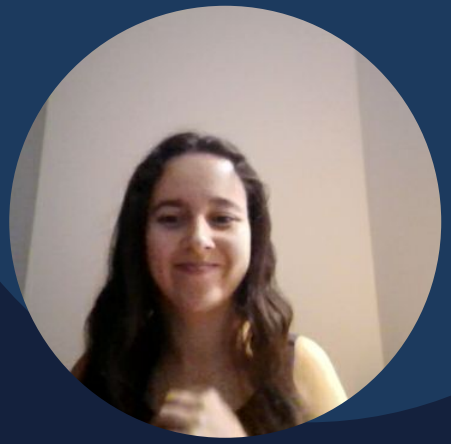


GRBIO RETREAT 2026 · SYNTHETIC DATA RESEMBLANCE

Complementary resemblance tests for synthetic data

A multiple-comparisons approach to combining marginal and joint discrepancy

Nora Amama-BenHassun

Supervisors: Daniel Fernández and Jordi Cortés

Research stay · Georgetown University, with Joshua Snoke and Claire Bowen

SIEMENS energy · IRIS Institute for Research and Innovation in Health · GRBIO · UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

OVERVIEW

Agenda

1. The problem: validating synthetic data resemblance
2. The first idea: a hybrid resemblance metric
3. From one metric to a union of tests
4. The framework and the multiplicity adjustments
5. Results: calibration, complementarity, statistical power
6. Future work

THE PROBLEM

Does the synthetic data resemble the original?

Resemblance is multi-faceted: marginals, pairwise dependence, and joint structure each tell a different story, and any single metric sees only part of the picture.

Marginal distributions

Per-variable shape: shifts in location, scale and tails.

Pairwise dependence

Correlation and association that marginal checks cannot see.

Joint global structure

Whether a model can still tell real from synthetic data.

The goal is not to report one p-value on its own. We want a global decision on resemblance, together with which components flagged the difference.

THE STARTING POINT

The first idea: a hybrid resemblance metric

This grows out of the previous objective. SPECKS works well, but what it detects depends on the classifier model behind it.

Classifier-based · joint

SPECKS

Joint: a classifier learns to tell real from synthetic, then we measure the gap in its scores.

$$\text{SPECKS} = \sup_t \left| \hat{F}^o_\pi(t) - \hat{F}^s_\pi(t) \right|$$

π = propensity score. CART and logistic regression probe different structure, which is what researchers and companies want.

Marginal · per-variable

Aggregated KS

Marginal: compares each variable's distribution, one at a time.

$$\text{KS}_j = \sup_x \left| \hat{F}^o_j(x) - \hat{F}^s_j(x) \right|$$

$$\text{KS}_{\text{mean}} = \frac{1}{q} \sum_{j=1}^{q} \text{KS}_j \qquad \text{KS}_{\text{max}} = \max_j \text{KS}_j$$

The first idea: combine these complementary signals into one hybrid resemblance score.

Standardizing the components under H_0

Each statistic is put on a common null scale before any combination.

Building blocks

Effective sample size

$$\lambda = \sqrt{\frac{n_o\, n_s}{n_o + n_s}}$$

Kolmogorov null moments

$$\mu_{KS} = \sqrt{\frac{\pi}{2}}\ln 2 \qquad \sigma^2_{KS} = \frac{\pi^2}{12} - \frac{\pi}{2}(\ln 2)^2$$

Mean-aggregated KS, null law ($\Sigma = I_a$)

$$\widehat{KS}^{mean}_{agg} \sim \mathcal{N}\left(\frac{\mu_{KS}}{\lambda}, \frac{\sigma^2_{KS}}{q\lambda^2}\right)$$

Standardized statistics

$$Z_S = \frac{\lambda\,\mathrm{SPECKS} - \mu_{KS}}{\sigma_{KS}}$$

$$Z_{KS,\,mean} = \frac{\widehat{KS}^{mean}_{agg} - \mu_{KS}/\lambda}{\sigma_{KS}/(\lambda\sqrt{q})} \sim \mathcal{N}(0,1)$$

Max variant

$$P(\widehat{KS}^{max}_{agg} \le x) \approx [\,F_K(\lambda x)\,]^q$$

$$Z_{KS,\,max} = \frac{\widehat{KS}^{max}_{agg} - \mu_{KS,\,max}}{\sigma_{KS,\,max}} \sim \mathcal{N}(0,1)$$

Common null scale: after standardizing, Z_S and Z_KS are each approximately N(0, 1) under H_0, so both move in the same range and can be compared or combined directly.

One standardized, weighted combination

THE COMBINED STATISTIC

$$M(w) = w \cdot Z_S + (1 - w) \cdot Z_{KS}$$

Z_S standardized SPECKS · Z_{KS} standardized aggregated KS (mean or max)

Aggregated KS, two variants

Mean

average of the per-variable KS gaps; stable, reflects typical mismatch.

Max

largest single gap; conservative, driven by the worst variable.

The weight $w \in [0, 1]$ tunes the balance. Both Z_S and Z_{KS} are approximately N(0, 1) under H_0, so they share the same range of values and the weighted sum is meaningful. Here α is the significance level and w the mixing weight.

This is where the work starts: SPECKS comes from the previous objective, so the new theoretical focus is the aggregated KS component and how it joins SPECKS.

WHY COMBINE

Two signals that fail in different places

Complementarity: neither component dominates the other, so combining them was meant to catch failures that either one alone would miss.

Classifier-based metrics

Strong on joint and global structure.

Can miss marginal misspecification: detecting it would need the propensity model itself to change.

Marginal metrics

Scale fast as marginals worsen.

Blind to correlation-structure distortion: a wrong dependence with right marginals slips through.

WHERE IT BROKE DOWN

Why one mixed number was not enough

The weight w

No principled, data-independent value; the conclusion can depend on a tuning choice.

Lost interpretability

A single blended score hides which aspect of resemblance actually failed.

Intractable null

Its null distribution has no general form; it shifts with the setting, and resampling it is hard too.

No error control

Each metric is its own test; blending or 'reject if any fails' controls nothing.

The flag that changed direction: a naive union, rejecting if any single test fails, inflates the Type I error under independence. Multiplicity has to be handled first.

THE PIVOT

From mixing scores to uniting tests

BEFORE · HYBRID METRIC

- One weighted statistic M(w)
- A tuning choice for the weight
- A single, hard-to-read number
- No control of false positives

AFTER · UNION OF TESTS

- Keep each metric as its own test
- Unite them into one decision
- Adjust for multiplicity
- See which test broke the union

Interpretable by design: the union condenses everything into one decision, yet it still reports which component rejected, so we keep per-test interpretability while controlling the error. The hybrid is the first idea whose drawbacks motivated this framework.

Multiplicity adjustments: Bonferroni via Bland & Altman (1995), BMJ 310:170 · Holm (1979), Scand J Statist 6:65-70 · Hochberg (1988), Biometrika 75:800-802
Simes (1986), Biometrika 73:751-754 · Hommel (1988), Biometrika 75:383-386 · Benjamini & Hochberg (1995), JRSS-B 57:289-300
Benjamini & Yekutieli (2001), Ann Statist 29:1165-1188 · Review: Chen, Feng & Yi (2017), J Thorac Dis 9(6):1725-1729

THE OBJECTIVE

A statistically valid way to combine metrics

Select

A complementary panel of resemblance metrics, not redundant ones.

Unite

Turn the panel into a single, well-defined union decision.

Control

Adjust for multiplicity so the Type I error stays at the nominal level.

GOAL

Develop a union-of-tests framework with multiplicity adjustment that controls Type I error while keeping the statistical power to detect synthetic data that truly differ.

This is the through-line of the PhD: a general, reusable validation procedure for synthetic tabular data.

Six adjustments, one union rule

GLOBAL DECISION RULE

$$\text{reject } H_0 \iff \min_j p_j^{adj} < \alpha$$

reject if at least one component rejects

FWER control · our default for a small panel

Bonferroni *single-step*

α/m valid under any dependence, the most conservative

Holm *step-down*

$\frac{\alpha}{m-i+1}$ any dependence, uniformly beats Bonferroni

Hochberg / Hommel *step-up (Simes)*

$\frac{\alpha}{m-i+1}$ positive dependence, more powerful

FDR control

BH

$$p_{(k)} \le \frac{k}{m}\alpha$$

independence or positive dependence

BY

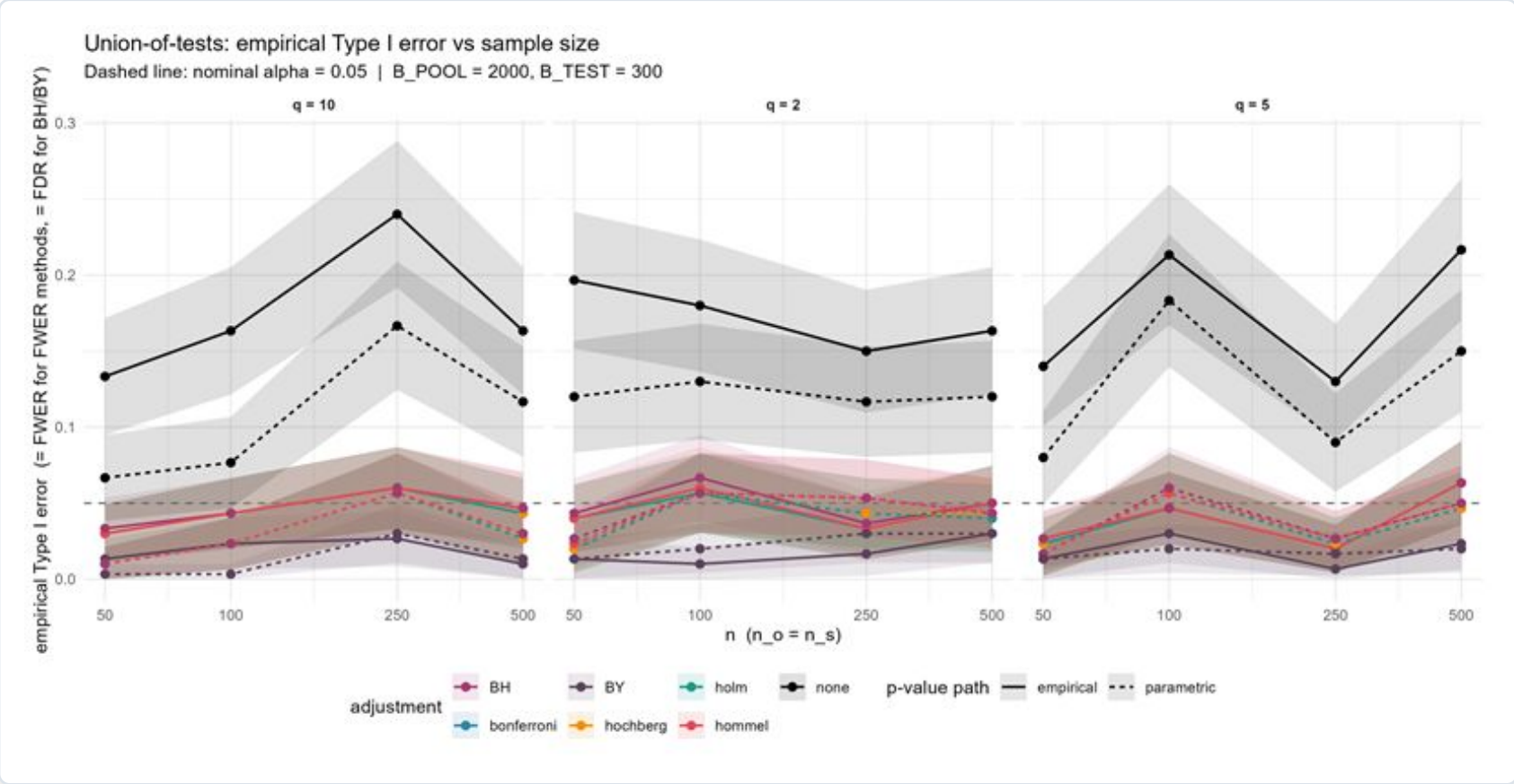
$$\frac{k}{m}\frac{\alpha}{c(m)},\quad c(m)=\sum_{i=1}^{m}\frac{1}{i}$$

any dependence, scales BH by c(m)

Ordered p-values p(1) ≤ … ≤ p(m), significance level α, panel size m. Under the global null these collapse to a few distinct rules; under the alternative the step-up methods can recover statistical power.

RESULTS · CALIBRATION

Naive union inflates, correction restores it



The y-axis is the Type I error estimated by simulation. Each method has two curves, the two ways of computing each test's p-value: resampling-based (labelled empirical in the legend) and parametric asymptotic. The dashed horizontal line marks the nominal α = 0.05.

0.24 → 0.05

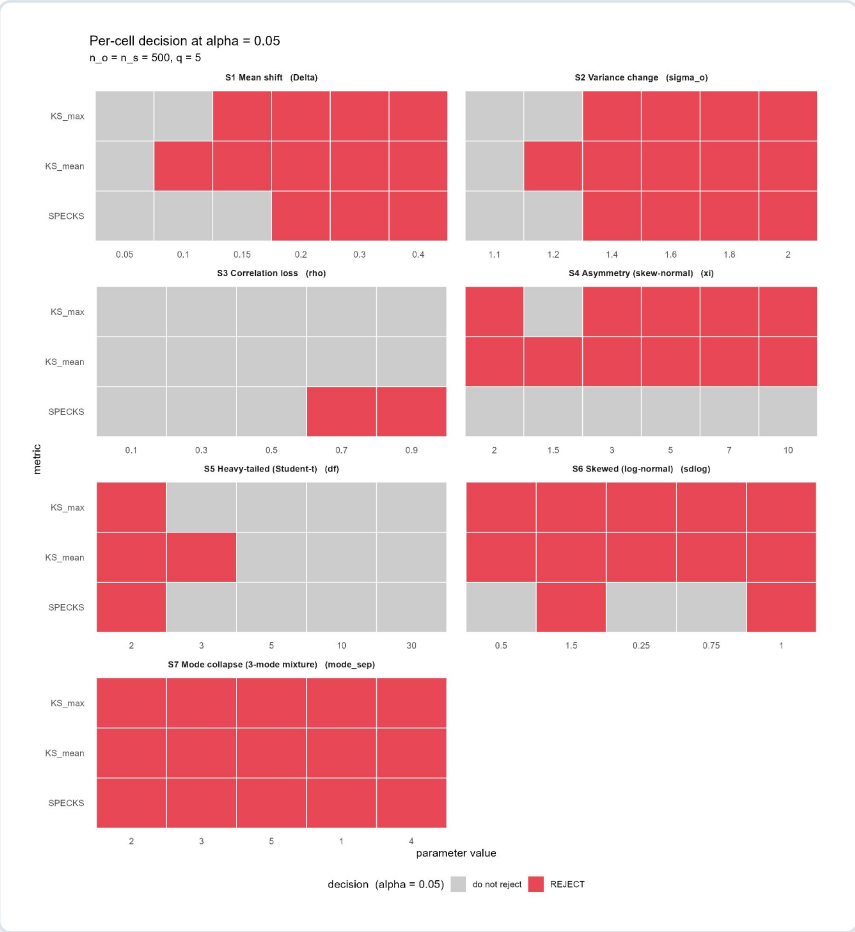
uncorrected (black) the union rejects far too often, peaking near 0.24; every FWER method and BH sit at the nominal line.

BY is too strict

it drops below α, losing statistical power. The FWER methods agree and are the safe default.

RESULTS · COMPLEMENTARITY

Different metrics catch different failures



Marginal KS

Catches location, scale, asymmetry, heavy tails and mode collapse (S1, S2, S4, S5, S7).

Classifier SPECKS

The only one that flags correlation loss (S3), which both KS variants miss entirely.

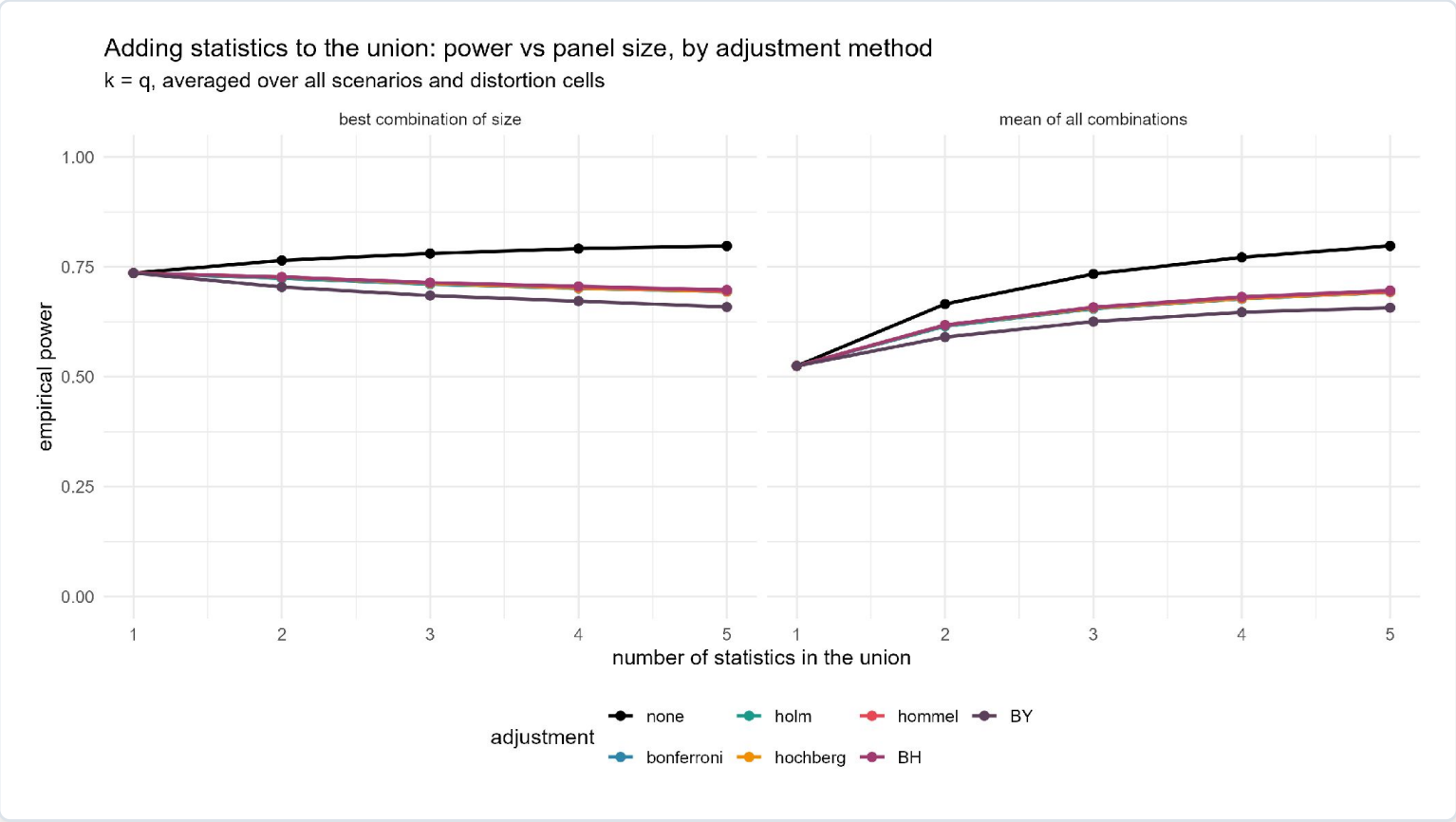
Together

The union covers all seven failure modes; no single metric does.

Gray means no rejection, red means rejection, read left to right as the distortion grows.

RESULTS · STATISTICAL POWER

More statistical power, same risk budget



+17 pp

statistical power vs one arbitrary metric

4 : 1
gain vs cost

−4 pp
from the oracle

Empirical statistical power as statistics are added to the union, averaged over all scenarios and distortion cells. The union recovers most of the oracle's power at a small cost, and adjustment barely changes it.

ON THE HORIZON

Extending and validating the framework

Integrate pMSE

Add the classifier-based pMSE results into the panel and the union decision.

REALISTIC PLAN · EARLY OCT 2026

Cramér–von Mises on copulas

A classifier-free test that isolates dependence after removing marginals via ranks.

CONCEPTUAL PROPOSAL · NOV 2026

Real datasets

Public data synthesized with different algorithms, the second simulation objective.

REALISTIC PLAN · EARLY OCT 2026

R package on CRAN

Full metric panel, adjustments, decision and diagnostics, with paired toy datasets.

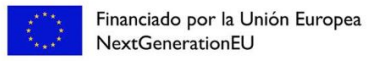
TOOLING

Positioning: the combined, multiplicity-adjusted procedure has to outperform the sequential inspection baseline (Raab, 2021).

Rémillard & Scaillet (2009), Testing for equality between two copulas, J. Multivariate Anal. 100(3), 377–386 · R package TwoCop.

Acknowledgements

This thesis is funded by the Siemens Energy AI Chair, Energy Sustainability for a Decarbonized Society 5.0 (TSI-100930-2023-5), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial within the Cátedras ENIA 2022 call, with support from the European Union (Next Generation EU).



This research was funded by MICIU / AEI /10.13039/501100011033 (Spain) and by FEDER (EU) [PID2023-148033OB-C21], and by grant 2021 SGR 01421 (GRBIO) administered by the Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain).

Funding support is also acknowledged to the project Generation of Reliable Synthetic Health Data for Federated Learning in Secure Data Spaces (PID2022-141045OB-C42, AEI/ERDF, EU), funded by MCIN / AEI /10.13039/501100011033 and by ERDF, a way of making Europe (European Union).

CLOSING

From isolated metrics to a unified validation framework

What began as a hybrid resemblance metric is now a general union-of-tests procedure that validates synthetic tabular data with controlled error and a reusable open-source tool.