Research Square

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

Measurement Reliability Confounds Cross-Regime Comparisons of Postprandial Glucose Prediction: Evidence from Bidirectional Cross-Cohort Transfer

Yulin Zhao

202334071131@stu.huel.edu.cn

Henan University of Economic and Law https://orcid.org/0009-0005-4623-8465

Research Article

Keywords: postprandial glucose; continuous glucose monitoring; measurement reliability; leave-one-subject-out; cross-cohort transfer; personalized nutrition

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11558064/v1

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

Additional Declarations: The authors declare no competing interests.

PREPRINT · NOT CERTIFIED BY PEER REVIEW

Measurement Reliability Confounds Cross-Regime Comparisons of Postprandial Glucose Prediction: Evidence from Bidirectional Cross-Cohort Transfer

Elijah Zhao

School of Data Science and E-Commerce, Henan University of Economics and Law, Zhengzhou 450046, China

ORCID: 0009-0005-4623-8465

29 October 2026

ABSTRACT

Postprandial glucose response (PPGR) prediction is a central promise of personalized nutrition, yet published models are rarely evaluated against the reliability of the measurement they try to predict. We reproduce and extend the CGMacros baseline (45 subjects, 1,699 meals; leave-one-subject-out) and subject it to a cross-cohort stress test on BIG IDEAs (16 subjects, 656 meals). The replication matches the published breakfast result (iAUC r = 0.655 against ≈ 0.64) and extends it to all analysable meals (iAUC 0.451; AUC 0.838). Duplicate reference breakfasts in CGMacros give a repeatability of ρ = 0.799, a ceiling the model stays below (√ρ ≈ 0.894) and so does the published baseline. A frozen model transfers forward at r = 0.227 for iAUC; trained on the external cohort and transferred back it reaches 0.085, so iAUC transfer fails in both directions. AUC behaves differently: it survives transfer (0.569 forward, 0.748 reverse) and is largely a baseline-glucose quantity — pre-prandial glucose alone recovers 0.807 of AUC across all meals, 0.861 at breakfast, and 0.963 at the person level.

Read against its own regime, the within-cohort breakfast model captures 0.54 of the reliable variance, a same-device cross-population transfer in a prior study also 0.54, and the cross-device transfer 0.11. Cross-regime comparisons of PPGR models are therefore confounded by measurement reliability and cannot be read as comparisons of model quality: reliability explains part of the gap, and the remainder is attributable to features, device, annotation and model class in proportions this design cannot separate. The claim is conditional and falsifiable, and we state the study that would overturn it. We recommend that PPGR reports name their measurement regime, place a reliability estimate beside every correlation, and report per-subject alongside marginal interval coverage.

Keywords: postprandial glucose; continuous glucose monitoring; measurement reliability; leave-one-subject-out; cross-cohort transfer; personalized nutrition

1. Introduction

1.1 What PPGR prediction promises, and what it is measured on

Two people can eat the same meal and mount very different glucose excursions. Zeevi et al. (2015) turned that observation into a research programme: an individual's postprandial response was reproducible enough to be predicted, and features of the person—microbiome, clinical markers—carried part of the signal. The measurement substrate has since changed. Continuous glucose monitoring (CGM) records the postprandial curve at 5–15 min resolution, and public cohorts now pair those traces with meal records. CGMacros (Das et al., 2025) contributes 45 adults, 1,699 meals, blood-panel markers and anthropometrics, measured on an Abbott Libre Pro with study-recorded meal annotations. BIG IDEAs (Cho et al., 2023; Bent et al., 2021) contributes 16 adults and 656 meals on a Dexcom G6, with free-living, self-reported food logs. Predicting the postprandial glucose response (PPGR)—2-h incremental area under the curve (iAUC), total AUC, or peak rise—from meal composition and subject phenotype is therefore an active supervised-learning problem with public data.

1.2 The evaluation gap

Published PPGR models are scored with correlations and errors against the observed response: r, R^2, RMSE. Reliability studies of the same meals are scored with intraclass correlations and duplicate-meal repeatability. The two literatures measure the same object—how well a prediction tracks a noisy observation—yet almost never appear on the same page. The separation has a consequence that is easy to miss. Under the classical measurement model, a correlation with a single observation cannot exceed the reliability of that observation, and a model that uses only stable information is bounded by $\sqrt{\rho}$, where ρ is the repeatability of the outcome. Part of an apparently modest r may therefore be a property of the measurement rather than of the model.

The size of that ceiling is not a biological constant. Hengist et al. (2025) presented duplicate meals about a week apart in an inpatient feeding study of 30 adults without diabetes and found $r = 0.46$ (Abbott Libre Pro) and 0.45 (Dexcom G4 Platinum), with ICCs of 0.28 and 0.17. Those are not free-living numbers, and because free-living self-report adds portion and timing error they are an *optimistic* benchmark for a free-living target; they nevertheless show that duplicate meals, even under controlled presentation, agree far from perfectly. In the CGMacros cohort the same device and the same participants yield 0.799 for a repeated reference breakfast but 0.374 and 0.376 for free-choice dinners and snacks (Table 1). If ρ moves by a factor of two inside one cohort, it can move across cohorts, and a cross-cohort comparison that does not measure ρ in both regimes cannot separate model quality from measurement quality.

One study made this reading explicitly. Mendes-Soares et al. (2019) used their cohort's duplicate-meal reproducibility ($R = 0.66$, $n = 111$) as an upper bound on explainable variance and read their cross-cohort transfer ($R = 0.596$ trained on the Israeli cohort alone; 0.618 with both cohorts combined) against it. Their design licensed that reading: both cohorts used the same device family (iPro2) and comparable logging protocols, so the ceiling could be treated as stable. A transfer that changes device and annotation protocol at the same time does not inherit that licence.

Two further complications compound the problem. First, AUC and iAUC are not interchangeable: with a 2-h window and a 15-min grid, AUC ≈ 120 × baseline glucose + iAUC, so a high AUC correlation can be earned by predicting who is fasting high rather than what was eaten (§2.3). Second, cross-cohort comparisons change device, annotation protocol and population at once, which makes a performance drop between cohorts unattributable by construction.

Concurrent work has improved PPGR accuracy with new inputs: meal images combined with clinical and microbiome variables (Kondratyeva et al., 2026; MICCAI 2026 MultiTab workshop), and frozen CGM foundation-model representations combined with meal nutrition and subject context (Li et al., 2026; GlucoFM, Google Research). Both are evaluated on CGMacros and both target accuracy. The reliability question is orthogonal to the input modality, and to our knowledge it remains unaddressed: no published study connects a change in measurement reliability across regimes to the collapse of cross-cohort PPGR performance, or tests that connection in both directions. That is the gap this paper fills.

1.3 Contributions

C1 — Reproduction and extension. Under strict leave-one-subject-out (LOPO) validation, the published CGMacros baseline is reproduced on breakfast iAUC (r = 0.655 against a published ≈0.64) and breakfast AUC (r = 0.890 against ≈0.89), then extended to all meal types (1,557 meals) and three targets: iAUC 0.451, AUC 0.838, peak rise 0.542. Code, fixed-seed experiment definitions and all derived tables are released; field audits find that only 23.9% of 67 deep-learning glucose-prediction papers released code (Lu et al., 2026).

C2 — A within-cohort reliability ceiling. Duplicate reference breakfasts in CGMacros correlate at r = 0.799 (95% CI 0.702–0.867; 174 pairs from 44 subjects; ICC 0.798), whereas duplicate free-choice dinners and snacks sit at 0.374 and 0.376 and no duplicate lunches exist. The model stays below both this conservative benchmark and the corresponding √ρ bound, as does the published baseline. The contrast separates a standardized-protocol regime from a free-choice regime inside one device and one cohort.

C3 — Bidirectional frozen-model transfer. A model frozen on CGMacros transfers to BIG IDEAs at r = 0.227 (iAUC), 0.569 (AUC) and 0.223 (peak); a model trained on BIG IDEAs transfers back at 0.085, 0.748 and 0.174. iAUC transfer fails in *both* directions, which rules out a sample-size explanation based on BIG IDEAs' 16 subjects, since the 45-subject cohort also fails when it is the target. Forward, a carbohydrate-only Ridge regression transfers better on iAUC (0.362) and peak (0.357) than the tree ensemble, but worse on AUC (0.288), and its iAUC R^2 is negative (−0.040): ranking better, calibration worse.

C4 — The baseline-glucose ablation. Removing pre-meal glucose leaves iAUC skill essentially unchanged (0.451 → 0.457) but cuts AUC from 0.838 to 0.638; baseline glucose alone reaches 0.807–0.861 on meal-level AUC and 0.963 on person-mean AUC. The AUC signal is therefore largely a fasting-state shadow, whereas iAUC carries meal-level information that is not circular.

C5 — Person-level aggregation and uncertainty. Aggregating predictions to the person level sharpens within-cohort rankings (iAUC 0.451 → 0.501; AUC 0.838 → 0.921) but does not rescue cross-cohort iAUC (r = −0.03, 95% CI [−0.43, 0.43]); cross-cohort person-mean

AUC reaches 0.715 [0.353, 0.898]. Distribution- free conformal intervals attain their nominal marginal coverage (0.800 and 0.898) yet fall to 0.00 for some individuals, and a mixed model locates 69.4% of AUC variance between subjects versus 32.8% for iAUC and 32.1% for peak rise.

C6 — An interpretive claim with consequences. Cross-regime comparisons of PPGR models are confounded by measurement reliability and cannot be read as comparisons of model quality. Reliability explains part of the gap between published numbers and ours; the residual reflects model, feature and cohort differences that reliability alone cannot capture. We instantiate the claim as a same-device versus cross-device contrast—Mendes-Soares et al.'s 0.596 under a measured, stable iPro2 ceiling versus our 0.227 under a device- and annotation-shifted regime whose ceiling we cannot measure—and we state explicitly that the two numbers are not like-for-like statistics (§4.5).

1.4 Scope and non-claims

We do not propose a new predictor. The model is deliberately the published baseline, and the contribution is the interpretation of its numbers, not its architecture. We make no causal claims: the features are observational, and nothing here implies that changing a macronutrient will change a response. The reliability framework bounds models that rely on stable information; models that consume personal CGM history or adaptive same-day baselines face a different, higher bound and are outside our scope (§7). We measure ρ in CGMacros only; the BIG IDEAs regime's ρ is unmeasured, and we treat that as part of the argument rather than as a gap to be papered over. Finally, one external cohort cannot establish that linear models generalise better than tree ensembles; the carbohydrate-only result is a cautionary observation, not a ranking.

Section 2 develops the framework, §3 the within-cohort analysis, §4 the cross-cohort stress test, §5 uncertainty and variance decomposition, §6 the implications for readers, builders and study design, and §7 the limitations.

2. A measurement-reliability framework for PPGR prediction

2.1 What a reliability ceiling is, and what it is not

Postprandial glucose response (PPGR) is not observed directly. We adopt the standard classical measurement model. For meal i in person j, the measured response X_{ij} (e.g., iAUC) can be written as $X_{ij} = T_{ij} + \varepsilon_{ij}$, where T_{ij} is the latent response under a noiseless sensor and an identical physiological state, and ε captures sensor error, site effects, calibration drift, and unobserved day-to-day physiology. When the same meal is eaten twice under the same protocol, the correlation between the two measurements is the repeatability ρ. Under the parallel-measurement assumption (equal error variance across repeats), a model that predicts T_{ij} from stable information (meal composition, person characteristics) can correlate with a single observed X_{ij} at most √ρ. Using ρ itself as the benchmark is therefore conservative. A model that also uses same-day information (e.g., pre-meal glucose) can exceed √ρ, because that information belongs to the measurement context rather than the stable meal signal.

ρ is an empirical, regime-specific quantity. It depends on the sensor, the meal protocol, the population, the meal type, and the level of aggregation. It is not a biological constant, and

it is not a strict upper bound on every model. We therefore use ρ as the primary, directly estimable benchmark and report √ρ as the model-relevant bound. A claim that a model "beats the ceiling" is meaningful only relative to the appropriate bound and regime; it is not, by itself, evidence of a super-measurement model.

2.2 The reliability ladder

Table 1 collects published duplicate-meal reliabilities and the estimate from this cohort. All values are correlations on iAUC-family targets unless noted.

Table 1. Duplicate-meal reliability across protocols and devices.

Study	Cohort / device	Protocol	ρ	Notes
Zeevi et al. 2015	Israeli; iPro2	Four standardized meals	0.71–0.77	—
Mendes-Soares et al. 2019	Israeli → Midwestern; iPro2	Standardized bagel and cream cheese	0.66	n = 111
This study (E2)	CGMacros; 45 subjects; Libre Pro	Reference breakfast (six macro profiles)	**0.799**	95% CI 0.702–0.867; 174 pairs from 44 subjects
This study (E2)	CGMacros; 45 subjects; Libre Pro	Same-type all meals	0.746	95% CI 0.657–0.807; 291 pairs
This study (E2)	CGMacros; 45 subjects; Libre Pro	Dinner	0.374	95% CI 0.088–0.569; 40 pairs from 13 subjects
This study (E2)	CGMacros; 45 subjects; Libre Pro	Snack	0.376	95% CI 0.069–0.656; 77 pairs from 18 subjects
This study (E2)	CGMacros; 45 subjects; Libre Pro	Lunch	—	No duplicate pairs
Hengist et al. 2025	Inpatient feeding study; Abbott Libre Pro	Presented duplicate meals, ~1 wk apart	0.46	95% CI 0.38–0.54; ICC 0.28
Hengist et al. 2025	Inpatient feeding study; Dexcom G4 Platinum	Presented duplicate meals, ~1 wk apart	0.45	95% CI 0.39–0.50; ICC 0.17

The ladder spans 0.37–0.80 (Figure 1A), a more than twofold range in ρ and a roughly 1.5-fold range in √ρ. The within-CGMacros contrast is the cleanest evidence that the ceiling is not a fixed biological property: with device and population held constant, breakfast is highly repeatable (0.799), whereas free-choice dinners and snacks are not (0.374 and 0.376). The breakfast records collapse to six macronutrient profiles and same-subject breakfasts match exactly (§3.3), so they behave as a repeated reference meal; whether that reflects a provided menu or a strongly habitual meal, the contrast with free-choice meals is a protocol contrast. Across studies, device, population, and protocol co-vary, so the ladder is descriptive rather than a causal decomposition; we do not attribute cross-study differences to any single factor.

One attribution matters for what follows. The Hengist et al. 2025 estimates come from an inpatient feeding study in 30 adults without diabetes, with duplicate meals presented about a week apart — not from free-living conditions, and not from BIG IDEAs. They are the closest device-level estimates available for the target device family, and because free-living self-report adds portion and timing error, they are an *optimistic* benchmark for a free-living target. We therefore never treat 0.45–0.46 as a measured ceiling for BIG IDEAs; we use it only as a generous reference point in §4.5.

Data note. The all-meals estimate pools same-type pairs only (0.746); allowing cross-type pairs, as in an earlier version of the analysis, gives 0.715. The breakfast estimate is unchanged. Because one subject can contribute many pairs, the 95% CI is a subject-clustered bootstrap (resampling subjects and retaining all their pairs); a naive pair-level CI

would be too narrow. As a subject-balanced sensitivity check, drawing one pair per subject per replicate (2,000 draws) gives 0.803 [0.684, 0.888] for breakfast and 0.746 [0.559, 0.867] for all meals, so the estimate is not an artefact of a few heavily represented subjects. For duplicate-meal reliability, we report r rather than ICC because the pairs are not independent.

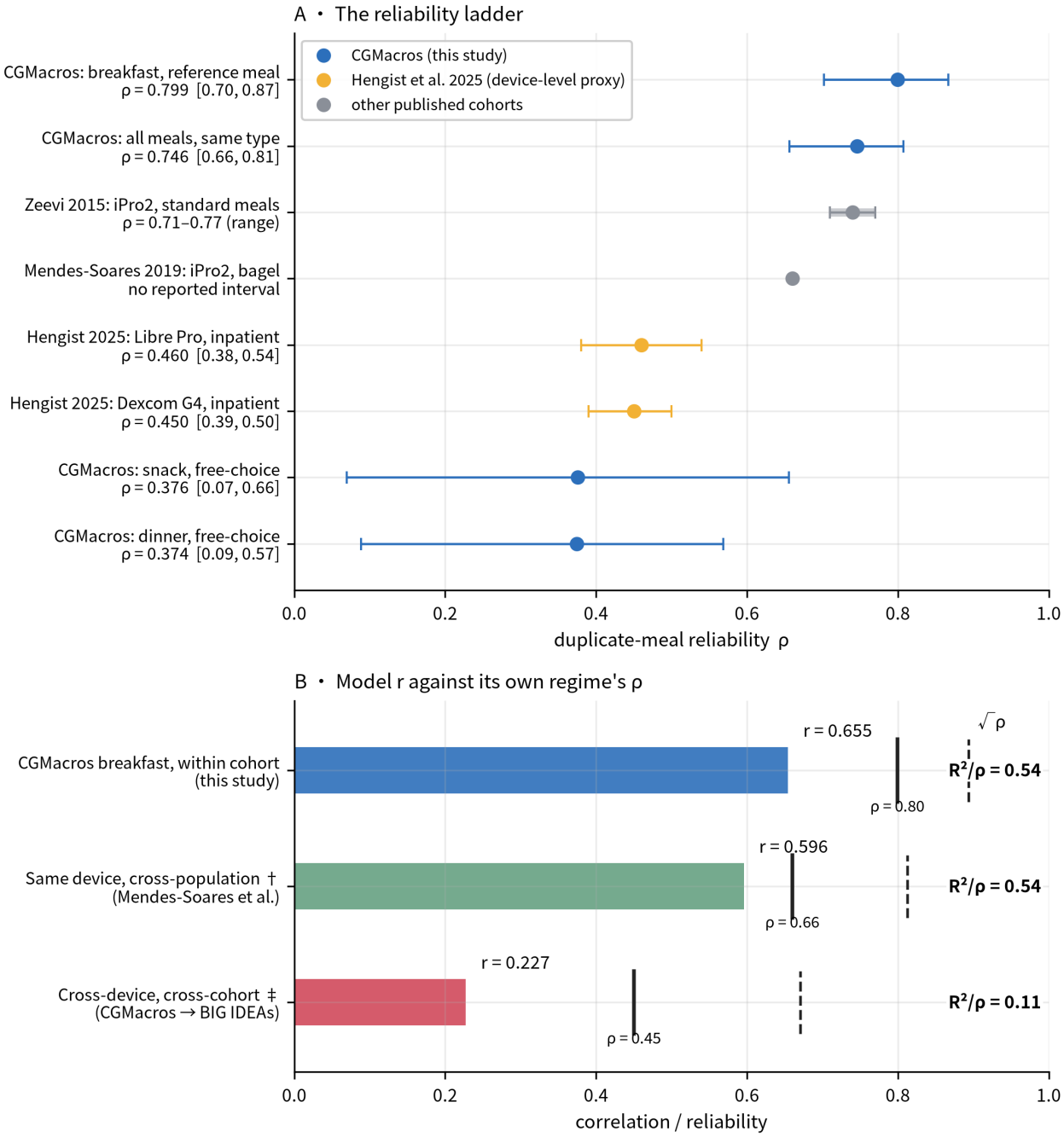


Figure 1. Duplicate-meal reliability and the position of three model correlations. (A) Duplicate-meal reliability ρ, the Pearson correlation between repeated presentations of the same meal. Markers are point estimates; horizontal bars are 95% CIs where the source reports one (Zeevi et al. report a range, drawn as a thick segment). Lunch is absent because CGMacros contains no duplicate lunch pairs. **(B)** Model r for three regimes, each plotted against the reliability of the regime it was evaluated in: solid vertical line, ρ; dashed line, √ρ, the maximum correlation available to a model that uses stable meal and person information only (§2.1); bar, the model's r; right-hand column, R^2/ρ, the share of reliable variance captured. † The ρ for the same-device transfer is measured on a bagel duplicate in the target cohort, not on the transfer meals. ‡ The reliability of BIG IDEAs was not measured; the ρ shown is the Hengist et al. (2025) Dexcom G4 estimate, used as a device-level proxy (§4.5). Panel B is an ordering, not a like-for-like statistic: the three fractions rest on different cohorts and, in the third row, on a proxy rather than a measurement. *Sources:* `ceiling_results.csv` (E2), `results.csv` and `external_results.csv`; published values as in Table 1.

2.3 The AUC identity trap

AUC and iAUC are not interchangeable. With a 2-h window and a 15-min trapezoidal grid, AUC ≈ 120 × baseline glucose + iAUC, where the first term is fasting state and the second

is the excursion above baseline. The identity is exact when glucose never falls below baseline and approximate otherwise. A model evaluated on AUC can therefore achieve a high correlation by predicting who is fasting high, without learning meal composition. iAUC removes this structural dependence. This is an accounting identity, not an empirical regularity; it motivates iAUC as the primary target and predicts that baseline-only models will perform well on AUC and poorly on iAUC.

2.4 What the framework predicts—and what it does not

The framework yields four testable expectations.

1. *Within cohort.* A model using stable information should sit below the cohort/protocol's √ρ, and typically below ρ. The gap is the stable signal the model fails to capture.
2. *Across regimes.* Model performance should fall when the target regime's ρ falls, but the fall need not be proportional. Feature and model mismatch add a residual that reliability alone cannot explain.
3. *Aggregation.* Averaging repeated measurements increases reliability (Spearman–Brown), so subject-level or meal-averaged correlations can rise without any improvement in meal-level prediction.
4. *Target dependence.* Because AUC is baseline-dominated, model rankings can differ between AUC and iAUC; iAUC is the appropriate target for meal-composition claims.

The framework does not claim that meal composition is unimportant, that ρ is constant across individuals, or that reliability is the sole cause of cross-cohort collapse. It claims that reliability is a first-order confound: a cross-regime comparison that does not measure ρ in both regimes cannot be read as a comparison of model quality. The claim is falsifiable. If a model's performance were invariant across regimes with very different ρ, the confounding account would be weakened. The framework also does not apply unchanged to models that use same-day baseline or continuous CGM; for those, the relevant bound includes same-day information and is therefore higher.

2.5 Relation to prior work

Mendes-Soares et al. 2019 used duplicate-meal reproducibility as an upper bound for model performance within a single device family (iPro2 on both sides), and found that their transfer R = 0.596/0.618 sat near their cohort's ceiling R = 0.66 on their own accounting (on the √ρ bound used here, the same transfer reaches 0.54 of the reliable variance; §4.5). The ceiling concept is theirs. We add: (i) a cross-device transfer in which the target regime's reliability is unmeasured and cannot be assumed equal to the source's; (ii) bidirectional transfer, showing that the iAUC failure is symmetric; (iii) the AUC identity, which separates baseline state from meal response; (iv) a subject-level aggregation analysis; and (v) a subject-clustered reliability estimate. The cross-regime confounding argument is ours; the ceiling concept is not.

3. Within-cohort analysis (CGMacros, 45 subjects / 1,699 meals)

3.1 Cohort, targets and evaluation protocol

The CGMacros release provides continuous glucose monitoring (CGM) traces with time-stamped meal annotations and macronutrient estimates for 45 adults, plus a subject-level

table of laboratory and anthropometric measurements. The cohort spans the glycemic range from healthy to type-2 diabetes (median HbA1c 5.9%, range 4.6–8.5%), which is deliberate: a model must handle a wide dynamic range rather than a narrow clinical subgroup.

Each meal annotation defines a two-hour postprandial window sampled every 15 minutes (nine points). We compute three targets: the **incremental area under the curve** (iAUC), the positive trapezoidal area above the pre-meal baseline; the **total area** (AUC) over the same window; and **peak rise**, the maximum elevation above baseline. Following the official replication protocol, all prediction analyses keep only meals with iAUC > 0, leaving 1,557 of 1,699 meals (breakfast 423, lunch 380, dinner 451, snack 303). The dependency AUC ≈ 120 × baseline + iAUC (§2.3) holds on this grid and motivates treating AUC and iAUC as separate targets rather than two views of one quantity.

The feature set has 19 columns: the four macronutrients (carbohydrate, protein, fat, fibre), the meal-time pre-prandial glucose, and 14 subject-level variables (age, sex, BMI, HbA1c, fasting glucose, insulin, HOMA-IR, triglycerides, total cholesterol, HDL, non-HDL, LDL, VLDL, and the cholesterol/HDL ratio). Subject variables are joined on subject id, never positionally. Participants wore two sensors concurrently; we use the **Libre Pro** channel (15-minute sampling), matching the official baseline's `Baseline_Libre` feature, so that the replication is like-for-like and the device regime is explicit.

Evaluation is strictly **leave-one-subject-out** (LOPO). In each of the 45 folds, every meal from one subject is held out; imputation and standardisation are fit on the training folds only, and metrics are computed on the pooled out-of-fold predictions. This matters because subject-level response baselines are strongly individual: a random meal-level split would let the same person appear on both sides and would overstate performance. We compare four models: a fold-mean predictor (the honest no-skill floor), a carbohydrate-only Ridge regression, an energy-only Ridge regression on carbohydrate, protein and fat, and XGBoost (Chen & Guestrin, 2016) with the official hyperparameters (depth-1 stumps, 80 rounds, learning rate 0.2, L1 penalty 1.0). Ridge, not ordinary least squares, is used for the linear baselines: with 19 collinear features and subjects carrying extreme values, unregularised OLS extrapolates catastrophically on a held-out person.

The primary metric is Pearson *r* between observed and predicted responses, which is what the published baseline reports and therefore what makes the replication directly comparable. We also report Spearman ρ, R^2, RMSE and MAE; R^2 and RMSE are on the original scale (iAUC in mg/dL·min) and are more sensitive to the long right tail, so we treat them as context rather than headline. To connect continuous predictions to classification-style reporting, we add a median-split ROC-AUC computed from the same out-of-fold predictions. All seeds are fixed and the full pipeline runs with one command (Section 9 of the technical report, which ships with the code repository as `research/reports/technical_report.pdf`).

3.2 Replicating the published baseline and extending it to all meal types

On the breakfast subset (423 meals, 45 subjects), XGBoost reaches **r = 0.890 for AUC** and **r = 0.655 for iAUC**, against the published ≈0.89 and ≈0.64 (Table 2). We treat the agreement as validation of the whole chain — meal extraction, target definitions, the iAUC > 0 filter, and the LOPO split — rather than as a merely similar-looking re-run.

Table 2. LOPO Pearson r by model, subset and target.

Model	Breakfast iAUC	Breakfast AUC	Breakfast peak	All-meal iAUC	All-meal AUC	All-meal peak
Mean predictor	−0.820	−0.924	−0.826	−0.553	−0.780	−0.552
Carb-only Ridge	0.107	−0.014	0.079	0.147	−0.032	0.150
Energy-only Ridge	0.255	0.100	0.247	0.143	−0.035	0.146
XGBoost (official)	**0.655**	**0.890**	**0.632**	**0.451**	**0.838**	**0.542**

Breakfast n = 423; all meals n = 1,557. Full table including Spearman ρ, R^2, RMSE, MAE and ROC-AUC: `research/experiments/results.csv`.

The negative correlation of the mean predictor is a structural diagnostic, not a modelling failure. Under LOPO, each fold predicts the mean of the *other* 44 subjects; a subject with an unusually large response pulls that training mean down precisely because they are excluded, which induces an anti-correlation. The sign confirms how much of the pooled variance is between people, and it is one reason a subject-wise split is mandatory.

The linear baselines expose the shape of the problem. Carbohydrate alone barely correlates with breakfast iAUC (r = 0.107); adding protein and fat helps (r = 0.255), consistent with fat and protein modulating gastric emptying and the shape of the excursion. Yet the best linear model explains about 6% of iAUC variance (R^2 = 0.062) against 43% for XGBoost (R^2 = 0.427). That gap is the empirical case for a non-linear model on this feature set.

Removing the breakfast restriction to all 1,557 meals costs 0.20 in iAUC correlation (0.655 → 0.451) but only 0.05 in AUC (0.890 → 0.838), with peak rise at 0.542. Two readings are possible: either AUC is a more portable target, or AUC is partly a baseline-glucose signal that survives any meal mix. Section 3.4 shows the second reading is correct. Discrimination remains strong on every target — median-split ROC-AUC is 0.875 for breakfast iAUC and 0.744 for all-meal iAUC (0.956 and 0.910 for AUC) — so the lower iAUC correlation is not a threshold or scale artefact: the model genuinely ranks meals less reliably for the excursion than for the total area.

3.3 The ceiling check

The framework's first question is whether r = 0.655 is near the best achievable under this protocol. The cohort provides an unusually clean answer. The 429 breakfasts in the unfiltered table collapse to **six distinct macronutrient profiles**, and every same-subject breakfast pair is an *exact* match on all four macronutrients: 174 pairs at ±2 g, and the same 174 at ±5 g. Breakfast in CGMacros is therefore effectively a repeated reference meal, and its duplicate-meal correlation is a genuine test–retest estimate rather than a loose nutrient matching. For dinners and snacks the matching is approximate (40 pairs at ±2 g versus 95 at ±5 g), so those rows answer a different and noisier question.

Duplicate-meal reliability for breakfast iAUC is **ρ = 0.799** (95% CI 0.702–0.867; 174 pairs from 44 subjects; Table 1). The prediction analyses use only iAUC > 0 meals, so we also computed the matched estimate on pairs with iAUC > 0 on both sides: **0.810** (0.715–0.876; 169 pairs). The model's 0.655 lies below both estimates, and below the parallel-measurement bound $\sqrt{\rho} \approx 0.894$ as well. There is no "model beats measurement" paradox, and the published baseline (≈0.64) is consistent for the same reason. The comparison is

conservative: using the unfiltered 0.799 as the benchmark understates the matched reliability.

Reliability is strongly meal-type-dependent. With device and cohort held constant, dinner duplicates correlate at 0.374 (40 pairs, 13 subjects) and snack duplicates at 0.376 (77 pairs, 18 subjects), against 0.799 for the reference breakfast. The contrast is between a repeated reference meal and free-choice meals, so it cannot identify which protocol feature — recipe control, portion control, timing, or context — produces the difference. What it does establish is that ρ is not a fixed biological constant: it tracks how controlled the meal regime is. Two consequences follow. First, the breakfast ceiling cannot be transplanted to other meal types, and we do not do so. Second, **CGMacros contains no duplicate lunches at all** (n = 0), so any within-cohort reliability claim is necessarily breakfast-centric by protocol. We flag this gap rather than paper over it.

3.4 The baseline-glucose ablation

AUC and iAUC differ in one structural respect: AUC contains the fasting baseline, iAUC subtracts it. If the model's AUC skill were mostly meal prediction, removing pre-prandial glucose should hurt both targets similarly. Table 3 shows that it does not.

Table 3. LOPO Pearson r with and without meal-time baseline glucose.

Target	Subset	Full (19)	No baseline (18)	Baseline only (1)
iAUC	All meals	0.451	**0.457**	0.204
iAUC	Breakfast	0.655	**0.657**	0.542
AUC	All meals	0.838	0.638	**0.807**
AUC	Breakfast	0.890	0.738	**0.861**
Peak rise	All meals	0.542	0.545	0.231
Peak rise	Breakfast	0.632	0.647	0.547

From `research/experiments/ceiling_results.csv`*, experiment E5/P2.*

For iAUC, removing baseline glucose changes nothing: 0.451 → 0.457 across all meals and 0.655 → 0.657 at breakfast. iAUC skill is therefore not a baseline artefact, which rules out the circularity that would arise if the target contained its own predictor. For AUC, the same removal costs 0.20 (0.838 → 0.638) across all meals and 0.15 at breakfast (0.890 → 0.738), while a model using **baseline glucose and nothing else** reaches 0.807 and 0.861. In variance terms, the baseline-only model's R^2 is 93% of the full model's at all meals (0.651 / 0.700) and 94% at breakfast (0.742 / 0.787). For iAUC the same ratio is 19% and 68%. This is the AUC identity made empirical: most of the "total area" signal is who is fasting high, not what was eaten.

Two qualifications keep the claim honest. First, baseline-only iAUC is not zero — 0.542 at breakfast — because fasting glucose also indexes how reactive a person is; the invariance under removal, not the level of the baseline-only model, is what excludes the mechanical loop. Second, the no-baseline model still contains subject-level variables such as fasting glucose and HbA1c. The ablation removes the *meal-time* pre-prandial measurement, not all glycemic state information. This is the right comparison for the ceiling argument, because it isolates the model that uses only stable, person- and meal-level information; that model sits at 0.657 at breakfast, still below the 0.799–0.810 reliability estimates. The ceiling conclusion therefore does not depend on same-day baseline glucose.

3.5 Who responds more versus which meal is worse

Aggregating out-of-fold predictions to one value per subject answers a different question: not "which of this person's meals will spike them?" but "who tends to spike?". Person-mean aggregation raises every correlation — iAUC 0.451 → 0.501, AUC 0.838 → 0.921, peak rise 0.542 → 0.697 (n = 45) — because averaging removes within-person measurement noise and because between-person differences are large.

The most informative number is what baseline glucose achieves at this level: **person-mean AUC r = 0.963** from baseline glucose alone. Knowing a person's pre-prandial glucose is almost sufficient to rank their average total exposure. For person-mean iAUC the picture is weaker and more honest: baseline-only reaches 0.712 across all meals and 0.761 at breakfast, while the full model reaches 0.501 and 0.693. At n = 45 these differences are small, so we read them as "the 18 non-baseline features add no detectable person-level value" rather than as evidence that meal features are harmful.

The aggregation result is therefore double-edged. It shows that person-level ranking is cheap and robust, and it shows that a higher correlation after aggregation is not evidence of better meal-level prediction — it is a change of question. Section 4.4 tests whether the same trick rescues cross-cohort iAUC; it does not.

4. Cross-cohort stress test (BIG IDEAs, 16 subjects / 656 meals)

4.1 Design of the stress test

The within-cohort analysis fixes a protocol and estimates its ceiling. The cross-cohort test asks what happens when the protocol changes — the comparison the field makes whenever it sets one cohort's correlation against another's. We freeze the CGMacros model (same 80 depth-1 trees, same preprocessing, scaler fitted on CGMacros only) and apply it to BIG IDEAs without any refitting.

Only five features exist in both cohorts: carbohydrate, protein, meal-time baseline glucose, sex and HbA1c. They are complete for all 16 BIG IDEAs subjects, so no imputation is needed. A seven-feature variant adds fat and fibre and covers 13 subjects (558 meals); it is a sensitivity check, not the primary result.

Three differences change together: the device (Libre Pro sampled every 15 min vs Dexcom G6 every 5 min), the annotation (study-recorded meal annotations vs a free-living, self-reported food log grouped within 20 minutes), and the population (CGMacros spans healthy to type-2 diabetes, HbA1c 4.6–8.5%, whereas BIG IDEAs is narrower, per-subject HbA1c 5.3–6.4%). Because the three co-vary, the design cannot attribute the transfer result to any one of them. That is a limitation, and it is also the structure of the comparison the field routinely performs.

Two internal references make the external drop interpretable. Under the same five features and LOPO, CGMacros reaches r = 0.487 for iAUC, 0.816 for AUC and 0.508 for peak rise (n = 1,699), and BIG IDEAs reaches 0.463, 0.625 and 0.483 (n = 656). BIG IDEAs is therefore learnable on its own terms: a poor external number cannot be blamed on unusable target data. These internal rows keep all meals, whereas the headline analysis in §3 applies the official iAUC > 0 filter (1,557 meals), so they are not directly comparable with Table 2; they serve only as like-for-like references for the transfer rows.

Table 4. Frozen-model transfer in both directions (five shared features, Pearson r).

Train → test	Test meals	iAUC	AUC	Peak rise
CGMacros LOPO (internal)	1,699	0.487	0.816	0.508
BIG IDEAs LOPO (internal)	656	0.463	0.625	0.483
CGMacros → BIG IDEAs (frozen)	656	**0.227**	**0.569**	**0.223**
BIG IDEAs → CGMacros (frozen)	1,699	**0.085**	**0.748**	**0.174**
CGMacros → BIG IDEAs, carb-only Ridge	656	0.362	0.288	0.357

From `research/experiments/external_results.csv` (forward transfer and internal references) and `ceiling_results.csv` (reverse transfer, E3). The mean predictor's correlation is undefined because its prediction is constant; its R^2 is negative on every target.

4.2 Forward transfer: AUC survives, the excursion does not

Frozen transfer to BIG IDEAs yields r = 0.227 for iAUC, 0.569 for AUC and 0.223 for peak rise, with coefficients of determination of 0.008, 0.205 and 0.025. On the two excursion targets the model explains about one per cent and two and a half per cent of the variance in a cohort it has never seen.

The AUC row looks like a success, and in a narrow sense it is: 21% of external variance from a model fitted in another study is not nothing. But §3.4 and §3.5 already identify what is being transferred. AUC contains the fasting baseline, baseline glucose is measured in the same units in both cohorts, and a model that ranks people by glycemic level will rank their total areas. The transferable component is baseline compatibility, not meal-response skill. The iAUC and peak rows, which remove the baseline, show what remains of the meal signal: very little.

The carb-only comparison is the uncomfortable result. On external iAUC a single-feature Ridge reaches r = 0.362 against the ensemble's 0.227; on peak rise, 0.357 against 0.223. On AUC the ordering reverses (0.288 vs 0.569). The pattern is consistent with the ensemble having learned CGMacros-specific interactions — the mapping from a macronutrient profile to an excursion in this menu, this population and this device — that do not survive the move, whereas the monotone "more carbohydrate, larger excursion" relationship does. Two qualifications keep this from being over-read. First, the carb-only model ranks better but calibrates worse: its iAUC R^2 is negative (−0.040) and its RMSE is higher (1,876 vs 1,832 mg/dL·min), so the higher r reflects ordering, not usable predictions. Second, one external cohort cannot establish that linear models generalise better; it establishes that the ensemble's within-cohort advantage is not a portable property.

The seven-feature variant changes little and does not alter the ordering: XGBoost reaches 0.272, 0.581 and 0.257 for iAUC, AUC and peak rise (n = 558), against carb-only 0.389, 0.301 and 0.389. The mean predictor's prediction is constant, so its correlation is undefined and its R^2 is negative on all three targets.

4.3 Reverse transfer: iAUC fails in both directions

The forward direction can be explained away by the target cohort: 16 subjects, self-reported meals, a different device. The reverse direction removes that explanation. A model frozen on BIG IDEAs and applied to all 1,699 CGMacros meals reaches r = 0.085 for iAUC (R^2 = −0.139), 0.748 for AUC (R^2 = 0.418) and 0.174 for peak rise (R^2 = −0.043). iAUC now fails on the cohort that was easy in §3 — 45 subjects, 1,699 meals, study-

recorded annotations. The size of the evaluation set is therefore not the explanation, and neither is "BIG IDEAs is unlearnable", since its internal LOPO reaches 0.463.

The failure is symmetric in direction: 0.227 forward, 0.085 backward (Figure 2). What the two models share is a training regime, not a target; each carries the response mapping of its own cohort and loses it on the other. This is the operational meaning of a cross-regime comparison: the number depends on which direction it is run, and there is no direction in which iAUC survives.

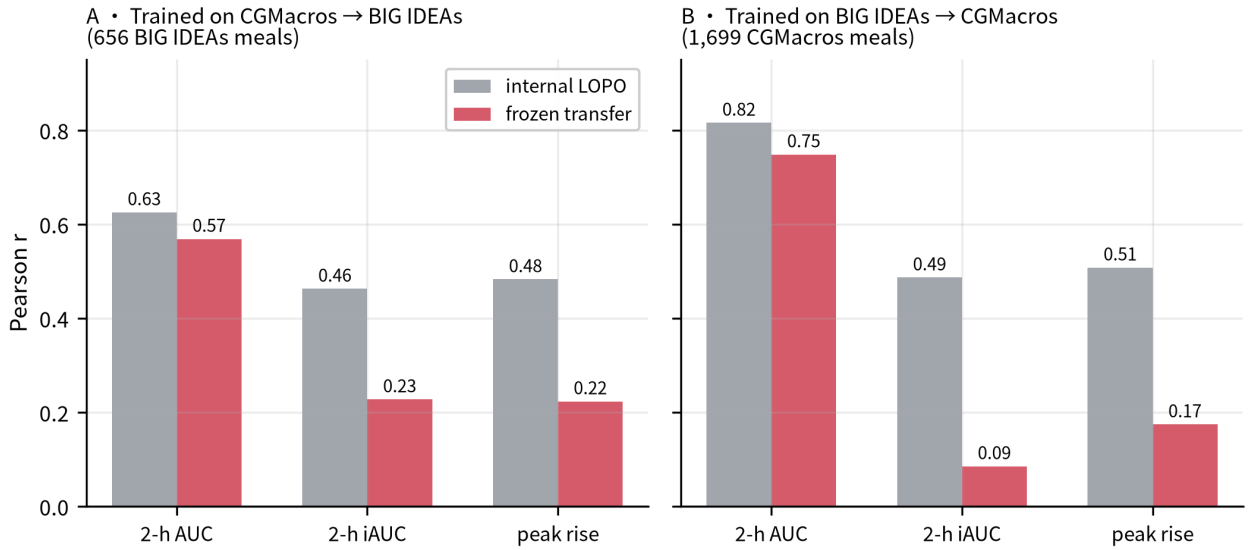


Figure 2. Bidirectional transfer of a frozen model. Grey, LOPO within the cohort being evaluated (the internal reference for that evaluation set); red, the model trained on the other cohort and applied without refitting. Both directions use the five shared features and all meals in the evaluation cohort. AUC survives on both sides of the boundary (0.57 forward, 0.75 reverse, against internal references of 0.63 and 0.82); iAUC (0.23 / 0.09) and peak rise (0.22 / 0.17) do not. *Sources:* `external_results.csv` (panel A) and `ceiling_results.csv`, experiment E3 (panel B).

AUC again transfers, and more strongly in reverse (0.748) than forward (0.569). We resist reading this as evidence that BIG IDEAs teaches a better AUC model. The simpler account is distributional: AUC variance is dominated by between-person differences in glycemic level, CGMacros spans a wider range (HbA1c 4.6–8.5%, including type-2 diabetes) than BIG IDEAs (5.3–6.4%), and ranking 45 subjects across a wide range is an easier task than ranking 16 across a narrow one. Our design does not test this account; we flag it as a plausible contributor, not a finding.

4.4 Person-level aggregation does not rescue cross-cohort iAUC

Within CGMacros, aggregation raised iAUC from 0.451 to 0.501 (§3.5). The natural next question is whether the same operation rescues the cross-cohort number. It does not (Table 5).

Table 5. Meal-level versus person-mean transfer to BIG IDEAs (five shared features).

Target	Meal-level r (n = 656)	Person-mean r (n = 16)	95% CI (person)
iAUC	0.227	**−0.029**	[−0.427, 0.431]
AUC	0.569	**0.715**	[0.353, 0.898]
Peak rise	0.223	0.054	[−0.325, 0.442]

From E4 in `research/experiments/ceiling_results.csv`; the interval is a subject bootstrap (2,000 resamples).

Aggregation lifts AUC by 0.15 — the baseline story once more — and leaves iAUC and peak rise at zero. The iAUC point estimate is in fact slightly negative: the model's person-level ordering is, if anything, inverted relative to the observed ordering. With 16 subjects the confidence interval is wide, so the correct statement is that cross-cohort person-level iAUC skill is indistinguishable from zero, not that it is proven absent. The design cannot resolve moderate correlations at this sample size.

The contrast with §3.5 is the point. Aggregation helped within cohort because averaging removed within-person measurement noise around a signal the model had already learned. Across cohorts there is no shared signal to recover: the gain was a property of the noise structure, not a free lunch that transfers with the model.

4.5 The same-device contrast: what the number needs to be read against

The comparison that makes the argument concrete comes from a study designed before ours. Mendes-Soares et al. (2019) trained a model on an Israeli cohort and applied it to a Midwestern US cohort. Both cohorts wore the same device (Medtronic iPro2, 5-min sampling), logged meals in comparable mobile applications, and had responses computed the same way (2-h iAUC, baseline taken as the median of the 30 minutes before the meal). Under those conditions the cross-cohort correlation was R = 0.596. Their own duplicate-meal reproducibility was R = 0.66 (bagel and cream cheese, n = 111), which they explicitly treat as an upper bound on explainable variance ($R^2 = 0.44$), and they report that the joint-data model captures 88% of that explainable variance (R = 0.618); their within-Israeli cross-validation was R = 0.671 against 0.68 in the original Zeevi cohort. On that accounting the Israeli-only transfer captures $0.596^2/0.66^2 \approx 82\%$ of the explainable variance. Under the stricter $\sqrt{\rho}$ bound this paper adopts — the maximum R^2 for a stable-information model is ρ, not ρ^2 (§2.1) — the same transfer captures $0.596^2/0.66 \approx 0.54$. We use the stricter figure throughout.

Our design changes the device and the annotation at the same time and does not measure the target's reliability at all. That asymmetry is the finding. Where the measurement regime is held fixed, a cross-population transfer can land close to the ceiling; where it is not, the number falls to 0.227 and there is no measured denominator against which to judge it. The best available denominator is the device-level estimate from Hengist et al. (2025): 0.46 on Abbott Libre Pro and 0.45 on Dexcom G4 Platinum, from inpatient duplicate meals — a different device generation, a controlled feeding protocol, and therefore an optimistic stand-in for a free-living, self-reported target. Even against that generous benchmark, the transferred model captures $0.227^2/0.45 \approx 0.11$ of the reliable variance, against 0.54 for the within-cohort breakfast model ($0.655^2/0.799$) and 0.54 for the same-device transfer ($0.596^2/0.66$).

These three fractions are not a like-for-like statistic. Each rests on a different ρ, estimated in a different cohort, and the third uses a proxy rather than a measurement. They are useful only as an ordering: the same model family, read against its own regime's reliability, reaches about half of it within cohort and about half across a population boundary when the device is fixed, but about a tenth when the device and the annotation change together. The cross-regime number 0.227 therefore cannot be read as "this model is worse than R = 0.596". The two numbers have different denominators, and one denominator is

unmeasured. That is precisely why cross-regime comparisons cannot be read as comparisons of model quality.

Two points of comparability deserve statement. Their outcome definition matches ours (2-h iAUC, pre-meal baseline), and their exclusion of very-low-response high-carbohydrate meals (>40 g carbohydrate with PPGR < 5 mg/dL·h) is directionally similar to our iAUC > 0 filter, so the target definitions are close. Their ceiling, however, is measured on a bagel-and-cream-cheese duplicate in the Midwestern cohort, not on the meals used for transfer, and our 0.799 is measured on CGMacros breakfasts; the difference between 0.66 and 0.799 confounds device, population and protocol and must not be read as a device ranking.

What the contrast supports is narrow and, we think, sufficient: transfer performance tracks the measurement regime, and a cross-regime number without a regime-specific reliability estimate is not interpretable as model quality.

5. Uncertainty, decomposition and the mixed model

Section 4 asked how far a frozen model travels. This section asks two different questions about the within-cohort result: how honest are the intervals that would accompany a prediction, and how much of the apparent skill is a between-person rather than a meal-level achievement. Neither analysis refits a model or alters the LOPO protocol; both are descriptive.

5.1 Conformal intervals: marginal coverage is not individual coverage

We attach cross-conformal intervals (Barber et al., 2021) to the existing out-of-fold predictions. For a held-out subject, the calibration set is the absolute residual of all *other* subjects; the interval is $\hat{y} \pm q$, with q the finite-sample conformal quantile. No model is refitted and no point prediction changes.

Marginal coverage lands on target at both nominal levels and in both subsets (Table 6): 0.800–0.804 at 80% and 0.896–0.898 at 90%. That is the guarantee the method offers, and it is a guarantee about the *population of meals*, not about any one person. Exchangeability is violated here by construction — meals within a subject are correlated — so we measure coverage instead of assuming it, and we report the worst individual. Per-subject coverage falls to 0.00 in three of the four breakfast cells (both targets at 80%, AUC at 90%) and to 0.30 in the fourth (90% breakfast iAUC); the all-meal subsets are less extreme but not uniform (minimum 0.18–0.32 at 80%, 0.40–0.50 at 90%). We therefore describe these as *marginal* intervals and make no person-level validity claim.

Widths make the same point in clinical units. At 90%, the all-meal iAUC interval averages 7,909 mg/dL·min — about 3.9× the cohort median iAUC of 2,036 mg/dL·min (n = 1,557) — and the breakfast interval averages 8,885 mg/dL·min against a breakfast median of 2,760. An interval that spans several typical responses is an honest statement of what a meal-composition model can say about one meal.

Table 6. Cross-conformal interval coverage and width.

Subset	Target	Nominal	Empirical coverage	Mean width	Median width	Lowest subject coverage	n
Breakfast	AUC	80%	0.801	6,986	6,992	0.00	423
Breakfast	AUC	90%	0.898	9,808	9,939	0.00	423
Breakfast	iAUC	80%	0.804	6,000	6,071	0.00	423
Breakfast	iAUC	90%	0.896	8,885	8,952	0.30	423
All meals	AUC	80%	0.800	6,119	6,133	0.18	1,557
All meals	AUC	90%	0.898	8,751	8,784	0.40	1,557
All meals	iAUC	80%	0.800	5,357	5,389	0.32	1,557
All meals	iAUC	90%	0.897	7,909	7,978	0.50	1,557

From `research/experiments/uncertainty.csv`. Widths are mean/median interval widths (2q) in mg/dL·min. "Lowest subject coverage" is the minimum, over subjects, of the fraction of that subject's meals falling inside the interval; we report the minimum because it is what the CSV stores, not a full distribution.

5.2 Within- versus between-subject skill

A pooled Pearson r mixes two different questions, and the mixture is not neutral. Subtracting each subject's mean from both outcome and prediction and correlating the centred series asks whether the model ranks *this person's* meals; correlating the 45 subject means asks whether it recovers *who* responds more (Table 7).

Table 7. Skill decomposition of the LOPO XGBoost model.

Subset	Target	Pooled r	Within-subject r	Between-subject r	n meals
All meals	AUC	0.838	0.654	0.921	1,557
All meals	iAUC	0.451	0.430	0.501	1,557
All meals	Peak rise	0.542	0.455	0.697	1,557
Breakfast	AUC	0.890	0.601	0.926	423
Breakfast	iAUC	0.655	0.519	0.693	423
Breakfast	Peak rise	0.632	0.544	0.667	423

From `research/experiments/results.csv` (rows `xgboost (official baseline)`). The between-subject column is the person-mean correlation of §3.5 computed from the same predictions, not an independent estimate.

The ordering is informative. For AUC the between-subject component (0.921) exceeds the within-subject component (0.654) by 0.27: the headline correlation is largely an achievement in ranking people, which is what the baseline-glucose result of §3.4 and the AUC identity of §2.3 predict. For iAUC the two components are nearly equal (0.430 within, 0.501 between), so the pooled 0.451 is not much inflated by between-person variance — iAUC is the target whose correlation comes closest to a genuine meal-level claim. Peak rise sits in between (0.455 within, 0.697 between). Breakfast reproduces the same ordering at a higher level for every target.

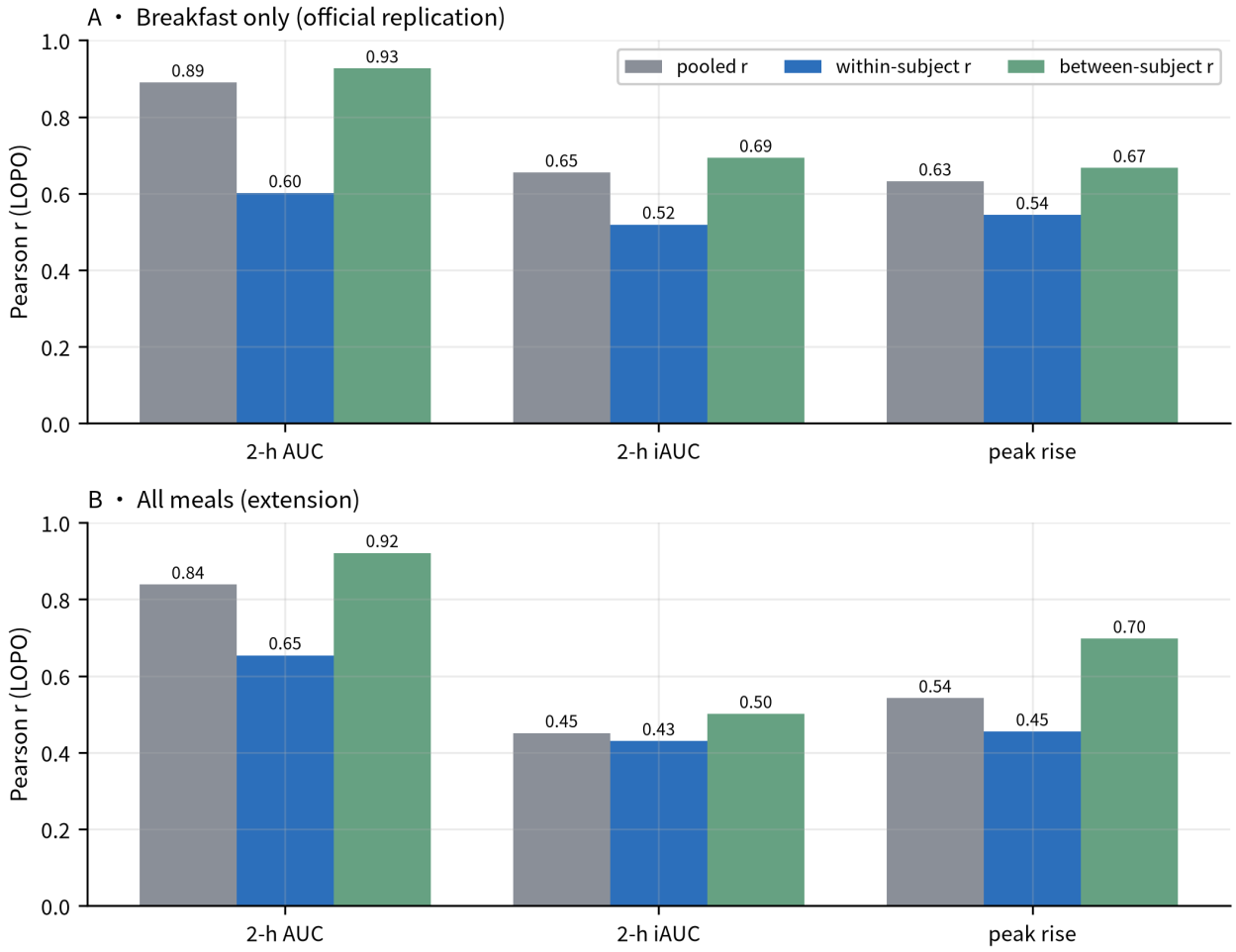


Figure 3. Pooled skill versus within- and between-subject skill. Pearson r for the LOPO XGBoost model computed three ways: on the pooled series, on subject-centred series, and on the 45 subject means. For AUC the between-subject component dominates the pooled figure; for iAUC the two components are nearly equal, so the pooled number is not much inflated by between-person variance. The between-subject bars are the person-mean correlations of §3.5 recomputed from the same predictions, not an independent estimate. *Source:* `results.csv`, rows `xgboost (official baseline)`.

5.3 The mixed model: variance lives where the decomposition says it does

The predictive decomposition has an inferential counterpart. We fit a linear mixed model with a subject random intercept and a Mundlak within-between specification (Mundlak, 1978) on the `iAUC > 0` subset: each standardised predictor enters twice, once centred within subject and once as that subject's mean. This separates "if this person eats more carbohydrate than usual" from "do people who habitually eat more carbohydrate respond differently", while the random intercept absorbs repeated measures. The model is fitted to 1,556 meals (one meal lacks a fibre value and is dropped, not imputed) and is descriptive; it never touches the LOPO protocol.

Table 8. Mixed-model variance components and ICC.

Target	Between-subject SD	Residual SD	ICC	Meals	Subjects
AUC	4,307	2,863	0.694	1,556	45
iAUC	1,536	2,201	0.328	1,556	45
Peak rise	19.4	28.3	0.321	1,556	45

From `research/experiments/mixed_effects_variance.csv`. SDs are in mg/dL·min for AUC and iAUC and mg/dL for peak rise; ICC = σ²_subject / (σ²_subject + σ²_residual).

Two readings follow. First, the ICC ordering matches §5.2: about 69% of AUC variance is between people, against 33% for iAUC and 32% for peak rise. The high AUC correlation is therefore what one expects when most of the outcome variance is a stable person-level quantity that fasting glucose already indexes. Second, the quantities are not the same: §5.2 decomposes *predictive skill*, §5.3 decomposes *outcome variance*. They agree in ordering, and they answer different questions.

The within-person coefficients are the stable part of the model. Holding the person fixed, carbohydrate raises the response for every target — AUC +582 (SE 89), iAUC +606 (SE 68) mg/dL·min, peak +7.9 (SE 0.88) mg/dL per 1 SD — and protein lowers it: −340 (SE 86), −215 (SE 66), −2.9 (SE 0.85). Fat and fibre are indistinguishable from zero within person for all three targets (largest |coefficient| 98 for AUC fat; smallest $p = 0.20$ for iAUC fibre).

The between-person terms are far less precise, because they are identified from 45 subject means: standard errors are 1,600–5,500 for AUC against 80–95 for the within-person terms. Only two reach $p < 0.05$ — protein for iAUC (+3,702, SE 1,800) and for peak rise (+47.9, SE 22.8) — and both carry the opposite sign to the within-person protein effect. We do not interpret them causally. With 45 clusters, habitual protein intake is entangled with overall energy intake and body size, and the contrast is too imprecise to support a between-person claim. The defensible statement is the within-person one: more carbohydrate and less protein at a given meal means a larger excursion, in all three targets, for the same person.

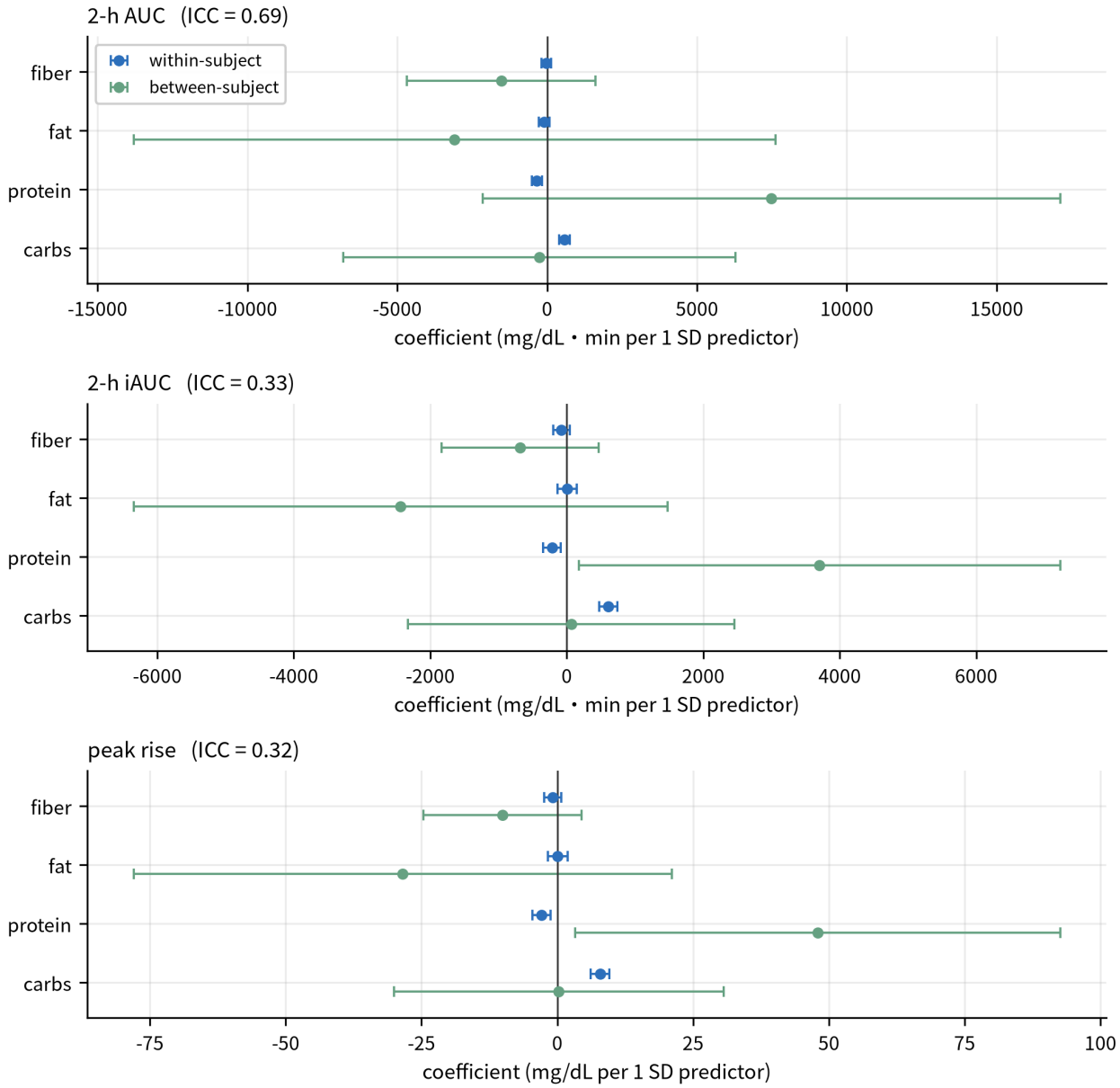


Figure 4. Within- versus between-subject effects in the mixed model. Predictors are standardised; each enters twice, centred within subject (blue) and as that subject's mean (green). Markers are coefficients and bars are ±1.96 SE, from the subject-random-intercept model with a Mundlak specification (1,556 meals, 45 subjects). x-scales differ by target because the coefficients are in mg/dL·min (AUC, iAUC) or mg/dL (peak rise) per 1 SD of the predictor, and the three outcomes differ by orders of magnitude. The between-person terms are identified from 45 subject means, so their intervals are wide and only two exclude zero (§5.3). ICC in each panel title is from Table 8. *Source:* `mixed_effects_coefficients.csv`.

6. Discussion: what follows for the field

The claim of this paper is narrow enough to state in one sentence: a PPGR model's correlation is bounded by the repeatability of the response it predicts, and that repeatability is neither measured nor constant across regimes. Sections 3–5 supply the evidence for the first half; Section 4 supplies the evidence for the second. What follows is an account of what a reader, a builder, and a study designer should do differently, and — more importantly — an account of what the cross-regime numbers do *not* license.

6.1 Three questions to ask of a PPGR correlation

Published PPGR correlations are routinely compared as if they shared a denominator. They do not. Three questions, asked in this order, resolve most of the ambiguity.

First: what is the duplicate-meal reliability under this protocol? Protocol alone moves ρ from 0.799 for a reference breakfast to 0.374 for a free-choice dinner and 0.376 for a snack, with device and population held fixed (§3.3). Across studies the ladder runs from 0.45–0.46 for inpatient duplicate meals on Libre Pro and Dexcom G4 to 0.71–0.77 for four standardized meals on iPro2 (Table 1). The practical consequence is large. A model reporting r = 0.60 is at 89% of the model-relevant bound $\sqrt{\rho} \approx 0.67$ in the low-reliability regime and at 67% of $\sqrt{\rho} \approx 0.894$ in the high-reliability one. The same number describes two different achievements, and without a stated ρ the reader cannot tell which.

Second: does the metric pool between-person variance? A pooled r mixes the question "which of this person's meals will spike them?" with "who tends to spike?". Our decomposition makes the mixture visible: for AUC the between-subject component (0.921) exceeds the within-subject component (0.654) by 0.27, while for iAUC the two are nearly equal (0.430 and 0.501, §5.2). Aggregating to person means raises AUC from 0.838 to 0.921 and iAUC from 0.451 to 0.501 without any change to the model (§3.5). A high pooled correlation on a between-person-dominated target is therefore a statement about ranking people, and should be reported as such.

Third: does the target contain its own predictor? With a 2-h window and a 15-min grid, AUC ≈ 120 × baseline glucose + iAUC (§2.3), so the answer is structural for AUC and not for iAUC. Empirically, a model using pre-prandial glucose and nothing else reaches 0.807 on all-meal AUC and 0.861 at breakfast, and 0.963 on person-mean AUC; removing meal-time baseline costs AUC 0.20 and 0.15 respectively, while leaving iAUC untouched (0.451 → 0.457, §3.4). An AUC correlation that is not accompanied by a baseline-only ablation cannot be distinguished from a fasting-state measurement.

Applied to our own results, these three questions leave one number that we would defend as a meal-level achievement: breakfast iAUC r = 0.655 (and its all-meal counterpart, 0.451). It survives all three filters, and it still sits below both the measured repeatability (0.799) and the model-relevant bound ($\sqrt{0.799} \approx 0.894$).

6.2 What the cross-regime result does and does not establish

Two things are established. First, iAUC transfer fails in both directions: 0.227 forward and 0.085 in reverse, with the reverse failure occurring on the larger, better-annotated cohort, which removes evaluation-set size as an explanation (§4.3). Second, the failure is not an aggregation artefact: person-mean iAUC transfer is −0.029 with a 95% CI of [−0.427, 0.431], so the honest statement is "indistinguishable from zero", not "proven absent" (§4.4).

Reliability is part of the account. The full reasoning is in §4.5, and it deliberately stops short of a decomposition: the only published reliability estimates for the target device family come from an inpatient study on a different device generation, so any arithmetic that uses them as a denominator produces an ordering, not an estimate. That ordering is what matters here: where the measurement regime is held fixed, a cross-population transfer lands near its ceiling; where the device and the annotation change together, the transfer retains a small fraction of the reliable variance against even the most generous available denominator.

What reliability does *not* do is explain the whole gap. Three confounds travel together in our design and cannot be separated with one external cohort: the feature set shrinks from 19 to the 5 shared features; the sensor changes from Libre Pro at 15-min sampling to Dexcom G6 at 5-min; and the annotation changes from study-recorded meals to free-living self-report, whose dates required our own repair procedure. Model class is a fourth: the carb-only Ridge outperforms the ensemble on external iAUC (0.362 vs 0.227) while calibrating worse ($R^2 = -0.040$), which says that the ensemble's within-cohort advantage is not portable, not that linear models generalise better. A reviewer is entitled to ask which of these dominates; with one external cohort we cannot answer, and we do not claim that reliability alone accounts for the collapse.

The AUC rows show what does travel. AUC reaches 0.569 forward and 0.748 in reverse, and the reverse advantage is consistent with distributional breadth rather than with better learning (§4.3). Read together with the ICC of 0.694 for AUC against 0.328 for iAUC (§5.3), the pattern is coherent: what crosses a cohort boundary is glycemic level, which is measured comparably in both cohorts; what does not cross is the mapping from meal composition to excursion. The transferable component and the reliable component are, to a first approximation, the same component — but we state that as an interpretation, not as a tested mechanism.

6.3 What a study would have to report

If the argument is accepted, the minimum reporting standard for a PPGR prediction benchmark changes in five ways.

1. *Measure ρ in the evaluation regime.* Duplicate meals do not require a new study: our breakfast subset already contained 174 same-subject pairs from 44 subjects, and the all-meal subset 291 pairs. The cost is analytical, not experimental.
2. *State the benchmark and the bound.* Report ρ and √ρ for the protocol, and say which one the model is being compared against. A model described as a fraction of "the ceiling" is ambiguous by construction unless the ceiling is named.
3. *Decompose the correlation.* Within- and between-subject components, or an ICC from a mixed model, separate a meal-level claim from a person-level one. Pooled r alone cannot.
4. *Run the transfer in both directions.* A one-directional transfer confounds model portability with target-cohort difficulty. The reverse direction costs one refit and removes the most common alternative explanation.
5. *Report individual coverage, not only marginal coverage.* Our conformal intervals hit nominal coverage marginally (0.800–0.804 at 80%, 0.896–0.898 at 90%) while per-subject coverage falls to 0.00 in three of four breakfast cells, and the 90% all-meal iAUC interval averages 7,909 mg/dL·min against a cohort median response of 2,036 (§5.1). Marginal guarantees are real guarantees; they are not person-level ones, and the distinction should be stated.

The reporting environment into which these recommendations land is not encouraging. An audit of 67 deep-learning glucose-prediction papers found that 23.9% released code, 37.3% used private data, and 55.2% relied on OhioT1DM, a 12-participant cohort (Lu et al., 2026). Those practices make the denominator problem worse, because a reader cannot recompute a reliability estimate for a pipeline that was never released, on data that was never shared.

6.4 What this implies for personalized nutrition

The results draw a sharp line between two products that are often conflated.

Person-level ranking is nearly free. Baseline glucose alone reaches 0.963 on person-mean AUC and 0.712 (all meals) to 0.761 (breakfast) on person-mean iAUC (§3.5). At n = 45 the person-mean iAUC figure for the baseline-only model is not distinguishable from the full model's 0.501 and 0.693, which we read as "the 18 additional features add no detectable person-level value", not as evidence that they are harmful. Anyone who wants to know who responds strongly can get most of the answer from a fasting measurement and a person mean.

Meal-level ranking is bounded by measurement, and the bound is now quantified. The model reaches 0.655 at breakfast iAUC against a repeatability of 0.799, and its honest interval spans several typical responses. A product that promises per-meal guidance is therefore making a claim the measurement regime cannot underwrite at the individual level — not because the models are bad, but because the target itself is only partly reproducible under free-living conditions. The defensible framing is population- or person-level triage, with meal-level predictions presented as distributions rather than point advice.

We state this as a research implication, not as product guidance. The distinction matters because the same arithmetic that limits a product also limits a paper: a benchmark whose target has $\rho = 0.45$ cannot produce a meal-level correlation much above 0.67, however good the model.

6.5 Why the negative result is the result

It would have been easy to publish §3 and stop. The within-cohort numbers are respectable, the replication matches the published baseline (0.655 against ≈ 0.64), and the extension to all meal types is new. The external test was supposed to be a formality, and it failed. We report the failure because it is the part of the study that carries information: a single-cohort benchmark flatters every method that is fitted to it, and the flattering is invisible without a transfer test and a reliability denominator.

Two 2026 papers evaluate PPGR prediction on the same CGMacros cohort and illustrate how easily the two literatures can pass each other. Kondratyeva et al. (2026) use the cohort's 45 adults and 10 free-living days with meal images and microbiome features; Li et al. (2026), in GlucoFM, evaluate 2-h PPGR on 874 matched meal events from 34 participants. Both ask how much accuracy is obtainable. This paper asks what an obtained accuracy means, given that the target is itself only partly reproducible. The questions are complementary rather than competing: a better model raises the numerator, and a reliability estimate fixes the denominator. Neither literature, as published, supplies the second for the regime it evaluates.

The broader point is about falsifiability. The framework of §2 predicts that performance should fall when the target regime's ρ falls, and that the fall need not be proportional. Our data are consistent with that prediction but do not test it cleanly, because we never measured ρ in BIG IDEAs. The test that would settle the question is a cohort with duplicate meals, two devices, and enough subjects to estimate ρ and model performance in the same protocol. Until such a study exists, the defensible conclusion is the conditional one: cross-regime comparisons of PPGR models are confounded by measurement reliability, and

reliability explains part of the observed gap while the residual remains attributable to features, devices, annotation and model class in proportions this design cannot separate.

7. Limitations

This design supports one claim — that comparisons of PPGR models across measurement regimes are confounded by reliability — and it bounds that claim on five sides. Each bound is stated with the number that fixes it, so a reader can judge how much of the argument survives a different protocol.

7.1 Two cohorts, one regime pair, and a narrower glycemic range on one side

The evidence comes from two cohorts: CGMacros (45 subjects, 1,699 meals, Libre Pro at 15 min) and BIG IDEAs (16 subjects, 656 meals, Dexcom G6 at 5 min). We test transfer in both directions — a frozen five-feature CGMacros model scored on BIG IDEAs (§4.2) and a frozen BIG IDEAs model scored on CGMacros (E3, §4.3) — but both directions use the same pair of cohorts. That is one regime contrast, not a sample of regime contrasts, and nothing in the design estimates how much the drop would vary across other device or protocol pairs.

The pair also differs in population, not only in device. CGMacros spans healthy, prediabetic and type-2-diabetic meals (573 / 576 / 408 of the 1,557 analysable meals) with HbA1c 4.6–8.5 %; BIG IDEAs sits at HbA1c 5.3–6.4 % (median 5.65 %). Device and glycemic range therefore move together, and no analysis here can attribute the iAUC loss between them. The AUC asymmetry — 0.569 forward, 0.748 reverse — is consistent with the range difference, but with a single pair it is equally consistent with other explanations, and we do not claim to have identified one.

7.2 Reliability is measured in one cohort and inferred in the other

Every ρ in this paper comes from CGMacros duplicate meals. We do not estimate ρ in BIG IDEAs: that cohort is free-living, and no repeat-meal analysis was attempted there. The account in §6.2 is therefore a consistency argument — the direction and rough size of the drop match what a lower-ρ regime would produce — rather than a mediation analysis. We cannot decompose the observed loss into a reliability term and a residual term with these data, and the title should be read with that limit in view.

A repeat-meal correlation is also not a device-agreement measure. Re-eating the same meal a week later changes the physiology as well as the sensor reading, so ρ as estimated here bundles biological day-to-day variability with measurement error, and the √ρ ceiling inherits both. Two devices worn on the same meal would separate them; this design cannot.

The estimate is thinner than the headline figure suggests, in three ways. First, "duplicate" is a tolerance choice: pairs are same-subject, same-meal-type events whose macros agree within ±2 g, and at ±5 g the pooled value falls from 0.746 to 0.685 (455 pairs; §3.3). Second, the ceiling is unevenly supported across meal types: lunch contributes zero pairs, dinner 40 pairs from 13 subjects (r = 0.374, CI 0.088–0.569) and snacks 77 pairs from 18 subjects (0.376, CI 0.069–0.656), so only breakfast (174 pairs, 44 subjects, 0.799, CI 0.702–0.867) and the same-type pooled value (291 pairs, 0.746, CI 0.657–0.807) rest on most of the cohort. Third, with 45 subjects the breakfast interval spans 0.165 in correlation. Any

ratio of model skill to the ceiling — including the percentages in §4.5 — inherits the uncertainty of ρ, which is not the uncertainty of the model correlation; we report point estimates and do not read them as precise fractions.

7.3 Annotation, preprocessing and target definition

Three choices bound what the numbers mean.

CGMacros meals are study-annotated; the BIG IDEAs record is a free-living food log that we group into events with a 20-minute rule. Portion estimation and event boundaries therefore differ between cohorts, so part of the transfer drop is a difference in what was measured, not only in whom.

The BIG IDEAs 1.1.2 release has misaligned food-log dates. We repair them per subject (no shift for 12 subjects, roughly five months for 4) and validate the repair against the study's standardised-breakfast days (median $\Delta \approx +14$ mg/dL), but a validated repair is not ground truth: a different rule could move individual meals across day boundaries and change their targets.

iAUC is coded to positive excursions only: 1,557 of 1,699 meals, with 142 (8.4 %) dropped. Postprandial dips are real and may be predictable; this coding discards them by construction.

Two further properties of the target are worth stating because they affect interpretation rather than validity. Targets are computed on a 120-minute window anchored at the annotated meal time on a 15-minute grid (9 points), and CGM lags blood glucose by roughly 9–10 minutes, so the anchor shifts the curve relative to the meal. AUC also inherits baseline by identity (AUC ≈ 120 × baseline + iAUC), which is why baseline-only models reach 0.807–0.861 and why AUC values are not pure meal-response skill.

In CGMacros we use the Libre channel only, although participants wore a Dexcom G6 Pro simultaneously and the two channels agree only moderately on meal ranking (mean Kendall $\tau \approx 0.43$; Howard et al., 2020). The second channel is an unused resource in this study: it could have supplied a cross-device reliability estimate, and it is the most obvious extension of E2.

7.4 Model class, and where the ceiling argument stops

Every model here is feature-based: 80 depth-1 boosted stumps (`learning_rate = 0.2`, `reg_alpha = 1.0`), ridge with `alpha = 1.0`, and 19 features (14 of them subject-level) or the 5 shared ones. At 45 subjects this is the defensible class, and sequence models, representation learning and meal-image pipelines are not identifiable on this sample — so nothing here speaks to what a larger model would do with the same features.

The ceiling argument needs an explicit boundary. The √ρ bound applies to a model that predicts one noisy realisation from information that does not share that realisation's measurement error. A personal-history model that consumes the same sensor's preceding trace violates that premise: part of what it learns is sensor behaviour shared between predictor and target, so its correlation with a single measured meal is not bounded by √ρ and is not comparable with ours. This cuts both ways. The paper does not claim that personal-history models cannot do better per meal; it claims that their gains require a design that separates shared sensor error from physiology — two devices, or a held-out

device — and that subject-wise splits are mandatory, because with 45 subjects a model allowed to see a subject's own history can also memorise that subject.

Finally, the feature set omits meal timing, the interval since the previous meal, physical activity, sleep and menstrual cycle, all of which influence the response. The implications in §6.4 are therefore implications about ranking under this protocol, not a causal menu.

7.5 Statistical precision, multiplicity and what would change the conclusion

We report 24 rows in `results.csv` and 77 in `ceiling_results.csv` without a multiplicity correction. The argument rests on contrasts that are large relative to their intervals — breakfast iAUC 0.655 against a ceiling of 0.799, the AUC split of 0.921 between subjects against 0.654 within, the iAUC transfer of 0.227 against within-cohort 0.655 — and the smaller contrasts should be read as exploratory. The cluster bootstrap respects the fact that one subject contributes several pairs, but 44 independent units is itself a small sample; the subject-balanced sensitivity check (one pair per subject, breakfast 0.803, CI 0.684–0.888) sits close to the primary estimate, so the pairing structure is not driving the result.

The transfer evidence is asymmetric in precision and in target. On the BIG IDEAs side, 16 subjects give a person-mean iAUC correlation of −0.03 with a CI of −0.43 to 0.43, which cannot distinguish "no transfer" from "moderate transfer"; only the meal-level estimate (0.227) is precise enough to show failure. In the reverse direction, iAUC 0.085 (R^2 = −0.139) shows failure while AUC 0.748 (R^2 = 0.418) shows partial survival, so "transfer fails" is target-specific rather than global. Mixed-model ICCs (0.694 / 0.328 / 0.321) are point estimates without confidence intervals, and the model assumes Gaussian subject random intercepts.

The conformal intervals carry a further limit. Marginal coverage matches nominal (0.800 / 0.898), but per-subject coverage reaches 0.00 in three of the four breakfast cells, and the all-meal iAUC 90 % interval has a mean width of 7,909 against a median iAUC of 2,036. These are population-level guarantees, not per-person ones, and the intervals are too wide to act on for an individual meal.

What would change the conclusion is stated in §6.5. We add one requirement there: the cohort that settles it should also cover a glycemic range wider than BIG IDEAs', and should evaluate a personal-history model under subject-wise splits with a device-separated target. Until such a study exists, the claim stays conditional, and the observational status of the data means none of the associations above is causal.

8. Conclusion

The question here was not how accurately postprandial glucose can be predicted, but what such an accuracy means. Reproducing the CGMacros baseline settled the first half: breakfast iAUC reaches r = 0.655 against the published ≈ 0.64, and the same features and model extended to all 1,557 analysable meals give 0.451 for iAUC and 0.838 for AUC. The rest of the paper is about the second half.

Within cohort, the breakfast model lands below its own measurement ceiling (0.655 against ρ = 0.799), capturing 0.54 of the reliable variance. Across cohorts, with device and annotation changing at once, the same five-feature model reaches 0.227 for iAUC — 0.11 of the best available denominator — and the reverse transfer fails as well (0.085). Where

device and annotation are held fixed, as in Mendes-Soares et al. (2019), the corresponding fraction is also 0.54. These are not a ranking of models. They are the same model family read against three different denominators, one of which is unmeasured.

Which target is predicted matters as much as which model predicts it. Within CGMacros a baseline-glucose-only model already recovers 0.807 of AUC across all meals and 0.861 at breakfast, and AUC is the target that survives transfer (0.569 forward, 0.748 in reverse, R^2 = 0.418). iAUC, the excursion itself, survives in neither direction. "Transfer works" and "transfer fails" are therefore both true of this study depending on the target, and a reader who reports only one of them is describing a target-definition effect as a model-quality effect.

The claim we defend is the conditional one. Cross-regime comparisons of PPGR models are confounded by measurement reliability; reliability explains part of the observed gap, and the remainder is attributable to features, device, annotation and model class in proportions this design cannot separate. Nothing here is causal, and nothing here proposes a new predictor.

The claim is also falsifiable, which is the point of stating it this way. A cohort with duplicate meals, two devices worn concurrently, a glycemic range wider than BIG IDEAs', and a personal-history model evaluated under subject-wise splits with a device-separated target would settle it. If such a study finds no transfer loss once ρ is matched, the argument narrows to feature-based models; if the loss persists, reliability is not the whole story — a possibility §6.2 and §7.5 leave open rather than exclude.

Three reporting habits would make the literature comparable without new data: name the regime (device, sampling interval, annotation, meal filter), report ρ or a duplicate-meal reliability estimate beside every r, and report per-subject coverage beside marginal coverage. None is expensive; each turns an uninterpretable number into an interpretable one.

Performance claims in PPGR prediction are claims about a measurement regime. Once the regime is stated, the published numbers stop conflicting, and the field's bottleneck becomes visible: not model capacity alone, but measurement reliability and protocol heterogeneity.

Declarations

Ethics

This study is a secondary analysis of two publicly available, de-identified datasets. No participants were recruited, no intervention was performed, and no direct identifiers were accessed. Ethics approval and informed consent for the original collections were obtained by the original study teams. CGMacros participants were recruited at the Sansum Diabetes Research Institute and gave written informed consent under Advarra IRB approval Pro00049227 (ClinicalTrials.gov NCT04991142), and the released files carry a 365-day date shift. BIG IDEAs participants provided written informed consent under the Duke University study protocol, and its released files are date-shifted in the same way. The present analysis used only the public, de-identified files. No additional ethics approval was sought: the study involves no new participants and no identifiable data.

Data availability

Both datasets are public and were obtained from PhysioNet under their own licenses: CGMacros, https://doi.org/10.13026/3z8q-x658 (v1.0.0), CC BY-NC-SA 4.0; and the BIG IDEAs Lab Glycemic Variability and Wearable Device Data, https://doi.org/10.13026/zthx-5212 (v1.1.2), ODC-By 1.0.

No raw or row-level data are redistributed with this manuscript. The two analysis tables (1,699 meals from 45 subjects; 656 meals from 16 subjects) are rebuilt from the public downloads by `research/src/build_dataset.py` and `research/src/build_external.py`, with the raw CGMacros archive fetched and SHA256-verified by `research/src/download_data.sh`. The result tables behind every number in this paper (`results.csv`, `ceiling_results.csv`, `external_results.csv`, `uncertainty.csv`, `mixed_effects_variance.csv`, `mixed_effects_coefficients.csv`) are distributed with the code.

Code availability

The analysis code is at https://github.com/ElijahZhao/MetaNutri---AI-, under `research/`, and `research/src/run_all.sh` reproduces the result tables. Code is MIT-licensed. Tables and figures derived from CGMacros are distributed under CC BY-NC-SA 4.0; material derived from BIG IDEAs carries its ODC-By 1.0 attribution.

Funding

No funding was received for this work. The institutional affiliation is listed only so that the submission is associated with an organization able to provide oversight of the research; that organization had no role in the design, the analysis, or the decision to submit.

Competing interests

The author declares no competing interests.

Author contributions

The author is the sole author, conceived the study, specified the analyses, and is responsible for the interpretation and the final text. Assistance with code execution, figure rendering and drafting is disclosed below.

Use of generative AI

The author designed and directed this study. An AI coding assistant (Trae; large language model) was used as a tool: it repaired and re-ran the analysis code (including correcting the ICC(1,1) estimator in §3.3, after which all experiments were re-run), rebuilt the two analysis tables from the public downloads, wrote the plotting code and rendered the four figures from the committed result files, and wrote the scripts that check each reported value against those files. Draft prose was produced with the assistant and then rewritten and approved by the author, who selected the analyses and the interpretation presented and is solely responsible for the content. No AI system is an author.

References

- Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. (2021). Predictive inference with the jackknife+. *The Annals of Statistics* 49(1):486–507. doi:10.1214/20-AOS1965

- Bent B, Cho PJ, Henriquez M, Wittmann A, Thacker C, Feinglos M, et al. (2021). Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *npj Digital Medicine* 4:89. doi:10.1038/s41746-021-00465-w
- Chen T, Guestrin C. (2016). XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. doi:10.1145/2939672.2939785
- Cho P, Kim J, Bent B, Dunn J. (2023). BIG IDEAs Lab Glycemic Variability and Wearable Device Data (version 1.1.2). *PhysioNet*. doi:10.13026/zthx-5212
- Das A, Kerr D, Glantz N, Bevier W, Santiago R, Gutierrez-Osuna R, Mortazavi BJ. (2025). CGMacros: a pilot scientific dataset for personalized nutrition and diet monitoring. *Scientific Data* 12:1557. doi:10.1038/s41597-025-05851-7
- Hengist A, Ong JA, McNeel K, Guo J, Hall KD. (2025). Imprecision nutrition? Intraindividual variability of glucose responses to duplicate presented meals in adults without diabetes. *The American Journal of Clinical Nutrition* 121(1):74–82. doi:10.1016/j.ajcnut.2024.10.007
- Howard R, Guo J, Hall KD. (2020). Imprecision nutrition? Different simultaneous continuous glucose monitors provide discordant meal rankings for incremental postprandial glucose in subjects without diabetes. *The American Journal of Clinical Nutrition* 112(4):1114–1119. doi:10.1093/ajcn/nqaa198
- Kondratyeva V, Zaripova K, Navab N, Farshad A. (2026). Predicting postprandial glycemic response from meal images, clinical variables, and gut microbiome information. arXiv:2609.24453 [cs.AI].
- Li Z, Natarajan K, Zhang W, Zhou M, Lee SA, Zhang Y, et al. (2026). GlucoFM: a dual-stream foundation model for continuous glucose monitoring. arXiv:2605.30865.
- Lu B, Wagle B, Liang Z, Cui Y, Prioleau T. (2026). Deep learning for blood glucose prediction: reproducibility challenges and factors affecting differential performance. *PLOS Digital Health* 5(9):e0001633. doi:10.1371/journal.pdig.0001633
- Mendes-Soares H, Raveh-Sadka T, Azulay S, Ben-Shlomo Y, Cohen Y, Ofek T, et al. (2019). Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in Midwestern American individuals. *The American Journal of Clinical Nutrition* 110(1):63–75. doi:10.1093/ajcn/nqz028
- Mundlak Y. (1978). On the pooling of time series and cross section data. *Econometrica* 46(1):69–85. doi:10.2307/1913646
- Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163(5):1079–1094. doi:10.1016/j.cell.2015.11.001