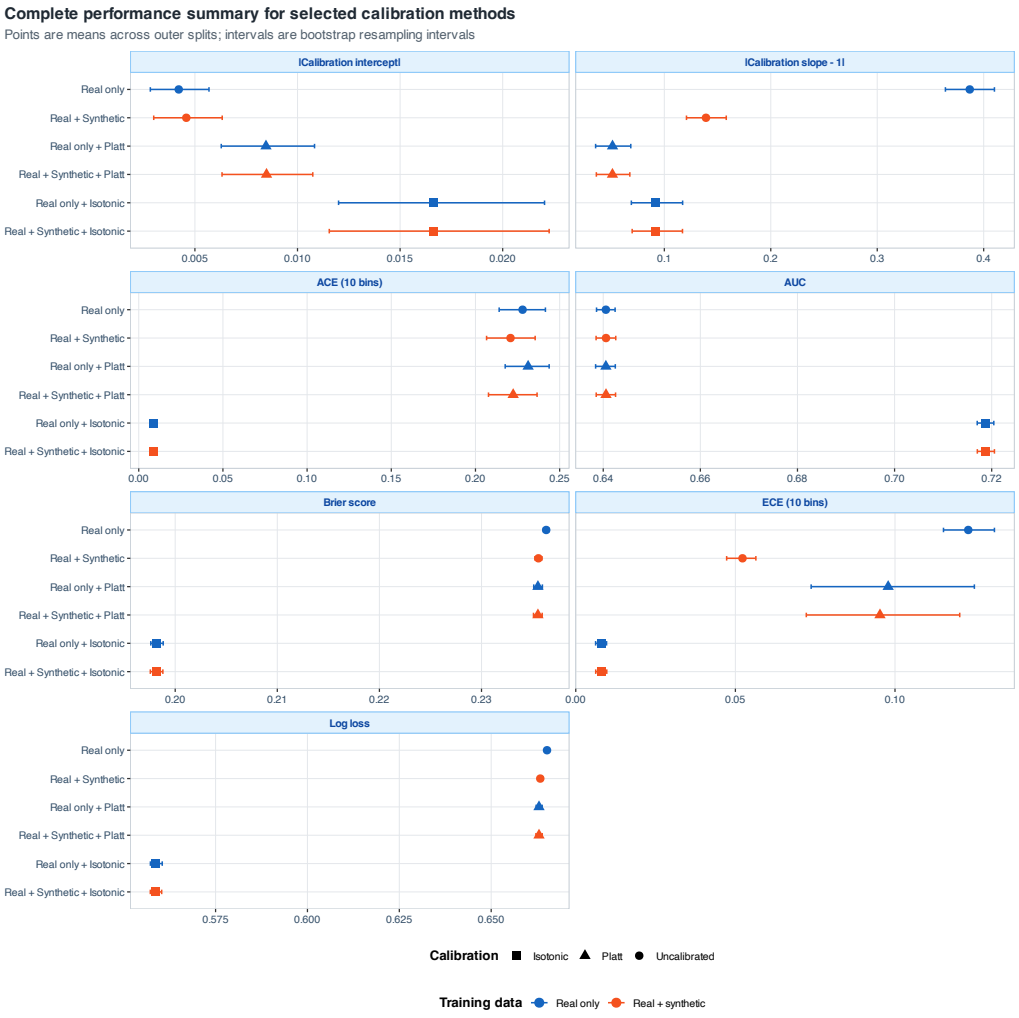

Figure A.10: Complete performance summary across selected training sources and calibration strategies.

Table A.5: Complete performance summary for all models and calibration methods.

Method	Metric	n	Mean	SD	Median	2.5%	97.5%	Training	Calibration
Real + Synthetic	ace10	20	0.2208	0.0339	0.2385	0.2066	0.2355	augmented	none
Real + Synthetic	ace15	20	0.3045	0.0244	0.3164	0.2934	0.3136	augmented	none
Real + Synthetic	ace20	20	0.3051	0.0232	0.3164	0.2944	0.3138	augmented	none
Real + Synthetic	ace5	20	0.0487	0.0266	0.0368	0.0384	0.0611	augmented	none
Real + Synthetic	AUC	20	0.6406	0.0047	0.6409	0.6385	0.6426	augmented	none
Real + Synthetic	Brier	20	0.2356	0.0008	0.2356	0.2352	0.2359	augmented	none
Real + Synthetic	Cal. intercept	20	-0.0012	0.0061	-0.0011	-0.0037	0.0015	augmented	none
Real + Synthetic	Abs. cal. intercept	20	0.0046	0.0040	0.0033	0.0030	0.0063	augmented	none
Real + Synthetic	Cal. slope	20	1.1391	0.0442	1.1390	1.1197	1.1586	augmented	none
Real + Synthetic	Abs. slope error	20	0.1391	0.0442	0.1390	0.1208	0.1582	augmented	none
Real + Synthetic	ece10	20	0.0522	0.0105	0.0552	0.0473	0.0564	augmented	none
Real + Synthetic	ece15	20	0.0944	0.0132	0.0989	0.0885	0.0994	augmented	none
Real + Synthetic	ece20	20	0.1573	0.0247	0.1699	0.1465	0.1669	augmented	none
Real + Synthetic	ece5	20	0.0113	0.0037	0.0106	0.0097	0.0129	augmented	none
Real + Synthetic	Log loss	20	0.6634	0.0017	0.6634	0.6627	0.6641	augmented	none
Real + Synthetic + Beta	ace10	20	0.2313	0.0289	0.2378	0.2185	0.2434	augmented	beta
Real + Synthetic + Beta	ace15	20	0.3052	0.0208	0.3146	0.2958	0.3129	augmented	beta
Real + Synthetic + Beta	ace20	20	0.3044	0.0223	0.3146	0.2945	0.3133	augmented	beta
Real + Synthetic + Beta	ace5	20	0.0709	0.0196	0.0638	0.0637	0.0803	augmented	beta
Real + Synthetic + Beta	AUC	20	0.6368	0.0121	0.6372	0.6316	0.6420	augmented	beta
Real + Synthetic + Beta	Brier	20	0.2343	0.0011	0.2341	0.2338	0.2348	augmented	beta
Real + Synthetic + Beta	Cal. intercept	20	0.0029	0.0111	0.0029	-0.0019	0.0075	augmented	beta
Real + Synthetic + Beta	Abs. cal. intercept	20	0.0095	0.0062	0.0089	0.0069	0.0122	augmented	beta
Real + Synthetic + Beta	Cal. slope	20	0.9882	0.0480	0.9941	0.9662	1.0080	augmented	beta
Real + Synthetic + Beta	Abs. slope error	20	0.0352	0.0339	0.0262	0.0216	0.0506	augmented	beta
Real + Synthetic + Beta	ece10	20	0.1176	0.0115	0.1199	0.1125	0.1218	augmented	beta
Real + Synthetic + Beta	ece15	20	0.2151	0.0295	0.2289	0.2014	0.2278	augmented	beta

Continued on next page

Table A.5 continued from previous page

Method	Metric	n	Mean	SD	Median	2.5%	97.5%	Training	Calibration
Real + Synthetic + Beta	ece20	20	0.2340	0.0217	0.2415	0.2241	0.2425	augmented	beta
Real + Synthetic + Beta	ece5	20	0.0239	0.0054	0.0237	0.0216	0.0261	augmented	beta
Real + Synthetic + Beta	Log loss	20	0.6583	0.0024	0.6578	0.6573	0.6593	augmented	beta
Real + Synthetic + Isotonic	ace10	20	0.0085	0.0037	0.0089	0.0070	0.0100	augmented	isotonic
Real + Synthetic + Isotonic	ace15	20	0.0085	0.0037	0.0089	0.0069	0.0100	augmented	isotonic
Real + Synthetic + Isotonic	ace20	20	0.0085	0.0037	0.0089	0.0070	0.0101	augmented	isotonic
Real + Synthetic + Isotonic	ace5	20	0.0085	0.0037	0.0089	0.0069	0.0101	augmented	isotonic
Real + Synthetic + Isotonic	AUC	20	0.7188	0.0040	0.7190	0.7170	0.7206	augmented	isotonic
Real + Synthetic + Isotonic	Brier	20	0.1982	0.0014	0.1978	0.1976	0.1988	augmented	isotonic
Real + Synthetic + Isotonic	Cal. intercept	20	0.0027	0.0207	0.0074	-0.0063	0.0112	augmented	isotonic
Real + Synthetic + Isotonic	Abs. cal. intercept	20	0.0166	0.0121	0.0152	0.0115	0.0223	augmented	isotonic
Real + Synthetic + Isotonic	Cal. slope	20	1.0018	0.1098	1.0152	0.9564	1.0478	augmented	isotonic
Real + Synthetic + Isotonic	Abs. slope error	20	0.0917	0.0566	0.0831	0.0698	0.1171	augmented	isotonic
Real + Synthetic + Isotonic	ece10	20	0.0079	0.0040	0.0072	0.0063	0.0097	augmented	isotonic
Real + Synthetic + Isotonic	ece15	20	0.0085	0.0037	0.0089	0.0069	0.0101	augmented	isotonic
Real + Synthetic + Isotonic	ece20	20	0.0085	0.0037	0.0089	0.0070	0.0100	augmented	isotonic
Real + Synthetic + Isotonic	ece5	20	0.0064	0.0035	0.0059	0.0048	0.0078	augmented	isotonic
Real + Synthetic + Isotonic	Log loss	20	0.5588	0.0037	0.5580	0.5572	0.5603	augmented	isotonic
Real + Synthetic + Platt	ace10	20	0.2224	0.0333	0.2405	0.2077	0.2366	augmented	platt
Real + Synthetic + Platt	ace15	20	0.3056	0.0247	0.3174	0.2943	0.3152	augmented	platt
Real + Synthetic + Platt	ace20	20	0.3062	0.0236	0.3174	0.2956	0.3160	augmented	platt
Real + Synthetic + Platt	ace5	20	0.0472	0.0255	0.0378	0.0380	0.0594	augmented	platt
Real + Synthetic + Platt	AUC	20	0.6406	0.0047	0.6409	0.6386	0.6425	augmented	platt
Real + Synthetic + Platt	Brier	20	0.2355	0.0010	0.2355	0.2351	0.2359	augmented	platt
Real + Synthetic + Platt	Cal. intercept	20	0.0019	0.0101	0.0035	-0.0026	0.0063	augmented	platt
Real + Synthetic + Platt	Abs. cal. intercept	20	0.0085	0.0055	0.0073	0.0063	0.0107	augmented	platt
Real + Synthetic + Platt	Cal. slope	20	0.9943	0.0638	0.9926	0.9664	1.0222	augmented	platt
Real + Synthetic + Platt	Abs. slope error	20	0.0513	0.0365	0.0421	0.0361	0.0676	augmented	platt

Continued on next page

Table A.5 continued from previous page

Method	Metric	n	Mean	SD	Median	2.5%	97.5%	Training	Calibration
Real + Synthetic + Platt	ece10	20	0.0953	0.0577	0.0831	0.0722	0.1203	augmented	platt
Real + Synthetic + Platt	ece15	20	0.1852	0.0861	0.1631	0.1497	0.2225	augmented	platt
Real + Synthetic + Platt	ece20	20	0.2253	0.0529	0.2210	0.2029	0.2462	augmented	platt
Real + Synthetic + Platt	ece5	20	0.0544	0.0431	0.0484	0.0359	0.0727	augmented	platt
Real + Synthetic + Platt	Log loss	20	0.6630	0.0020	0.6630	0.6622	0.6639	augmented	platt
Real + Synthetic + Temperature	ace10	20	0.2225	0.0332	0.2403	0.2081	0.2373	augmented	temperature
Real + Synthetic + Temperature	ace15	20	0.3058	0.0246	0.3176	0.2945	0.3149	augmented	temperature
Real + Synthetic + Temperature	ace20	20	0.3064	0.0234	0.3176	0.2962	0.3156	augmented	temperature
Real + Synthetic + Temperature	ace5	20	0.0472	0.0255	0.0378	0.0375	0.0590	augmented	temperature
Real + Synthetic + Temperature	AUC	20	0.6406	0.0047	0.6409	0.6385	0.6425	augmented	temperature
Real + Synthetic + Temperature	Brier	20	0.2355	0.0010	0.2355	0.2351	0.2359	augmented	temperature
Real + Synthetic + Temperature	Cal. intercept	20	-0.0012	0.0069	-0.0009	-0.0040	0.0020	augmented	temperature
Real + Synthetic + Temperature	Abs. cal. intercept	20	0.0052	0.0046	0.0037	0.0032	0.0070	augmented	temperature
Real + Synthetic + Temperature	Cal. slope	20	0.9946	0.0637	0.9926	0.9677	1.0218	augmented	temperature
Real + Synthetic + Temperature	Abs. slope error	20	0.0511	0.0366	0.0421	0.0360	0.0670	augmented	temperature
Real + Synthetic + Temperature	ece10	20	0.1047	0.0640	0.0806	0.0785	0.1342	augmented	temperature
Real + Synthetic + Temperature	ece15	20	0.1618	0.0778	0.1285	0.1303	0.1970	augmented	temperature
Real + Synthetic + Temperature	ece20	20	0.2367	0.0563	0.2254	0.2123	0.2613	augmented	temperature
Real + Synthetic + Temperature	ece5	20	0.0482	0.0424	0.0218	0.0303	0.0666	augmented	temperature
Real + Synthetic + Temperature	Log loss	20	0.6630	0.0020	0.6630	0.6622	0.6639	augmented	temperature

Continued on next page

Table A.5 continued from previous page

Method	Metric	n	Mean	SD	Median	2.5%	97.5%	Training	Calibration
Real only	ace10	20	0.2280	0.0330	0.2382	0.2141	0.2416	real	none
Real only	ace15	20	0.3162	0.0029	0.3167	0.3150	0.3174	real	none
Real only	ace20	20	0.3162	0.0029	0.3167	0.3149	0.3174	real	none
Real only	ace5	20	0.0451	0.0234	0.0405	0.0379	0.0567	real	none
Real only	AUC	20	0.6405	0.0047	0.6405	0.6386	0.6424	real	none
Real only	Brier	20	0.2363	0.0007	0.2363	0.2361	0.2366	real	none
Real only	Cal. intercept	20	-0.0002	0.0055	0.0008	-0.0024	0.0021	real	none
Real only	Abs. cal. intercept	20	0.0042	0.0033	0.0033	0.0028	0.0057	real	none
Real only	Cal. slope	20	1.3869	0.0559	1.3895	1.3643	1.4102	real	none
Real only	Abs. slope error	20	0.3869	0.0559	0.3895	0.3640	0.4101	real	none
Real only	ece10	20	0.1229	0.0191	0.1146	0.1152	0.1311	real	none
Real only	ece15	20	0.1511	0.0049	0.1516	0.1489	0.1531	real	none
Real only	ece20	20	0.2701	0.0290	0.2838	0.2573	0.2819	real	none
Real only	ece5	20	0.1139	0.0044	0.1146	0.1119	0.1158	real	none
Real only	Log loss	20	0.6652	0.0013	0.6651	0.6646	0.6658	real	none
Real only + Beta	ace10	20	0.2402	0.0269	0.2427	0.2279	0.2511	real	beta
Real only + Beta	ace15	20	0.3165	0.0032	0.3164	0.3152	0.3179	real	beta
Real only + Beta	ace20	20	0.3165	0.0032	0.3164	0.3151	0.3179	real	beta
Real only + Beta	ace5	20	0.0658	0.0202	0.0590	0.0588	0.0759	real	beta
Real only + Beta	AUC	20	0.6367	0.0121	0.6369	0.6316	0.6419	real	beta
Real only + Beta	Brier	20	0.2343	0.0011	0.2340	0.2338	0.2348	real	beta
Real only + Beta	Cal. intercept	20	0.0029	0.0111	0.0029	-0.0020	0.0077	real	beta
Real only + Beta	Abs. cal. intercept	20	0.0094	0.0062	0.0085	0.0068	0.0121	real	beta
Real only + Beta	Cal. slope	20	0.9882	0.0479	0.9940	0.9681	1.0078	real	beta
Real only + Beta	Abs. slope error	20	0.0351	0.0338	0.0254	0.0208	0.0499	real	beta
Real only + Beta	ece10	20	0.1190	0.0112	0.1207	0.1135	0.1228	real	beta
Real only + Beta	ece15	20	0.2136	0.0329	0.2374	0.1993	0.2277	real	beta
Real only + Beta	ece20	20	0.2350	0.0222	0.2421	0.2249	0.2437	real	beta

Continued on next page

Table A.5 continued from previous page

Method	Metric	n	Mean	SD	Median	2.5%	97.5%	Training	Calibration
Real only + Beta	ece5	20	0.0239	0.0054	0.0238	0.0217	0.0262	real	beta
Real only + Beta	Log loss	20	0.6582	0.0024	0.6578	0.6572	0.6593	real	beta
Real only + Isotonic	ace10	20	0.0085	0.0037	0.0089	0.0069	0.0101	real	isotonic
Real only + Isotonic	ace15	20	0.0085	0.0037	0.0089	0.0069	0.0101	real	isotonic
Real only + Isotonic	ace20	20	0.0085	0.0037	0.0089	0.0069	0.0101	real	isotonic
Real only + Isotonic	ace5	20	0.0085	0.0037	0.0089	0.0069	0.0100	real	isotonic
Real only + Isotonic	AUC	20	0.7188	0.0040	0.7190	0.7170	0.7204	real	isotonic
Real only + Isotonic	Brier	20	0.1982	0.0014	0.1978	0.1976	0.1988	real	isotonic
Real only + Isotonic	Cal. intercept	20	0.0027	0.0207	0.0074	-0.0066	0.0109	real	isotonic
Real only + Isotonic	Abs. cal. intercept	20	0.0166	0.0121	0.0152	0.0120	0.0220	real	isotonic
Real only + Isotonic	Cal. slope	20	1.0018	0.1098	1.0152	0.9540	1.0487	real	isotonic
Real only + Isotonic	Abs. slope error	20	0.0917	0.0566	0.0831	0.0690	0.1172	real	isotonic
Real only + Isotonic	ece10	20	0.0079	0.0040	0.0072	0.0062	0.0096	real	isotonic
Real only + Isotonic	ece15	20	0.0085	0.0037	0.0089	0.0069	0.0100	real	isotonic
Real only + Isotonic	ece20	20	0.0085	0.0037	0.0089	0.0069	0.0101	real	isotonic
Real only + Isotonic	ece5	20	0.0064	0.0035	0.0059	0.0048	0.0079	real	isotonic
Real only + Isotonic	Log loss	20	0.5588	0.0037	0.5580	0.5573	0.5605	real	isotonic
Real only + Platt	ace10	20	0.2313	0.0313	0.2411	0.2177	0.2438	real	platt
Real only + Platt	ace15	20	0.3193	0.0032	0.3190	0.3179	0.3206	real	platt
Real only + Platt	ace20	20	0.3193	0.0032	0.3190	0.3179	0.3206	real	platt
Real only + Platt	ace5	20	0.0377	0.0232	0.0325	0.0311	0.0487	real	platt
Real only + Platt	AUC	20	0.6405	0.0047	0.6405	0.6384	0.6425	real	platt
Real only + Platt	Brier	20	0.2355	0.0010	0.2355	0.2351	0.2360	real	platt
Real only + Platt	Cal. intercept	20	0.0019	0.0101	0.0034	-0.0026	0.0063	real	platt
Real only + Platt	Abs. cal. intercept	20	0.0085	0.0055	0.0073	0.0063	0.0108	real	platt
Real only + Platt	Cal. slope	20	0.9944	0.0637	0.9916	0.9666	1.0215	real	platt
Real only + Platt	Abs. slope error	20	0.0513	0.0363	0.0421	0.0354	0.0685	real	platt
Real only + Platt	ece10	20	0.0979	0.0608	0.0828	0.0737	0.1249	real	platt

Continued on next page

Table A.5 continued from previous page

Method	Metric	n	Mean	SD	Median	2.5%	97.5%	Training	Calibration
Real only + Platt	ece15	20	0.1927	0.0901	0.1907	0.1541	0.2331	real	platt
Real only + Platt	ece20	20	0.2265	0.0574	0.2197	0.2004	0.2507	real	platt
Real only + Platt	ece5	20	0.0545	0.0453	0.0578	0.0346	0.0726	real	platt
Real only + Platt	Log loss	20	0.6630	0.0020	0.6631	0.6622	0.6639	real	platt
Real only + Temperature	ace10	20	0.2312	0.0313	0.2411	0.2181	0.2441	real	temperature
Real only + Temperature	ace15	20	0.3194	0.0031	0.3195	0.3180	0.3207	real	temperature
Real only + Temperature	ace20	20	0.3194	0.0031	0.3195	0.3181	0.3207	real	temperature
Real only + Temperature	ace5	20	0.0376	0.0233	0.0324	0.0308	0.0486	real	temperature
Real only + Temperature	AUC	20	0.6405	0.0047	0.6405	0.6385	0.6425	real	temperature
Real only + Temperature	Brier	20	0.2355	0.0010	0.2355	0.2351	0.2359	real	temperature
Real only + Temperature	Cal. intercept	20	0.0006	0.0077	0.0020	-0.0027	0.0039	real	temperature
Real only + Temperature	Abs. cal. intercept	20	0.0061	0.0045	0.0045	0.0044	0.0082	real	temperature
Real only + Temperature	Cal. slope	20	0.9945	0.0638	0.9916	0.9666	1.0220	real	temperature
Real only + Temperature	Abs. slope error	20	0.0513	0.0364	0.0421	0.0364	0.0670	real	temperature
Real only + Temperature	ece10	20	0.1116	0.0676	0.0911	0.0832	0.1397	real	temperature
Real only + Temperature	ece15	20	0.1891	0.0787	0.1926	0.1557	0.2229	real	temperature
Real only + Temperature	ece20	20	0.2432	0.0529	0.2198	0.2213	0.2649	real	temperature
Real only + Temperature	ece5	20	0.0576	0.0458	0.0911	0.0389	0.0767	real	temperature
Real only + Temperature	Log loss	20	0.6630	0.0020	0.6631	0.6622	0.6639	real	temperature

Table A.6: Calibration-error sensitivity to estimator and bin count.

method	estimator	bins	mean	lower	upper	training_source	calibration_method	training_label	calibration_family
Real + Synthetic	ACE: equal-frequency bins	5.0000	0.0487	0.0261	0.1074	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ACE: equal-frequency bins	10.0000	0.2208	0.1710	0.2717	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ACE: equal-frequency bins	15.0000	0.3045	0.2589	0.3224	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ACE: equal-frequency bins	20.0000	0.3051	0.2628	0.3224	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ECE: equal-width bins	5.0000	0.0113	0.0053	0.0171	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ECE: equal-width bins	10.0000	0.0522	0.0346	0.0675	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ECE: equal-width bins	15.0000	0.0944	0.0611	0.1056	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic	ECE: equal-width bins	20.0000	0.1573	0.1028	0.1760	augmented	none	Real + synthetic	Uncalibrated
Real + Synthetic + Isotonic	ACE: equal-frequency bins	5.0000	0.0085	0.0024	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ACE: equal-frequency bins	10.0000	0.0085	0.0024	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ACE: equal-frequency bins	15.0000	0.0085	0.0024	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ACE: equal-frequency bins	20.0000	0.0085	0.0024	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ECE: equal-width bins	5.0000	0.0064	0.0016	0.0120	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ECE: equal-width bins	10.0000	0.0079	0.0023	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ECE: equal-width bins	15.0000	0.0085	0.0024	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Isotonic	ECE: equal-width bins	20.0000	0.0085	0.0024	0.0157	augmented	isotonic	Real + synthetic	Isotonic
Real + Synthetic + Platt	ACE: equal-frequency bins	5.0000	0.0472	0.0270	0.1045	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ACE: equal-frequency bins	10.0000	0.2224	0.1740	0.2715	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ACE: equal-frequency bins	15.0000	0.3056	0.2589	0.3241	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ACE: equal-frequency bins	20.0000	0.3062	0.2627	0.3241	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ECE: equal-width bins	5.0000	0.0544	0.0041	0.1061	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ECE: equal-width bins	10.0000	0.0953	0.0335	0.2196	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ECE: equal-width bins	15.0000	0.1852	0.0834	0.2970	augmented	platt	Real + synthetic	Platt
Real + Synthetic + Platt	ECE: equal-width bins	20.0000	0.2253	0.1276	0.3203	augmented	platt	Real + synthetic	Platt
Real only	ACE: equal-frequency bins	5.0000	0.0451	0.0290	0.1003	real	none	Real only	Uncalibrated
Real only	ACE: equal-frequency bins	10.0000	0.2280	0.1666	0.2724	real	none	Real only	Uncalibrated
Real only	ACE: equal-frequency bins	15.0000	0.3162	0.3103	0.3203	real	none	Real only	Uncalibrated
Real only	ACE: equal-frequency bins	20.0000	0.3162	0.3103	0.3203	real	none	Real only	Uncalibrated
Real only	ECE: equal-width bins	5.0000	0.1139	0.1044	0.1199	real	none	Real only	Uncalibrated
Real only	ECE: equal-width bins	10.0000	0.1229	0.1044	0.1586	real	none	Real only	Uncalibrated
Real only	ECE: equal-width bins	15.0000	0.1511	0.1430	0.1586	real	none	Real only	Uncalibrated
Real only	ECE: equal-width bins	20.0000	0.2701	0.2174	0.2928	real	none	Real only	Uncalibrated
Real only + Isotonic	ACE: equal-frequency bins	5.0000	0.0085	0.0024	0.0157	real	isotonic	Real only	Isotonic
Real only + Isotonic	ACE: equal-frequency bins	10.0000	0.0085	0.0024	0.0157	real	isotonic	Real only	Isotonic
Real only + Isotonic	ACE: equal-frequency bins	15.0000	0.0085	0.0024	0.0157	real	isotonic	Real only	Isotonic
Real only + Isotonic	ACE: equal-frequency bins	20.0000	0.0085	0.0024	0.0157	real	isotonic	Real only	Isotonic
Real only + Isotonic	ECE: equal-width bins	5.0000	0.0064	0.0016	0.0120	real	isotonic	Real only	Isotonic
Real only + Isotonic	ECE: equal-width bins	10.0000	0.0079	0.0023	0.0157	real	isotonic	Real only	Isotonic
Real only + Isotonic	ECE: equal-width bins	15.0000	0.0085	0.0024	0.0157	real	isotonic	Real only	Isotonic
Real only + Isotonic	ECE: equal-width bins	20.0000	0.0085	0.0024	0.0157	real	isotonic	Real only	Isotonic
Real only + Platt	ACE: equal-frequency bins	5.0000	0.0377	0.0270	0.0921	real	platt	Real only	Platt
Real only + Platt	ACE: equal-frequency bins	10.0000	0.2313	0.1713	0.2715	real	platt	Real only	Platt
Real only + Platt	ACE: equal-frequency bins	15.0000	0.3193	0.3139	0.3241	real	platt	Real only	Platt
Real only + Platt	ACE: equal-frequency bins	20.0000	0.3193	0.3139	0.3241	real	platt	Real only	Platt
Real only + Platt	ECE: equal-width bins	5.0000	0.0545	0.0041	0.1061	real	platt	Real only	Platt
Real only + Platt	ECE: equal-width bins	10.0000	0.0979	0.0322	0.2196	real	platt	Real only	Platt
Real only + Platt	ECE: equal-width bins	15.0000	0.1927	0.0834	0.2970	real	platt	Real only	Platt
Real only + Platt	ECE: equal-width bins	20.0000	0.2265	0.1276	0.3203	real	platt	Real only	Platt

Table A.7: Net benefit at prespecified reporting thresholds.

Method	Threshold	Mean net benefit	SD	2.5%	97.5%
Real + Synthetic	0.10	0.4438	0.0000	0.4438	0.4438
Real + Synthetic	0.20	0.3743	0.0000	0.3743	0.3743
Real + Synthetic	0.30	0.2849	0.0000	0.2849	0.2849
Real + Synthetic	0.40	0.1885	0.0039	0.1811	0.1944
Real + Synthetic	0.50	0.1369	0.0275	0.0761	0.1593
Real + Synthetic + Beta	0.10	0.4438	0.0000	0.4438	0.4438
Real + Synthetic + Beta	0.20	0.3743	0.0000	0.3743	0.3743
Real + Synthetic + Beta	0.30	0.3483	0.0078	0.3295	0.3534
Real + Synthetic + Beta	0.40	0.1885	0.0039	0.1811	0.1944
Real + Synthetic + Beta	0.50	0.0735	0.0043	0.0678	0.0817
Real + Synthetic + Isotonic	0.10	0.4610	0.0005	0.4599	0.4616
Real + Synthetic + Isotonic	0.20	0.4128	0.0011	0.4105	0.4142
Real + Synthetic + Isotonic	0.30	0.3510	0.0018	0.3470	0.3534

Continued on next page

Table A.7 continued from previous page

Method	Threshold	Mean net benefit	SD	2.5%	97.5%
Real + Synthetic + Isotonic	0.40	0.2685	0.0028	0.2623	0.2722
Real + Synthetic + Isotonic	0.50	0.1524	0.0038	0.1462	0.1588
Real + Synthetic + Platt	0.10	0.4438	0.0000	0.4438	0.4438
Real + Synthetic + Platt	0.20	0.3743	0.0000	0.3743	0.3743
Real + Synthetic + Platt	0.30	0.2902	0.0117	0.2849	0.3177
Real + Synthetic + Platt	0.40	0.1500	0.0391	0.1044	0.1944
Real + Synthetic + Platt	0.50	0.1190	0.0358	0.0709	0.1593
Real + Synthetic + Temperature	0.10	0.4438	0.0000	0.4438	0.4438
Real + Synthetic + Temperature	0.20	0.3743	0.0000	0.3743	0.3743
Real + Synthetic + Temperature	0.30	0.2901	0.0114	0.2849	0.3165
Real + Synthetic + Temperature	0.40	0.1551	0.0374	0.1044	0.1938
Real + Synthetic + Temperature	0.50	0.1369	0.0275	0.0761	0.1593
Real only	0.10	0.4438	0.0000	0.4438	0.4438
Real only	0.20	0.3743	0.0000	0.3743	0.3743
Real only	0.30	0.2849	0.0000	0.2849	0.2849
Real only	0.40	0.2685	0.0028	0.2623	0.2722
Real only	0.50	0.1329	0.0330	0.0760	0.1584
Real only + Beta	0.10	0.4438	0.0000	0.4438	0.4438
Real only + Beta	0.20	0.3743	0.0000	0.3743	0.3743
Real only + Beta	0.30	0.3494	0.0075	0.3316	0.3534
Real only + Beta	0.40	0.1885	0.0039	0.1811	0.1944
Real only + Beta	0.50	0.0735	0.0043	0.0678	0.0817
Real only + Isotonic	0.10	0.4610	0.0005	0.4599	0.4616
Real only + Isotonic	0.20	0.4128	0.0011	0.4105	0.4142
Real only + Isotonic	0.30	0.3510	0.0018	0.3470	0.3534
Real only + Isotonic	0.40	0.2685	0.0028	0.2623	0.2722
Real only + Isotonic	0.50	0.1524	0.0038	0.1462	0.1588
Real only + Platt	0.10	0.4438	0.0000	0.4438	0.4438
Real only + Platt	0.20	0.3743	0.0000	0.3743	0.3743
Real only + Platt	0.30	0.2897	0.0117	0.2849	0.3177
Real only + Platt	0.40	0.1500	0.0408	0.1044	0.1944
Real only + Platt	0.50	0.1183	0.0396	0.0709	0.1593
Real only + Temperature	0.10	0.4438	0.0000	0.4438	0.4438
Real only + Temperature	0.20	0.3743	0.0000	0.3743	0.3743
Real only + Temperature	0.30	0.2913	0.0131	0.2849	0.3177
Real only + Temperature	0.40	0.1460	0.0401	0.1044	0.1938
Real only + Temperature	0.50	0.1329	0.0330	0.0760	0.1584
Treat all	0.10	0.4438	0.0000	0.4438	0.4438
Treat all	0.20	0.3743	0.0000	0.3743	0.3743
Treat all	0.30	0.2849	0.0000	0.2849	0.2849
Treat all	0.40	0.1657	0.0000	0.1657	0.1657
Treat all	0.50	-0.0011	0.0000	-0.0011	-0.0011
Treat none	0.10	0.0000	0.0000	0.0000	0.0000
Treat none	0.20	0.0000	0.0000	0.0000	0.0000
Treat none	0.30	0.0000	0.0000	0.0000	0.0000
Treat none	0.40	0.0000	0.0000	0.0000	0.0000
Treat none	0.50	0.0000	0.0000	0.0000	0.0000

Table A.8: Exploratory paired augmentation effects within eligible subgroups.

subgroup_variable	subgroup_level	metric	n_splits	mean_improvement	lower	upper	subgroup
age_group	30-34	AUC	20.0000	0.0002	-0.0017	0.0031	age_group: 30-34
age_group	30-34	Brier score	20.0000	0.0002	-0.0002	0.0007	age_group: 30-34
age_group	30-34	ECE (10 bins)	20.0000	0.0502	0.0404	0.0618	age_group: 30-34
age_group	<=29	AUC	20.0000	0.0000	-0.0002	0.0004	age_group: <=29
age_group	<=29	Brier score	20.0000	0.0011	0.0008	0.0014	age_group: <=29
age_group	<=29	ECE (10 bins)	20.0000	0.0943	0.0621	0.1531	age_group: <=29
alcohol_group	>3	AUC	20.0000	0.0000	-0.0002	0.0005	alcohol_group: >3
alcohol_group	>3	Brier score	20.0000	0.0008	0.0005	0.0011	alcohol_group: >3
alcohol_group	>3	ECE (10 bins)	20.0000	0.0707	0.0457	0.1142	alcohol_group: >3
bmi_group	Normal	AUC	20.0000	0.0000	-0.0003	0.0006	bmi_group: Normal
bmi_group	Normal	Brier score	20.0000	0.0007	0.0004	0.0011	bmi_group: Normal
bmi_group	Normal	ECE (10 bins)	20.0000	0.0713	0.0463	0.1137	bmi_group: Normal
bmi_group	Overweight	AUC	20.0000	-0.0000	-0.0010	0.0008	bmi_group: Overweight
bmi_group	Overweight	Brier score	20.0000	0.0008	0.0002	0.0017	bmi_group: Overweight
bmi_group	Overweight	ECE (10 bins)	20.0000	0.0693	0.0379	0.1233	bmi_group: Overweight
bmi_group	Underweight	AUC	20.0000	0.0000	-0.0001	0.0004	bmi_group: Underweight
bmi_group	Underweight	Brier score	20.0000	0.0009	0.0003	0.0019	bmi_group: Underweight
bmi_group	Underweight	ECE (10 bins)	20.0000	0.0656	0.0268	0.1153	bmi_group: Underweight
prior_miscarriage_group	None	AUC	20.0000	0.0001	-0.0002	0.0006	prior_miscarriage_group: None
prior_miscarriage_group	None	Brier score	20.0000	0.0009	0.0003	0.0017	prior_miscarriage_group: None
prior_miscarriage_group	None	ECE (10 bins)	20.0000	0.0751	0.0534	0.1239	prior_miscarriage_group: None
prior_miscarriage_group	One or more	AUC	20.0000	0.0000	-0.0003	0.0005	prior_miscarriage_group: One or more
prior_miscarriage_group	One or more	Brier score	20.0000	0.0007	0.0005	0.0011	prior_miscarriage_group: One or more
prior_miscarriage_group	One or more	ECE (10 bins)	20.0000	0.0687	0.0412	0.1172	prior_miscarriage_group: One or more

Table A.9: Feature-level fidelity statistics across outer splits.

feature	mean_ks	ks_lower	ks_upper	mean_smd
Drinving_z	0.5116	0.4147	0.6284	0.0477
Drunk_z	0.4709	0.4266	0.5123	0.1513
Sitting_z	0.3332	0.3179	0.3457	0.0336
Binking_z	0.3074	0.2890	0.3374	0.0331
Walking_z	0.3066	0.2868	0.3308	0.0306
bp_z	0.3051	0.2643	0.3331	0.0474
stress_z	0.2218	0.1874	0.2531	0.0301
Nmisc_z	0.1827	0.1638	0.2276	0.0525
Location_z	0.1625	0.1451	0.1892	0.0554
temp_z	0.1298	0.1125	0.1600	0.0408
BMI_z	0.1259	0.1097	0.1512	0.0375
Age_z	0.1044	0.0800	0.1278	0.0353
bpm_z	0.0940	0.0803	0.1085	0.0433
Alcohol.Comsumption_z	0.0673	0.0531	0.0894	0.0549
Activity_z	0.0000	0.0000	0.0000	

Table A.10: Complete constrained Bayesian-optimization history.

lambda_gp	lambda_mom	lambda_corr	lambda_dom	lambda_label	latent_dim	n_critic	prod_alpha	prod_log10_lambda	objective	auc	brier	logloss	ece10	c2st_auc	memorization_fraction	domain_violation_rate	constraint_c2st	constraint_mem	constraint_domain	phase	feasible	iteration
11.8270	0.6999	2.5638	2.4546	1.3086	10.0000	3.0000	0.9463	-3.2628	0.7873	0.6350	0.2394	0.6707	0.1357	0.8905	0.0000	0.9442	-0.0695	-0.0500	0.9242	initial	FALSE	1.0000
2.3451	2.2207	1.2877	0.9338	2.9712	10.0000	3.0000	0.5919	-2.1215	0.8128	0.6346	0.2393	0.6703	0.1863	0.8365	0.0000	0.8817	-0.0635	-0.0500	0.8617	initial	FALSE	2.0000
10.8138	1.1001	1.5943	1.3243	2.7369	16.0000	4.0000	0.0821	-4.0305	0.7927	0.6347	0.2393	0.6704	0.1462	0.8041	0.0000	0.8683	-0.0959	-0.0500	0.8483	initial	FALSE	3.0000
13.0801	2.6241	0.2842	0.5135	2.0383	28.0000	4.0000	0.7931	-2.5671	0.7996	0.6346	0.2385	0.6687	0.1618	0.8040	0.0000	0.9275	-0.0960	-0.0500	0.9075	initial	FALSE	4.0000
3.9984	0.4543	2.3368	1.7995	1.9673	13.0000	2.0000	0.6761	-3.6771	0.7863	0.6347	0.2401	0.6721	0.1316	0.9307	0.0000	0.9500	0.0307	-0.0500	0.9300	initial	FALSE	5.0000
8.9977	2.8558	0.7688	2.9183	1.2318	20.0000	4.0000	0.5024	-4.8638	0.7972	0.6336	0.2387	0.6692	0.1544	0.7978	0.0000	0.8825	-0.1022	-0.0500	0.8625	initial	FALSE	6.0000
4.2658	1.5470	0.1661	1.0091	0.1930	29.0000	2.0000	0.1000	-3.4538	0.7544	0.6309	0.2375	0.6670	0.0661	0.7583	0.0000	0.9333	-0.1417	-0.0500	0.9133	initial	FALSE	7.0000
10.3045	1.4586	1.8399	0.1978	0.3629	21.0000	4.0000	0.8814	-2.9367	0.7796	0.6356	0.2369	0.6656	0.1274	0.6992	0.0000	0.8733	-0.2008	-0.0500	0.8533	initial	FALSE	8.0000
5.5620	0.1291	2.9483	0.3981	2.3213	32.0000	3.0000	0.2933	-1.3765	0.7884	0.6346	0.2368	0.6655	0.1433	0.8145	0.0000	0.7992	-0.0855	-0.0500	0.7792	initial	FALSE	9.0000
14.9216	1.7553	2.0252	2.2315	0.8033	25.0000	2.0000	0.2308	-1.2018	0.7283	0.6343	0.2366	0.6653	0.0230	0.6778	0.0000	0.9567	-0.2222	-0.0500	0.9367	initial	FALSE	10.0000
7.9821	0.9266	0.5964	1.6245	1.6499	23.0000	1.0000	0.4645	-1.7395	0.7887	0.6349	0.2371	0.6660	0.1440	0.7968	0.0000	0.9950	-0.1032	-0.0500	0.9750	initial	FALSE	11.0000
7.1515	2.2558	1.0995	2.5344	0.6793	16.0000	5.0000	0.4069	-4.3903	0.7728	0.6349	0.2375	0.6669	0.1110	0.5552	0.0000	0.8350	-0.3448	-0.0500	0.8150	initial	FALSE	12.0000
14.4119	0.0557	1.4097	2.9987	1.6742	32.0000	5.0000	0.0285	-1.1880	0.7450	0.6335	0.2369	0.6657	0.0542	0.8360	0.0000	0.8758	-0.0640	-0.0500	0.8558	bo	FALSE	13.0000
12.9765	2.6275	2.7695	0.1758	0.1629	27.0000	1.0000	0.1238	-1.5339	0.7308	0.6345	0.2367	0.6655	0.0281	0.8457	0.0008	0.9717	-0.0543	-0.0492	0.9517	bo	FALSE	14.0000
2.4413	1.1629	2.9519	2.4262	2.9296	10.0000	5.0000	0.1745	-1.9043	0.8056	0.6345	0.2385	0.6687	0.1736	0.7686	0.0000	0.8708	-0.1314	-0.0500	0.8508	bo	FALSE	15.0000
2.0242	2.5495	2.4649	1.2038	1.0113	31.0000	5.0000	0.1632	-2.2291	0.7873	0.6339	0.2378	0.6674	0.1374	0.7711	0.0008	0.8650	-0.1289	-0.0492	0.8450	bo	FALSE	16.0000
3.2513	0.2922	0.0369	2.1934	0.3177	23.0000	4.0000	0.3363	-2.6850	0.7713	0.6330	0.2374	0.6667	0.1046	0.8559	0.0000	0.8858	-0.0441	-0.0500	0.8658	bo	FALSE	17.0000
9.8527	2.7232	0.2178	0.0149	0.0783	9.0000	1.0000	0.2930	-3.8258	0.7725	0.6343	0.2373	0.6664	0.1097	0.8406	0.0000	0.9525	-0.0594	-0.0500	0.9325	bo	FALSE	18.0000
4.2312	0.9978	0.0895	0.0742	2.5528	29.0000	5.0000	0.1580	-4.6759	0.7958	0.6355	0.2387	0.6691	0.1555	0.5293	0.0008	0.8083	-0.3707	-0.0492	0.7883	bo	FALSE	19.0000
14.9417	2.7246	2.6413	2.5146	0.5318	30.0000	1.0000	0.3712	-4.6950	0.7831	0.6346	0.2376	0.6670	0.1308	0.7909	0.0000	0.9533	-0.1091	-0.0500	0.9333	bo	FALSE	20.0000

Table A.11: Prediction-score resolution and change in AUC after calibration. Large AUC changes after isotonic calibration can arise when the calibration map creates many ties.

method	training_source	calibration_method	mean_auc_change	minimum_auc_change	maximum_auc_change	mean_test_n	median_unique_probabilities	minimum_unique_probabilities	maximum_unique_probabilities	mean_repeated_score_fraction
Real + Synthetic	augmented	none	0.0000	0.0000	0.0000	8001.0000	13.0000	13.0000	91.0000	0.9967
Real + Synthetic + Beta	augmented	beta	-0.0038	-0.0267	0.0131	8001.0000	13.0000	13.0000	91.0000	0.9967
Real + Synthetic + Isotonic	augmented	isotonic	0.0782	0.0749	0.0809	8001.0000	4.0000	4.0000	4.0000	0.9995
Real + Synthetic + Platt	augmented	platt	0.0000	0.0000	0.0000	8001.0000	13.0000	13.0000	91.0000	0.9967
Real + Synthetic + Temperature	augmented	temperature	0.0000	0.0000	0.0000	8001.0000	13.0000	13.0000	91.0000	0.9967
Real only	real	none	0.0000	0.0000	0.0000	8001.0000	13.0000	13.0000	13.0000	0.9984
Real only + Beta	real	beta	-0.0039	-0.0267	0.0131	8001.0000	13.0000	13.0000	13.0000	0.9984
Real only + Isotonic	real	isotonic	0.0782	0.0752	0.0809	8001.0000	4.0000	4.0000	4.0000	0.9995
Real only + Platt	real	platt	0.0000	0.0000	0.0000	8001.0000	13.0000	13.0000	13.0000	0.9984
Real only + Temperature	real	temperature	0.0000	0.0000	0.0000	8001.0000	13.0000	13.0000	13.0000	0.9984

Table A.12: All paired synthetic-versus-corresponding-real comparisons.

method	reference	metric	n_pairs	difference_ci	p_wilcoxon_bh
Real + Synthetic	Real only	AUC	20.0000	0.0000 (-0.0000, 0.0001)	1.0000
Real + Synthetic	Real only	Brier score	20.0000	-0.0008 (-0.0008, -0.0007)	<0.001
Real + Synthetic	Real only	Log loss	20.0000	-0.0018 (-0.0020, -0.0017)	<0.001
Real + Synthetic	Real only	|Calibration intercept|	20.0000	0.0004 (-0.0004, 0.0012)	0.7559
Real + Synthetic	Real only	|Calibration slope - 1|	20.0000	-0.2478 (-0.2588, -0.2370)	<0.001
Real + Synthetic	Real only	ECE (10 bins)	20.0000	-0.0707 (-0.0807, -0.0613)	<0.001
Real + Synthetic	Real only	ACE (10 bins)	20.0000	-0.0072 (-0.0165, 0.0016)	0.7510
Real + Synthetic + Platt	Real only + Platt	AUC	20.0000	0.0000 (-0.0000, 0.0001)	1.0000
Real + Synthetic + Platt	Real only + Platt	Brier score	20.0000	0.0000 (-0.0000, 0.0000)	0.9398
Real + Synthetic + Platt	Real only + Platt	Log loss	20.0000	0.0000 (-0.0000, 0.0000)	0.8568
Real + Synthetic + Platt	Real only + Platt	|Calibration intercept|	20.0000	0.0000 (-0.0000, 0.0001)	0.3461
Real + Synthetic + Platt	Real only + Platt	|Calibration slope - 1|	20.0000	-0.0001 (-0.0004, 0.0002)	0.7559
Real + Synthetic + Platt	Real only + Platt	ECE (10 bins)	20.0000	-0.0026 (-0.0085, 0.0020)	0.7559
Real + Synthetic + Platt	Real only + Platt	ACE (10 bins)	20.0000	-0.0089 (-0.0181, -0.0001)	0.4644
Real + Synthetic + Isotonic	Real only + Isotonic	AUC	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Isotonic	Real only + Isotonic	Brier score	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Isotonic	Real only + Isotonic	Log loss	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Isotonic	Real only + Isotonic	|Calibration intercept|	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Isotonic	Real only + Isotonic	|Calibration slope - 1|	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Isotonic	Real only + Isotonic	ECE (10 bins)	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Isotonic	Real only + Isotonic	ACE (10 bins)	20.0000	0.0000 (0.0000, 0.0000)	NA
Real + Synthetic + Beta	Real only + Beta	AUC	20.0000	0.0001 (0.0000, 0.0003)	0.1138
Real + Synthetic + Beta	Real only + Beta	Brier score	20.0000	0.0000 (0.0000, 0.0000)	<0.001
Real + Synthetic + Beta	Real only + Beta	Log loss	20.0000	0.0000 (0.0000, 0.0000)	<0.001
Real + Synthetic + Beta	Real only + Beta	|Calibration intercept|	20.0000	0.0000 (0.0000, 0.0001)	0.3461
Real + Synthetic + Beta	Real only + Beta	|Calibration slope - 1|	20.0000	0.0001 (-0.0001, 0.0004)	0.1429
Real + Synthetic + Beta	Real only + Beta	ECE (10 bins)	20.0000	-0.0014 (-0.0036, 0.0001)	0.0511
Real + Synthetic + Beta	Real only + Beta	ACE (10 bins)	20.0000	-0.0089 (-0.0179, -0.0010)	0.7559
Real + Synthetic + Temperature	Real only + Temperature	AUC	20.0000	0.0000 (-0.0000, 0.0001)	1.0000
Real + Synthetic + Temperature	Real only + Temperature	Brier score	20.0000	-0.0000 (-0.0000, 0.0000)	0.5241
Real + Synthetic + Temperature	Real only + Temperature	Log loss	20.0000	-0.0000 (-0.0000, 0.0000)	0.6339
Real + Synthetic + Temperature	Real only + Temperature	|Calibration intercept|	20.0000	-0.0009 (-0.0021, 0.0003)	0.1487
Real + Synthetic + Temperature	Real only + Temperature	|Calibration slope - 1|	20.0000	-0.0002 (-0.0005, 0.0002)	0.1549
Real + Synthetic + Temperature	Real only + Temperature	ECE (10 bins)	20.0000	-0.0068 (-0.0233, 0.0081)	0.7510
Real + Synthetic + Temperature	Real only + Temperature	ACE (10 bins)	20.0000	-0.0088 (-0.0182, -0.0001)	0.4091

Table A.13: Implementation and reproducibility settings.

Component	Setting
Software target	R $\geq$ 4.3; Python 3.11 environment accessed through `reticulate`; `keras3` with TensorFlow 2.20 backend. Exact installed versions should be archived with `sessionInfo.txt` and the Python environment manifest.
Master seed	1234. TensorFlow and R seeds were initialized from the same master seed; deterministic split-specific offsets were used for tuning candidates, outer splits, generator fitting, C2ST, and memorization diagnostics.
Data sample	50,000 records; tuning fraction 0.20; final evaluation fraction 0.80.
Outer validation	20 repetitions; 60% real training, 20% real calibration, and 20% held-out real test records in each split.
BO design	12 initial Latin-hypercube candidates and 8 sequential constrained-BO evaluations; candidate pool 1,200; surrogate noise 10^{-6}.
BO internal data	Internal BO training fraction 0.80; at most 8,000 real BO-training rows; 1,200 synthetic records per BO evaluation; 8 GAN epochs.
Objective weights	$w_{\mathrm{AUC}} = 1.00$, $w_{\mathrm{Brier}} = 0.50$, $w_{\mathrm{NLL}} = 0.35$, and $w_{\mathrm{ECE}} = 0.50$.
Feasibility limits	C2ST AUC ≤ 0.90; memorization fraction ≤ 0.05; raw domain-violation rate ≤ 0.02.
BO search bounds	$\lambda_{\mathrm{gp}} \in [2, 15]$; $\lambda_{\mathrm{mom}}, \lambda_{\mathrm{corr}}, \lambda_{\mathrm{dom}}, \lambda_{\mathrm{label}} \in [0, 3]$; latent dimension $k \in [8, 32]$; $n_{\mathrm{critic}} \in [1, 5]$; elastic-net $\alpha \in [0, 1]$; $\log_{10} \lambda_\phi \in [-5, -1]$.
Locked configuration	$\lambda_{\mathrm{gp}} = 14.922$, $\lambda_{\mathrm{mom}} = 1.755$, $\lambda_{\mathrm{corr}} = 2.025$, $\lambda_{\mathrm{dom}} = 2.232$, $\lambda_{\mathrm{label}} = 0.803$, $k = 25$, $n_{\mathrm{critic}} = 2$, $\alpha = 0.231$, and $\log_{10} \lambda_\phi = -1.202$.
Generator architecture	Concatenated latent and conditioning inputs; two fully connected hidden layers with 128 ReLU units each; linear output layer with dimension equal to the processed feature space.
Critic architecture	Concatenated record and conditioning inputs; two fully connected hidden layers with 128 units each and leaky-ReLU activation with negative slope 0.2; one linear output unit.
Auxiliary classifier	Fully connected layers with 64 and 32 ReLU units followed by one sigmoid output; Adam optimizer with learning rate 10^{-3}; 8 epochs; batch size 256.
GAN optimizer	Separate Adam optimizers for generator and critic; learning rate 10^{-4} for each; $\beta_1 = 0.5$ and $\beta_2 = 0.9$.
Outer GAN training	Batch size 64; 12 epochs; at most 12,000 real training records used for GAN fitting in each outer split; 4,000 synthetic records generated per split.
Calibration methods	None, Platt scaling, isotonic regression, beta calibration, and temperature scaling; each fitted only on the real calibration partition.
Calibration metrics	Equal-width ECE and tie-preserving equal-frequency ACE with 5, 10, 15, and 20 nominal bins; main reporting at 10 bins.
Inference	5,000 bootstrap replicates; 95% resampling intervals; paired t tests and Wilcoxon signed-rank tests; Benjamini–Hochberg adjustment.
Decision curves	Thresholds 0.05–0.50 in increments of 0.01; reporting thresholds 0.10, 0.20, 0.30, 0.40, and 0.50.