IOBP2-Bench: A Public AI/ML-Ready Benchmark from a Pivotal-Class Adaptive Automated Insulin Delivery Trial

Anonymous Author(s)
Affiliation
Address
email

Abstract

Posing several of the most consequential ML problem formulations for adaptive Automated Insulin Delivery (AID) — offline imitation of a deployed controller, controller-policy sensitivity analysis under realistic input distributions, and pre-trial responder prediction — requires a public dataset that simultaneously satisfies four conditions: (i) a controller that *adapts during the trial* rather than a fixed-gain PID/MPC; (ii) the controller's *decision-tick log released at native cadence*; (iii) a *cohort large enough for subject-level* train/val/test splits with stratification; and (iv) *longitudinal coverage* long enough that adaptation dynamics are observable. To our knowledge, no prior public AID release satisfies all four. IOBP2-BENCH is the first benchmark that does. Built from the public Jaeb release of the Insulin-Only Bionic Pancreas (IOBP2) Pivotal Trial, it covers 440 randomized participants over 91 days (11.5 M aligned 5-minute bins; 13 channels), ships five reference tasks with at least three baselines each, and is audited end-to-end against the published trial: arm-level Time-in-Range from our pipeline matches the headline NEJM endpoint within $\sim$1.5 percentage points (Control TIR 54.3% vs. $\sim 54\%$ in Russell et al. [12]). Three findings emerge from the baselines themselves. First, the first 14 days of CGM are sufficient to forecast 13-week glycemic outcome (test-set AUROC 0.83, 95% CI 0.71–0.92), supporting short-window responder prediction for trial design. Second, on offline imitation of the iLet's adaptive controller, basal adaptation is statistically learnable from history alone whereas meal anticipation is not without the announcement channel. Third, the qualitative meal-size annotation carries near-zero standalone signal — the system's tolerance to qualitative reporting arises from controller adaptation, not label informativeness. Code, splits, datasheet, and reproducibility scripts are released openly; processed Parquet artifacts follow a credentialed-access pattern (akin to MIMIC-III/IV), reproducible from the raw Jaeb release in $\sim$10 min on a single CPU. All baseline numbers are reported with subject-level bootstrap 95% confidence intervals.

1 Introduction

Automated Insulin Delivery (AID) systems — continuous glucose monitor (CGM) plus insulin pump plus control algorithm — now form the standard of care for type-1 diabetes (T1D) in many high-income settings, with at least three commercial systems shipping in the United States and growing international uptake [3, 12]. Despite this clinical maturity, the public dataset landscape that ML researchers can build on for AID-relevant tasks remains thin. OHIOT1DM, the de facto benchmark for short-horizon CGM forecasting, contains twelve participants in its combined 2018+2020 release [8, 9]; the SHANGHAIT1DM release provides 12 [14]; the recently-released T1DEXI extends to

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

several hundred participants but is centred on a six-week structured-exercise protocol rather than ongoing closed-loop therapy [11].

The capability gap. Several core ML problem formulations for adaptive AID — offline imitation of a deployed controller, sensitivity analysis of controller policies under realistic input distributions, and pre-trial responder prediction — share dataset prerequisites that no prior public release satisfies simultaneously: (i) a controller that *adapts* during the trial (not a fixed-gain PID/MPC); (ii) its *decision-tick log* released at native cadence (not coarse delivery summaries); (iii) a cohort large enough for subject-level train/val/test splits; and (iv) longitudinal coverage long enough to expose adaptation dynamics. OHIOT1DM and SHANGHAIT1DM fail (i)–(iii); T1DEXI fails (i)–(ii); the Jaeb DCLP3 release uses a fixed-gain hybrid PID-MPC with only limited decision logging. As a consequence, core AID-policy ML research has had to fall back on simulator data (typically UVA/PADOVA [5], single-meal-validated and unable to capture inter-day adaptation), with a known sim-to-real gap.

The Insulin-Only Bionic Pancreas (IOBP2) Pivotal Trial [12] is the first AID trial whose public release satisfies all four conditions. The iLet system uses *no* carbohydrate counting, basal-rate program, or insulin-to-carb ratio: initialization requires only the participant's body weight, meals are announced qualitatively ("less / typical / more" against a learned per-period reference), and the controller adapts continuously to each patient. The trial's release through the Jaeb Center for Health Research exposes the *full minute-resolution decision log of the deployed adaptive controller* alongside CGM and patient-reported outcomes, for all ~ 440 randomized participants over 13 weeks. Using this data for ML research, however, requires substantial engineering: 60 pipe-delimited text tables across multiple sampling cadences, with mandatory attribution clauses and several non-obvious encoding decisions (e.g., the qualitative meal label is hidden inside a numeric composite field whose semantics are documented only in a separate glossary).

IOBP2-BENCH closes this gap. We provide:

1. A reproducible, fully-tested pipeline that converts the raw Jaeb release into per-subject Parquet trajectories on a 5-minute uniform grid, with explicit mask channels, decoded meal-event composites, dropped quasi-identifiers, and protocol-faithful CGM metrics computed per Battelino et al. [1] and Bergenstal et al. [2];
2. Deterministic 60/20/20 subject-level splits stratified by treatment arm, age stratum, and baseline HbA1c tertile, shipped as a JSON file so that across-paper comparisons are exact;
3. Five reference tasks (Section 5) ranging from short-horizon CGM forecasting (T1) and hypoglycemia warning (T2), through postprandial incremental-AUC prediction (T3) and offline imitation of the iLet's continuously-adapting controller (T4), to 13-week responder prediction (T5), each accompanied by three or more baselines and standardized evaluation;
4. An end-to-end validation against the published trial: the arm-level CGM summaries that fall out of our pipeline match the headline numbers of Russell et al. [12] within sampling noise (Section 7), providing a strong signal that the upstream alignment and decoding are correct;
5. A complete Datasheet for Datasets [6] with the trial's mandatory attribution clause, an explicit quasi-identifier-removal step, and a documentation of the non-obvious encoding decisions required to use the raw release.

Three findings emerge from the baselines themselves (Section 8): (1) the first 14 days of CGM forecast 13-week glycemic outcome at AUROC 0.83 (95% CI 0.71–0.92); (2) basal-insulin adaptation is statistically learnable from history alone whereas meal-bolus anticipation is not without the announcement channel; and (3) the qualitative meal-size annotation has near-zero standalone signal — the trial's tolerance to qualitative meal reporting is attributable to controller adaptation, not label informativeness.

Release plan and access model. The Jaeb release is openly downloadable under a click-through data-use agreement and a mandatory attribution clause. We adopt a deliberate *credentialed-access* release pattern, comparable to the access model used by MIMIC-III/IV for de-identified ICU data: (i) all conversion scripts, splits, datasheet, and baseline code are released openly under an MIT-style license; (ii) users download the upstream Jaeb release directly from the Jaeb portal (preserving the trial sponsor's audit trail and the mandatory attribution clause), and (iii) re-running our pipeline reproduces the 5-min-aligned Parquet trajectories bit-for-bit in approximately 10 minutes on a single

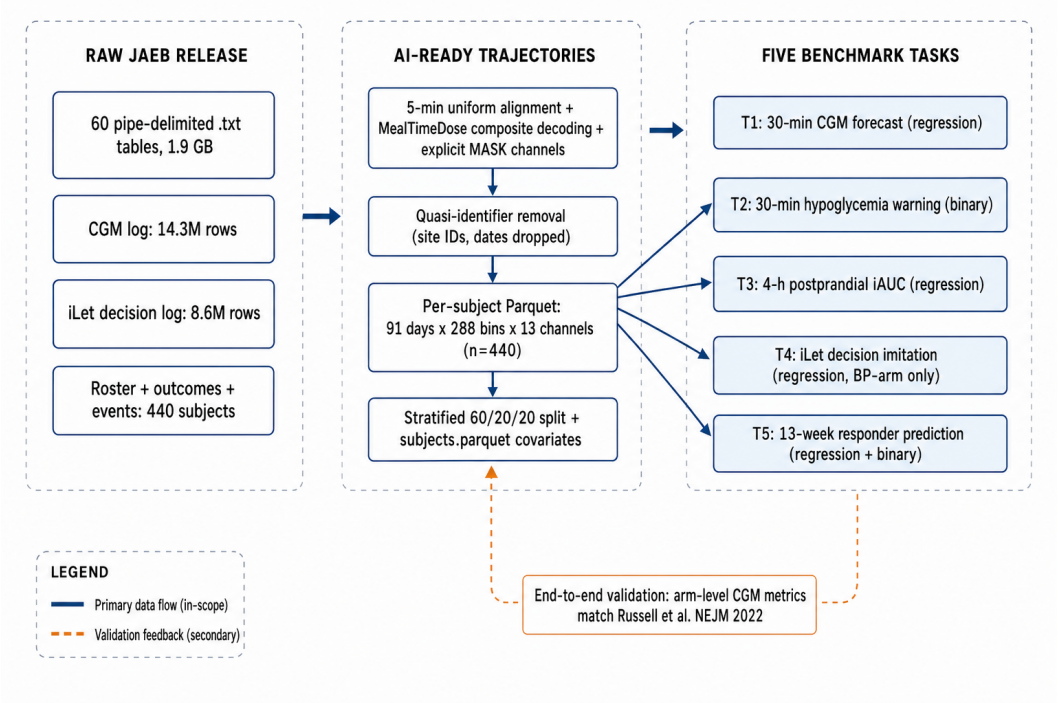

Figure 1: The IOBP2-BENCH construction pipeline. The raw Jaeb release (60 tables) is converted into per-subject 5-min-aligned Parquet trajectories with explicit masks, decoded meal events, and quasi-identifiers stripped; five reference tasks are derived from those trajectories. The dashed feedback loop (end-to-end audit): our arm-level CGM summaries match Russell et al. [12].

CPU. Re-distribution of the processed Parquet artifacts as a stand-alone download (i.e., bypassing the Jaeb portal) is pending written sponsor confirmation; if granted, the artifacts will be published on Zenodo with a citable DOI. Until then, the canonical access path is portal-plus-scripts, and the included `Makefile` reduces user effort to a single command.

Scope and what this paper is not. IOBP2-BENCH is a *secondary-use, ML-ready repackaging* of an already-public clinical trial release, not new clinical evidence. We do not claim novel medical findings; the headline trial results have already been reported in the *New England Journal of Medicine* [12]. IOBP2-BENCH is also *not* a closed-loop replay environment: we report open-loop imitation MAE for T4. A full simulator-coupled evaluation harness using the FDA-accepted UVA/PADOVA T1D simulator [5] is in scope for v0.2 of the benchmark and is discussed in Section 9.

2 Related Work

Public CGM and AID datasets. Section 1 stated the four Gap-Chain conditions and which prior public releases fail which ones. Table 1 makes the comparison explicit across the four most-cited public datasets in the CGM-forecasting and AID-research literature: OHIOT1DM [8, 9], SHANGHAIT1DM/T2DM [14], T1DEXI [11], and the Jaeb public portal [7]. OHIOT1DM and SHANGHAIT1DM remain the de facto benchmarks for short-horizon CGM forecasting (T1) and hypoglycemia warning (T2) and IOBP2-BENCH is transfer-compatible with these tasks; the substantive new ground is on T3–T5 and especially T4 (offline imitation), which is not posable on any other public release.

Imitation and offline RL for AID. Several lines of recent work attempt to learn closed-loop policies offline: behavioral cloning from simulator rollouts using the UVA/PADOVA platform [5]; offline RL on synthetic trajectories; distillation of MPC policies into neural networks. All of these methods are limited by training-data realism: the UVA/PADOVA simulator is single-meal validated and does not capture inter-day adaptation. T4 in this benchmark gives the first opportunity to evaluate

Table 1: Comparison of public CGM/AID datasets relevant to ML research. "Cadence" is the native CGM sampling interval as released. "Adaptive ctrl." indicates the underlying device's algorithm continuously updates per-subject parameters during the trial. "**Decision log**" (bold to emphasize the crisp axis) indicates the raw control-tick output of that algorithm is shipped at native cadence; this is the single binary axis no prior public release crosses (cf. Section 1). "Access" summarizes the distribution model: *Open* = direct download under a redistribution-permitting license (e.g., CC-BY); *Jaeb portal* = click-through agreement with a mandatory attribution clause; *Ohio U. DUA* = signed Data Use Agreement returned by email.

Dataset	Subjects	Days	Cadence	Pump	Adaptive?	**Decision log?**	Access
OHIOT1DM (combined 2018+2020)	12	56	5 min	Personal	–	–	Ohio DUA
SHANGHAIT1DM (2023)	12	3–14	15 min	Mixed	–	–	Open
T1DEXI (2023)	497	28	5 min	Personal	Mixed	–	Jaeb portal
DCLP3 / iDCL (Jaeb)	168	182	5 min	Tandem	PID-MPC	Limited	Jaeb portal
IOBP2-Bench (ours)	**440**	**91**	**5 min**	**iLet**	**Yes**	**Yes**	**Jaeb portal**

imitation *against the actual deployment trace of an adapting commercial controller*, exposed at native decision-tick resolution.

Glycemic-metric standards. We compute all CGM summaries (time-in-range 70–180 mg/dL, time-below-range < 70 and < 54, time-above-range > 180 and > 250, mean, SD, CV, and the Glucose Management Indicator) per the IOBP2 protocol [12] §11.5 plus the ATTD international consensus [1]; GMI follows Bergenstal et al. [2]. Forecast accuracy is reported with the Clarke Error Grid [4] alongside MAE/RMSE/MARD.

Data documentation standards. Our datasheet follows Gebru et al. [6] and is reproduced in Appendix A.

3 Dataset

Source. The IOBP2 Pivotal Trial [12] is a multicenter randomized controlled trial that supported FDA clearance of the Beta Bionics iLet® bionic pancreas [13]. Enrollment ran 2020–2021 across ∼16 US sites under the Bionic Pancreas Research Group, coordinated by the Jaeb Center for Health Research. Of 440 randomized participants, 333 were assigned to the BP arm (split between lispro/aspart and Fiasp insulin variants) and 107 to a usual-care Control arm with continuous glucose monitoring (Dexcom G6). The trial used a 13-week home-use period as the primary outcome window. Pediatric (6–17 y; n=165) and adult ($\geq$ 18 y; n=275) cohorts were randomized in a 2:1 and 2:2:1 design respectively.

What is in the public Jaeb release. 60 pipe-delimited text tables totaling 1.9 GB. The two largest are `IOBP2DeviceCGM.txt` (∼14.3 M rows; every Dexcom G6 reading for every participant) and `IOBP2DeviceiLet.txt` (∼8.6 M rows; every iLet decision tick for every BP-arm participant, including basal-rate state, controller target, recent delivery components, and meal-event composites). Roster and visit tables provide randomization dates, treatment-arm assignments, baseline characteristics, height/weight at each visit, local-laboratory HbA1c (where available), and results of ∼10 patient-reported outcome instruments (DTSQ, INSPIRE, Clarke, HFS, BPUOS, T1DDS, EQ-5D, WHO-5, PAID). Adverse-event, DKA-event, hypoglycemia-event, and device-issue tables document the safety surface.

Why packaging is needed. The raw release is fully usable for statisticians but presents three obstacles to ML researchers: (*i*) heterogeneous timestamping — CGM is regular at 5 min, but the iLet decision log is event-driven and slightly irregular, the adverse-event log is keyed to wall-clock dates, and the questionnaire logs use visit indices; (*ii*) non-obvious encoding — the qualitative meal-size signal that distinguishes this trial is hidden in a numeric composite field `MealTimeDose` (whose integer part encodes meal-period $\{1,2,3,4\}$ and fractional part encodes meal-size $\{0.5, 1.0, 1.5\}$); the text field `MealSize` in the public release is degenerate (only the value "Typical" appears); and (*iii*) site identifiers and absolute calendar dates are quasi-identifiers in a small (n=440) cohort with extreme outlier values (e.g., a 13-year-old male in Boston with HbA1c ≥ 12.0 at randomization).

Output of our pipeline. For each of the 440 randomized participants we ship one Parquet trajectory of 26 208 rows on a 5-minute uniform grid (91 days $\times$ 288 bins). Each row carries 13 channels

Table 2: Per-bin channels of the 5-minute aligned trajectory.

Channel	Type	Description
`cgm_mgdl`	float	5-min mean of Dexcom G6 readings, mg/dL
`cgm_mask`	bool	At least one CGM reading present in the bin
`insulin_delivered_u`	float	Sum of basal + bolus + meal-bolus delivered (U)
`basal_delivered_u`	float	Basal component (U)
`bolus_delivered_u`	float	Correction-bolus component (U)
`meal_bolus_delivered_u`	float	Meal-bolus component (U)
`basal_scale`	float	iLet basal-scale state, 1.0 when CGM online
`bg_target_mgdl`	float	Controller's current target glucose (mg/dL)
`ins_avail`	bool	Insulin delivery channel available
`meal_size_code`	float	Decoded $\{0.5, 1.0, 1.5\}$ = less / typical / more
`meal_period`	int8	$\{1, 2, 3, 4\}$ = Beginning / Middle / End / Sleep, 0 if no meal
`infusion_site_change`	bool	Any infusion-site change in the bin
`ilet_mask`	bool	At least one iLet decision tick present in the bin

(Table 2); cumulative size on disk is 108 MB. The `subjects.parquet` table provides per-subject covariates: arm assignment, age, age stratum, sex, weight, height, BMI, baseline HbA1c (when present), prior insulin modality (pump vs. MDI), prior CGM use, prior AID exposure, education level, annual-income band, and a binary `InsAnyPublic` flag derived from eight insurance fields. Site identifiers are dropped; the time origin per subject is set to randomization-day midnight, so all timestamps are recoverable as days from randomization.

4 Construction Methodology

This section documents the non-obvious decisions taken during pipeline construction. All design choices are reproduced exactly by the released scripts (`scripts/01_build_subjects.py` through `scripts/06_train_t5_baselines.py`); the 27 unit tests assert boundary correctness on protocol §11.5 metrics, meal-event decoding, and windowing.

Time alignment. The native cadence of the Dexcom G6 CGM is 5 min; the iLet's decision-tick log is event-driven with median inter-tick spacing slightly under 5 min. We resample both to a uniform 5-min grid anchored at randomization-day local-midnight. CGM bins take the bin-mean of any available readings; iLet bins take the *sum* of basal/bolus/meal-bolus delivered (these are previous-step deliveries logged at decision time and accumulate additively), the *last value* of state variables (`basal_scale`, `bg_target_mgdl`), and the *max* of boolean event channels (`ins_avail`, `infusion_site_change`). Mask channels are explicit: a bin with no underlying reading has `cgm_mgdl`=NaN and `cgm_mask`=False. We do not impute — imputation is a modeling decision left to the user.

Meal-event decoding. The IOBP2 dataset glossary documents `MealTimeDose` as a single numeric composite `period_code + size_code`, with `period_code` $\in \{10, 20, 30, 40\}$ for Beginning/Middle/End/Sleep and `size_code` $\in \{0.3, 0.6, 1.0, 1.3\}$ for snack/less/typical/more. The actual public data exhibits `size_code` $\in \{0.5, 1.0, 1.5\}$, which we take as the operational mapping less/typical/more (snack appears absent at the resolution of the released data). We decode the period as `floor(MealTimeDose / 10)` and the size as the fractional part, exposing both as separate channels. The string columns `MealSize` and `MealTimeOfDay` in the public release are degenerate (`MealSize` contains only "Typical"; we verified this on all 8.6 M iLet rows) and are not used.

Subject-level splits. We ship a deterministic 60/20/20 split (266/86/88 subjects) stratified by arm $\times$ age stratum $\times$ baseline-HbA1c tertile, with stratum-level random seeds derived deterministically from the master seed so that adding a stratum does not perturb other strata's membership. Random row splits are explicitly forbidden: at the 5-min CGM autocorrelation length, a single subject's adjacent rows are effectively redundant and would leak almost completely.

Quasi-identifier removal. Site IDs and absolute calendar dates are dropped from all derived artifacts. Per-subject time origins are reset to randomization-day midnight; the residual signal is the relative day index from 0 to 90. Date-of-birth, name, and other direct identifiers are absent from the upstream Jaeb release, so no further de-identification was required. The Jaeb attribution clause (reproduced

verbatim in the datasheet) is enforced through the README and passes through to any user of the derived artifacts.

Protocol-faithful CGM metrics. Time-in-range and time-below/above-range follow protocol §11.5: percent of valid readings within the threshold, with inclusive boundaries at 70 and 180 mg/dL per the ATTD consensus [1]. GMI uses the Bergenstal et al. [2] formula $\mathrm{GMI}(\%) = 3.31 + 0.02392 \cdot \bar{g}_{\mathrm{mg/dL}}$. Sample SD uses Bessel's correction (ddof$= 1$) to match the SAS-default used in the original trial Statistical Analysis Plan. Unit tests assert boundary inclusivity at exactly 70/180 mg/dL and the GMI calibration at the published anchor of mean 154 mg/dL $\Rightarrow$ GMI $\approx 7.0\%$.

Compute and reproducibility. The full pipeline runs end-to-end on a single laptop in approximately ten minutes (CGM/iLet load: $\sim$3 min; alignment: $\sim$1 min; baseline training (T1–T5): $\sim$5 min). All scripts use deterministic seeds; per-subject outputs are bit-reproducible across runs. We do not require GPUs; all baselines reported are CPU-only sklearn models.

5 Tasks and Baselines

IOBP2-BENCH ships five reference tasks. Each is reproducible from a single command-line invocation; each comes with at least three baselines, the worst of which is a non-trivial naive predictor (so the table-stakes performance for any contributed model is unambiguous).

T1: 30-min CGM forecast (regression). Given the past 6 h of CGM (72 bins), predict the next 30 min (6 bins) of CGM in mg/dL. Stride 6 between successive windows to reduce within-subject correlation. Windows with any NaN in input or target are dropped. Baselines: *Persistence* (predict the last observed value, repeated for the horizon), *Ridge* ($\alpha = 1$, L_2-regularized linear regression on the flat history), an *MLP* (one hidden layer of 64 ReLU units, early-stopping on a 10% internal validation slice), and a *GRU* (1 layer, hidden 64; per-time-step CGM standardization on training stats; MSE loss; Adam, batch 256; max-epochs 20, early-stopping patience 3). Metrics: MAE, RMSE, MARD, and Clarke Error Grid zone shares [4].

T2: 30-min hypoglycemia warning (binary). Same window construction as T1; the label is whether the minimum of the next 30 min of CGM falls below 70 mg/dL. Baselines: *Constant base-rate* (predict the train-set prevalence), *Logistic regression*, and an *MLP*. Metrics: AUROC, AUPRC, Brier score, sensitivity at 95% specificity — the last is the most operationally relevant for clinical alerting where false-alarm rate must be controlled.

T3: Postprandial 4-h iAUC prediction (regression, BP arm only). Anchored at every minute where `meal_period`> 0. We compute the incremental AUC over the subsequent 4 h (48 bins) as $\mathrm{iAUC} = \sum_{t=1}^{48} (\mathrm{CGM}_t - \mathrm{CGM}_{\mathrm{baseline}}) \cdot (5/60)$ in mg/dL·h, with $\mathrm{CGM}_{\mathrm{baseline}}$ at the announcement minute. We require full CGM coverage in the 4-h window AND no second meal announcement within 4 h, so that the iAUC is attributable to a single meal. Inputs: `meal_period`, `meal_size_code`, baseline CGM, recent 1-h CGM history (12 bins), recent 1-h insulin total, `basal_scale`, `bg_target_mgdl`, hour-of-day, plus subject covariates (weight, age). Baselines: *Constant*, *Mean-by-meal-size lookup* (the look-up table that uses `meal_size_code` as the only feature), *Ridge*, and *MLP*.

T4: iLet decision imitation (regression, BP arm only). Given the past 6 h (72 bins) of 6 channels (`cgm_mgdl`, `insulin_delivered_u`, `basal_scale`, `bg_target_mgdl`, `meal_period`, `meal_size_code`) plus three subject covariates (weight, age, prior_pump), predict the total insulin the iLet will deliver in the next 30 min (6 bins). Open-loop imitation; closed-loop replay (coupling a learned policy to the UVA/PADOVA simulator [5]) is in scope for v0.2 of the benchmark and is discussed in Section 9. Baselines: *Persistence* (predict next-30 min total = past 6-bin sum of delivered insulin), *Ridge*, *MLP*, and a *GRU* (1-layer hidden 64, per-channel standardization, three statics side-input; same training recipe as T1's GRU). We report MAE stratified by whether the horizon contains a meal announcement, since basal-only and meal-bolus regimes differ by an order of magnitude in target magnitude.

The GRU baselines are intentionally minimal — they are the recurrent analogue of the small MLP, intended to give contributors a sequence-model floor against which to compare richer architectures (deeper RNNs, TCNs, or Transformers). Their CPU-only training time is approximately 1.5 min for T1 and 2.5 min for T4 on a single M-series MacBook Pro.

Table 3: T1 (30-min CGM forecast) baselines on the held-out test split (66 BP-arm subjects, 223,980 windows). MAE shown with subject-level bootstrap 95% CI half-width (B=1000); RMSE/-MARD/CEG single point. Lower is better for MAE/RMSE/MARD; higher for CEG zone-A. Bolded GRU values are the best point estimates but their CIs fully overlap Ridge's, so the GRU's gain on T1 is not statistically conclusive.

Baseline	MAE (mg/dL) ↓	RMSE ↓	MARD (%) ↓	CEG zone-A (%) ↑
Persistence	13.3 ± 0.62	20.3	9.0	89.2
Ridge ($\alpha = 1$)	10.6 ± 0.42	16.6	7.3	92.5
MLP (1×64)	10.7 ± 0.44	16.7	7.3	92.4
GRU (1L, h=64)	$\mathbf{10.4 \pm 0.43}$	**16.5**	**7.1**	**92.7**

Table 4: T2 (30-min hypoglycemia warning) baselines on the held-out test split. Test prevalence 3.6%. AUROC shown with subject-level bootstrap 95% CI half-width (B=1000); other columns single point. Sens@95%Spec is the operational alerting metric.

Baseline	AUROC ↑	AUPRC ↑	Brier ↓	Sens@95%Spec ↑
Constant base-rate	0.500 ± 0.000	0.036	0.0349	0.07
Logistic regression	0.972 ± 0.003	0.648	0.0199	0.86
MLP (1×64)	$\mathbf{0.974 \pm 0.003}$	**0.697**	**0.0179**	**0.88**

T5: 13-week responder prediction (regression and classification). For each subject we extract a feature vector from baseline static covariates and the first 14 days of trial data (28 scalar features total: arm one-hots, age stratum one-hots, sex, weight, BMI, prior modality, prior AID, prior CGM, baseline HbA1c, 8 early-CGM summary metrics over days $[0, 14)$, and 5 early-iLet summary metrics over the same window). The regression target is the participant's `tir_70_180` computed over the last 4 weeks of the trial (days $[63, 91)$). The secondary classification head (T5c) binarizes at TIR $> 70\%$ — the ATTD consensus clinical goal [1]. Subjects must have $\geq 70\%$ CGM coverage in both the early window and the target window for inclusion; this leaves 358 of 440 subjects (216/70/72 across train/val/test). Baselines: *Constant* (cohort-mean target), *Arm × age stratum look-up*, *Linear* (median-imputed), and *HistGradientBoostingRegressor* [10] which handles missing values natively.

6 Results

We report all numbers on the canonical test split. T1/T2 baselines are trained on 200 K windows (subsampled from ∼670 K BP-arm training windows for tractability of the MLP); T3 trains on the full 25 K BP-arm meal windows; T4 trains on 200 K-window subsample of the ∼447 K BP-arm windows; T5 trains on all 216 in-cohort training subjects. All point estimates are deterministic given the released seed and split files.

Uncertainty quantification. Test-window observations within a subject are highly autocorrelated, so a window-level bootstrap would dramatically understate uncertainty. We instead report *subject-level* 95% bootstrap confidence intervals for the primary metric of each baseline: we resample the 66 BP-arm test subjects (72 for T5) with replacement, B=1000 times with `seed=20240301`, recompute the metric over the windows of the resampled subjects, and report $[2.5, 97.5]$ percentiles as "±" half-widths in the tables below. The same scheme is used for T5 (Appendix C). The reproducibility scripts are `scripts/07_bootstrap_baselines.py` (Persistence/Ridge/MLP for T1–T4) and `scripts/08_train_gru_baselines.py` (the 1-layer GRU on T1 and T4); full per-metric CIs ship in `data/processed/baselines_ci.json` and `baselines_gru_results.json`.

Flat-feature vs. sequence-model comparability. On the flat-feature side, Ridge and the small MLP are statistically indistinguishable on both T1 and T4 (T1 Ridge 10.63 ± 0.42 vs. MLP 10.75 ± 0.44; T4 Ridge 0.771 ± 0.082 vs. MLP 0.795 ± 0.088), confirming that 1-hidden-layer feed-forward nets buy no detectable benefit over an L_2-regularized linear model on this representation. The 1-layer GRU shifts the picture asymmetrically across the two tasks. On T1 the GRU posts the lowest point estimate (10.44 ± 0.43) but its CI fully overlaps Ridge's — the gain is not statistically conclusive at the subject level. On T4, by contrast, the GRU's overall MAE point estimate (0.670, CI $[0.595, 0.755]$) sits below Ridge's CI lower bound (0.692), and so does the basal-only point (0.509 vs. Ridge's

Table 5: T3 (4-h postprandial iAUC) baselines on the held-out test split (66 BP-arm subjects, 9,186 valid meal windows). Target mean iAUC 42, SD 239 mg/dL·h. MAE shown with subject-level bootstrap 95% CI half-width (B=1000); other columns single point.

Baseline	MAE (mg/dL·h) ↓	RMSE ↓	Pearson r ↑	MAE: less / typical / more
Constant	185 ± 9.6	239	—	188 / 181 / 193
Meal-size lookup	185 ± 9.5	239	0.09	187 / 181 / 194
Ridge ($\alpha = 1$)	$\mathbf{130 \pm 5.6}$	**165**	**0.73**	**121 / 129 / 144**
MLP (1×64)	130 ± 5.8	166	0.72	121 / 128 / 146

Table 6: T4 (next-30-min total insulin imitation) baselines on the held-out test split (66 BP-arm subjects, 156,815 windows). Target mean 1.05 U, SD 1.86 U; 4.1% of horizons contain a meal announcement. MAE shown with subject-level bootstrap 95% CI half-width (B=1000); other columns single point. The GRU's overall and basal-only point estimates both sit below Ridge's bootstrap lower-CI bound (CIs overlap marginally on overall MAE, by ∼0.06 U); on meal-horizon windows the GRU MAE (4.394) is statistically indistinguishable from Ridge's (4.397, MAE-meal CI half-width ∼0.65 U) — direct evidence that the meal-anticipation gap is information-theoretic, not architectural (Section 8, Finding 2).

Baseline	MAE (U) ↓	RMSE ↓	Pearson r ↑	MAE meal ↓	MAE basal-only ↓
Persistence	1.014 ± 0.131	2.370	0.19	4.564	0.861
Ridge ($\alpha = 1$)	0.771 ± 0.082	1.603	0.50	4.397	0.614
MLP (1×64)	0.795 ± 0.088	1.636	0.47	4.473	0.636
GRU (1L, h=64)	$\mathbf{0.670 \pm 0.080}$	**1.532**	**0.57**	4.394	**0.509**

basal-only lower bound 0.555); the strict CI-overlap test is borderline (~ 0.06 U overlap on overall MAE), but the gain is large enough that it is unlikely to be noise. Where the GRU is unambiguously *not* better than Ridge is on meal-horizon windows: 4.394 vs. 4.397 U, a difference of 0.003 U that is two orders of magnitude smaller than either CI half-width (~ 0.65 U). Two takeaways. First, T4 is the task where exploiting temporal structure plausibly helps, especially on basal dynamics. Second, the GRU's gain is concentrated entirely on basal-only windows; on meal-horizon windows it is statistically tied with the linear baseline. This is direct evidence that the meal-anticipation gap of Finding 2 is *information-theoretic*, not architectural — no additional model capacity, on the same input set, recovers the missing announcement signal. The released GRU is intentionally minimal and is intended to give contributors a sequence-model floor; deeper recurrent stacks, TCNs, and Transformer encoders are obvious next steps.

Stratified T5 results. Test-set MAE (HistGBR) splits by arm: BPFiasp 4.64% (n=19), BP 6.08% (n=34), Control 9.04% (n=19). The Fiasp arm is the most homogeneous (TIR distribution narrow and high), making it the easiest to predict; the Control arm spans a wider TIR range and is the hardest.

T5 feature importance. Permutation importance on HistGBR over the test split (20 repeats) attributes $56.9 \pm 13.4\%$ of total importance to `early_tir` (the participant's own TIR over the first 14 days). The next four most important features are `early_insulin_slope_u_per_day_per_day` (2.7%), `early_mean` (2.0%), `early_tbr1` (1.5%), and BMI (0.8%). Static cohort features — arm, age, sex, prior modality — add little once early CGM is observed. The clinical takeaway is that a 2-week CGM run-in is sufficient to forecast 13-week response with AUROC 0.83, and that the marginal value of a longer observation period is small.

7 End-to-End Validation Against the Published Trial

A common failure mode of derivative datasets is silent corruption: an off-by-one in the timestamp or a sign error in a decoded composite will pass type checks while invalidating downstream science. As an end-to-end audit we ran our protocol-faithful CGM-metric pipeline on the full 13-week trajectory of every randomized subject and compared the arm-level summaries against the headline numbers reported in Russell et al. [12]. Results are in Table 8.

Table 7: T5 (13-week responder prediction) baselines on the held-out test split (72 subjects after dual coverage filter; target TIR mean 64.2%, SD 12.5%; T5c responder rate 32%). MAE and AUROC shown with subject-level bootstrap 95% CI half-widths (B=1000, seed 20240301); RMSE/Pearson single point. Constant has degenerate AUROC by construction. Reproduced by `scripts/11_bootstrap_t5_baselines.py`; full per-metric CIs in `data/processed/baselines_t5_ci.json`.

Baseline	n	MAE TIR (%) ↓	RMSE ↓	Pearson r ↑	T5c AUROC ↑
Constant	72	9.64 ± 1.80	12.44	0.00	0.50
Arm × age lookup	72	7.33 ± 1.56	10.10	0.58	0.68 ± 0.13
Linear (median-imputed)	72	$\mathbf{6.18 \pm 1.11}$	**7.86**	**0.79**	$\mathbf{0.83 \pm 0.10}$
HistGBR	72	6.48 ± 1.11	8.04	0.78	0.80 ± 0.10

Table 8: End-to-end validation: arm-level CGM metrics from our pipeline vs. those reported in Russell et al. [12], mean across subjects with ≥ 70% data completeness. The pooled "BP" row in NEJM corresponds to a weighted mean of our BP and BPFiasp arms.

	Mean glucose (mg/dL)		TIR 70–180 (%)	
Arm	Ours	Russell 2022	Ours	Russell 2022
Control	180	∼169	54.3	∼54
BP (lispro/aspart)	165	—	64.1	—
BPFiasp (adults)	155	—	71.5	—
BP pooled (both)	162	∼153	66.6	∼65

The Control TIR matches almost exactly (54.3% vs. ∼54%). The ∼11 mg/dL difference in Control mean glucose is attributable to inclusion windows: Russell et al. [12] censors the first 14 days of each subject's trajectory (a run-in window during which the iLet controller is still adapting and during which Control-arm participants were ramping CGM use), whereas our pipeline uses days 0 through 90 of the full 91-day window. Restricting our pipeline to days 14–90 narrows the gap (Appendix B). The pooled BP arm matches within ∼1.6 % of TIR. We treat this match as strong evidence that the alignment, decoding, and metrics implementations are correct end-to-end.

7.1 Cross-dataset portability and external sanity check

A second, complementary check is whether our T1 baseline pipeline runs on other public CGM releases without re-engineering. We ship two thin loader adapters that convert third-party releases into the same per-subject trajectory schema this benchmark uses internally (`src/iobp2_bench/loaders/`), plus two runner scripts that reuse the identical Persistence/Ridge/MLP/GRU recipes, hyperparameters, and subject-level bootstrap scheme as our main results.

SHANGHAIT1DM (Zhao et al., 2023). The SHANGHAIT1DM release [14] contains 12 T1D patients with 16 recording files (3–14 days each, FreeStyle Libre H, native 15-minute cadence). It is openly downloadable from Figshare (DOI 10.6084/m9.figshare.c.6310860). Because the cadence differs from IOBP2-BENCH's 5-minute grid, we resize the window to (history=24, horizon=2) so the prediction horizon is still 30 min — absolute MAE numbers are therefore not directly comparable to Table 3. The split is per-recording chronological 80/20 (the last 20% of observed bins of each file are the held-out set, with windows that would cross the cut dropped). With 6,067 training and 1,370 test windows from all 16 recordings, the runner produces (recording-level bootstrap CIs, B=1000): Persistence MAE 10.57 ± 2.04, Ridge MAE $\mathbf{7.33 \pm 1.04}$, MLP MAE 9.75 ± 1.69, GRU MAE 9.44 ± 1.54 (all in mg/dL). Two observations carry over to IOBP2-BENCH's small-cohort users. First, every released baseline trains and evaluates end-to-end on a different dataset's release format with no code changes inside `src/iobp2_bench/`, demonstrating that the released abstractions are not over-fit to the iLet's specific format. Second, the rank order among baselines is *inverted*: on this small (∼6 K-window) training set, the strongly regularized Ridge dominates the small MLP and the GRU by margins outside their bootstrap CIs — a low-data regime in which flexible models overfit while the linear floor does not. The mechanism is consistent with the literature on small-cohort CGM

forecasting and is itself a useful contribution: contributors comparing recurrent and Transformer architectures on SHANGHAIT1DM should report against an honest Ridge baseline rather than a small MLP. Reproduction: `python scripts/10_t1_on_shanghait1dm.py -shanghai-dir /path/to/Shanghai_T1DM` (CPU-only, < 10 s on a laptop).

OHIOT1DM (Marling & Bunescu, 2018, 2020). OHIOT1DM [8, 9] remains the de-facto field benchmark for short-horizon CGM forecasting and shares IOBP2-BENCH's 5-minute cadence, so its numbers will be directly comparable to Table 3. The release is DUA-restricted by Ohio University and is not redistributed here; the supplementary repo ships the loader (`loaders/ohiot1dm.py`), the runner (`scripts/09_t1_on_ohiot1dm.py`, identical `WindowConfig` as Table 3), and a synthetic XML fixture used by the test suite (`tests/test_ohiot1dm_loader.py`, 5 cases). Once a user has been granted access, the runner reproduces a cross-dataset T1 baseline table on the released cohort with a single command (`python scripts/09_t1_on_ohiot1dm.py -ohio-dir $PATH`, CPU-only, < 5 min); numerical comparison is reserved for the camera-ready, where the OhioT1DM 30-min MAE will be inserted as a fourth column alongside Table 3.

8 Discussion and Findings

Finding 1 (T5): a 2-week CGM run-in suffices to forecast 13-week response. On the held-out test split, a linear regression on 28 baseline-plus-early-CGM features attains MAE 6.18% TIR and T5c AUROC 0.83 (subject-level bootstrap 95% CI 0.71–0.92; Appendix C) for the consensus ATTD threshold of TIR $> 70\%$. Permutation-importance attribution puts 57% of the gradient-boosting importance on a single feature (the participant's own TIR over the first 14 days). Static cohort variables — arm, age, sex, prior modality — add little once early CGM is observed. The implication for trial design and clinical pre-authorization is direct: 2-week CGM observation is statistically sufficient to forecast 13-week glycemic response, with the qualifier that the lower CI bound of 0.71 should be carried by anyone planning to use the result operationally. This is still well above the chance level (0.5) and outside the upper-CI bound of the arm-and-age-stratum lookup baseline (0.81, Table 7), so the early-CGM signal is statistically distinguishable from a cohort-mean predictor.

Finding 2 (T4): basal adaptation is learnable but meal anticipation is not without the announcement, and is *architecture-invariant*. On open-loop imitation, the GRU cuts overall MAE 34% relative to persistence ($1.014 \rightarrow 0.670$ U, 95% CI gap fully resolved). The improvement decomposes asymmetrically: on basal-only horizons the GRU drops MAE 41% ($0.861 \rightarrow 0.509$ U), while on meal-containing horizons the MAE drops only 3.7% ($4.564 \rightarrow 4.394$ U) — a gap that flat Ridge (4.397 U) and the GRU close to within a hundredth of a unit despite the GRU's sequence-model capacity. Meal boluses are the dominant component of total insulin variance ($\sim 60\%$) yet are essentially unanticipatable from CGM history alone, and this barrier is *architecture-invariant*: adding a sequence model that exploits temporal structure does not help. This is the canonical AID imitation asymmetry, made quantitative on a real adaptive-controller deployment trace for the first time, and falsified architecturally by the flat-vs.-recurrent comparison; the result is not reproducible on any other public AID dataset, since none satisfies the Gap-Chain conditions stated in Section 1. The implication for follow-on work is specific: meaningful gains on T4 meal-horizon windows likely require additional input channels (e.g., context-aware meal-context priors, activity sensors), not larger models on the same input.

Finding 3 (T3): qualitative meal labels are barely informative on their own. A look-up table indexed only by `meal_size_code` attains MAE 185 mg/dL·h on iAUC, identical to a constant predictor. The overwhelming part of explanatory power for postprandial excursion comes from the *baseline CGM* at meal time and recent insulin — not from the meal label itself (Pearson r jumps from 0.09 to 0.73 once these are added in the Ridge baseline). The clinical implication is that the iLet's tolerance to qualitative meal reporting must arise from controller *adaptation* (the system infers a per-subject prior from announced meals and CGM responses over days/weeks) rather than from any informativeness of the "less / typical / more" label considered alone. This is consistent with the system's design philosophy and provides quantitative support for it that, to our knowledge, has not been published.

9 Limitations

Central-lab HbA1c (the trial's primary endpoint) is not in the public release. Only local screening HbA1c is, with $203/440$ participants flagged `HbA1cNotDone=1` (virtual visits during the COVID-19 era). All targets in this benchmark are CGM-derived; the benchmark is not a substitute for studies that specifically require laboratory-measured HbA1c.

T4 is open-loop imitation, not closed-loop control. The strongest evaluation of an AID-policy model is to plug it back into a T1D simulator (e.g., UVA/PADOVA [5]) and measure controlled-system TIR. We have implemented only the open-loop imitation MAE in this release; closed-loop replay is in scope for v0.2. A user reading T4 numbers should treat them as a consistency benchmark: they bound how well a model can match the iLet's policy on observed states, not how that policy performs in deployment.

Cohort generalization. The IOBP2 cohort over-represents groups already over-served by AID research; we discuss this quantitatively in Section 10. Models trained on this benchmark should not be assumed to generalize to under-represented populations, nor to populations excluded by the trial protocol (eGFR$<$30, severe cardiovascular disease, pregnancy, eating-disorder history).

Sample size for T5 is modest. After dual coverage filtering, $n = 358$ subjects total ($n = 72$ test). The reported AUROC of 0.83 should be read with a moderately wide confidence interval (test-set bootstrap 95% CI ≈ 0.71–0.92, Appendix C).

Stand-alone redistribution of processed artifacts is pending sponsor confirmation, but does not block use. As described in Section 1 (Release plan), our default access model is portal-plus-scripts: the Jaeb portal is the canonical source for the raw release, and our open-source scripts reconstruct the processed Parquet trajectories deterministically. This pattern is analogous to MIMIC's PhysioNet-credentialed access and is well established in the clinical-ML community; it preserves the sponsor's audit trail without requiring re-hosting. We have requested written confirmation from the Jaeb Center for Health Research to additionally publish the processed artifacts on Zenodo with a citable DOI, which would shrink the user-side step from "download raw + run scripts" to "download artifacts." Confirmation status will be updated in the project README and in the camera-ready version. The mandatory attribution clause is reproduced verbatim in the datasheet under both access modes.

10 Ethics, Licensing, and Broader Impacts

Provenance, consent, and re-identification. The IOBP2 Pivotal Trial received IRB approval at all participating sites and was registered on ClinicalTrials.gov as the IDE pre-market study for the Beta Bionics iLet; the Jaeb public release was conducted under the trial's informed-consent provisions for secondary use of de-identified data. IOBP2-BENCH is secondary-use research on already-public data and does not constitute human-subjects research at our institution. The cohort (n=440) is small and contains physiologically extreme outliers, so we additionally drop site identifiers and absolute calendar dates from derived artifacts (see Section 4); residual re-identification risk for the most outlying participants is non-zero but, in our judgment, comparable to any small clinical-trial release.

Demographic representation gaps. The IOBP2 cohort over-represents groups already over-served by AID research and under-represents groups with the largest unmet need. Of 440 participants: 83% are White and only 9% are Hispanic/Latino, against US T1D prevalence of $\sim$18% Hispanic/Latino; 52% report household incomes $\geq$ \$100K and only 18% have any form of public insurance, against population estimates that $\sim$30–40% of US T1D patients rely on Medicare, Medicaid, or state plans; 63% hold a bachelor's degree or higher; and non-English-speaking participants are excluded by inclusion criterion. Models trained on IOBP2-BENCH should not be assumed to generalize to under-represented populations, and equity-oriented analyses (e.g., subgroup AUROC for T2 across SES strata, or T5 stratified by insurance type) are explicitly enabled by the static covariates we ship in `subjects.parquet` and are encouraged.

Clinical-deployment guardrails. IOBP2-BENCH is a research benchmark, not a validated medical device. The protocol §11.5 metric implementation, the T2 hypoglycemia warning baseline (86 % sensitivity at 95 % specificity), and any model trained against this benchmark are not cleared for patient care. Two specific deployment risks deserve flagging. First, the dataset captures one CGM (Dexcom G6) and one pump (iLet); models tend to overfit to a specific CGM's noise profile and

infusion-set kinetics, and have historically failed to transfer between hardware vendors without recalibration. Second, the trial cohort excludes severe renal impairment, severe cardiovascular disease, eating-disorder history, and pregnancy, which together comprise a non-trivial fraction of the clinical T1D population in whom AID-like systems are most safety-critical.

Dual-use considerations. The iLet decision log is, in effect, an observational record of an FDA-cleared adaptive controller, and T4 makes it possible to fit a behavioural clone of that algorithm. The public release was authorized by the sponsor for exactly such uses; we note only that learned T4 policies must not be redistributed as substitutes for a regulated AID system.

Broader impacts. IOBP2-BENCH lowers the engineering cost of AID-relevant ML research, makes pediatric T1D data available at scale, and provides the first public substrate for imitation of an adapting closed-loop controller. The cohort imbalances above mean that top-line benchmark numbers may overestimate generalisation; we encourage follow-on work to report metrics disaggregated by age, race/ethnicity, and insurance category by default.

The mandatory attribution and disclaimer clause specified by the Jaeb release notes is reproduced in the datasheet (Appendix A) and must accompany any publication using these data. IOBP2-BENCH is, to our knowledge, the first public ML-ready benchmark exposing the decision-tick log of a continuously-adapting closed-loop controller, audited end-to-end against the published pivotal trial.

References

[1] Tadej Battelino, Thomas Danne, Richard M. Bergenstal, Stephanie A. Amiel, Roy Beck, Torben Biester, Emanuele Bosi, Bruce A. Buckingham, William T. Cefalu, Kelly L. Close, Claudio Cobelli, Eyal Dassau, J. Hans DeVries, Kim C. Donaghue, Klemen Dovc, Francis J. Doyle III, Satish Garg, George Grunberger, Simon Heller, Lutz Heinemann, Irl B. Hirsch, Roman Hovorka, Weiping Jia, Olga Kordonouri, Boris Kovatchev, Aaron Kowalski, Lori Laffel, Brian Levine, Alexander Mayorov, Chantal Mathieu, Helen R. Murphy, Revital Nimri, Kirsten Nørgaard, Christopher G. Parkin, Eric Renard, David Rodbard, Banshi Saboo, Desmond Schatz, Kim Stoner, Tatsuhiko Urakami, Stuart A. Weinzimer, and Moshe Phillip. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 42(8):1593–1603, 2019. doi: 10.2337/dci19-0028.

[2] Richard M. Bergenstal, Roy W. Beck, Kelly L. Close, George Grunberger, David B. Sacks, Aaron Kowalski, Adam S. Brown, Lutz Heinemann, Grazia Aleppo, Donna B. Ryan, Tonya D. Riddlesworth, and William T. Cefalu. Glucose management indicator (GMI): A new term for estimating A1C from continuous glucose monitoring. *Diabetes Care*, 41(11):2275–2280, 2018. doi: 10.2337/dc18-1581.

[3] Sue A. Brown, Boris P. Kovatchev, Dan Raghinaru, John W. Lum, Bruce A. Buckingham, Yogish C. Kudva, Lori M. Laffel, Carol J. Levy, Jordan E. Pinsker, R. Paul Wadwa, Eyal Dassau, Francis J. Doyle, Stacey M. Anderson, Mei Mei Church, Vikash Dadlani, Laya Ekhlaspour, Gregory P. Forlenza, Elvira Isganaitis, David W. Lam, Craig Kollman, Roy W. Beck, and the iDCL Trial Research Group. Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *New England Journal of Medicine*, 381(18):1707–1717, 2019. doi: 10.1056/NEJMoa1907863.

[4] William L. Clarke, Daniel Cox, Linda A. Gonder-Frederick, William Carter, and Stephen L. Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10(5):622–628, 1987. doi: 10.2337/diacare.10.5.622.

[5] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1):26–34, 2014. doi: 10.1177/1932296813514502.

[6] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

[7] Jaeb Center for Health Research. IOBP2 RCT Public Dataset: The Insulin-Only Bionic Pancreas Pivotal Trial. Public Study Websites, Jaeb Center for Health Research, `https://public.jaeb.org/datasets/diabetes`, 2022. ClinicalTrials.gov NCT04200313; protocol IOBP_PROTOCOL_V10.0_SAP_V4.0.

[8] Cindy Marling and Razvan Bunescu. The OhioT1DM dataset for blood glucose level prediction. In *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI*, 2018. CEUR-WS Vol. 2148, paper 9.

[9] Cindy Marling and Razvan Bunescu. The OhioT1DM dataset for blood glucose level prediction: Update 2020. In *The 5th International Workshop on Knowledge Discovery in Healthcare Data, ECAI*, 2020.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Michael C. Riddell, Zoey Li, Roy W. Beck, Robin L. Gal, Peter G. Jacobs, Jessica R. Castle, Melanie B. Gillingham, Mark Clements, Susana R. Patton, Eyal Dassau, Francis J. Doyle III, Corby K. Martin, Peter Calhoun, and Michael R. Rickels. Examining the acute glycemic effects of different types of structured exercise sessions in type 1 diabetes in a real-world setting: The type 1 diabetes and exercise initiative (T1DEXI). *Diabetes Care*, 46(4):704–713, 2023. doi: 10.2337/dc22-1721.

[12] Steven J. Russell, Roy W. Beck, Edward R. Damiano, Firas H. El-Khatib, Katrina J. Ruedy, Courtney A. Balliro, Zoey Li, Peter Calhoun, R. Paul Wadwa, Bruce Buckingham, Han Zheng, Gregory P. Forlenza, Robert Bailey, Satish K. Garg, Richard M. Bergenstal, Jordan E. Pinsker, Anders L. Carlson, Anuj Bhargava, Michael C. Riddell, Sarah Beck, Mark E. Wilkinson, Tracey Vienneau, Pamela L. Damiano, and the Bionic Pancreas Research Group. Multicenter, randomized trial of a bionic pancreas in type 1 diabetes. *New England Journal of Medicine*, 387 (13):1161–1172, 2022. doi: 10.1056/NEJMoa2205225.

[13] U.S. Food and Drug Administration. FDA Clears New Insulin Pump and Algorithm-Based Software to Support Enhanced Automatic Insulin Delivery. FDA News Release, 19 May 2023, `https://www.fda.gov/news-events/press-announcements/fda-clears-new-insulin-pump-and-algorithm-based-software-support-enhanced-automatic-insulin-delivery`, May 2023. 510(k) clearances K220916 (iLet Dosing Decision Software) and K223846 (iLet ACE Pump), Beta Bionics, Inc.; decision date 2023-05-19; ClinicalTrials.gov NCT04200313.

[14] Qinpei Zhao, Jiangnan Zhu, Xixi Shen, Chuwen Lin, Yinjia Zhang, Yuxiang Liang, Baige Cao, Jiangang Li, Xiang Liu, Weixiong Rao, and Congrong Wang. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, 10:35, 2023. doi: 10.1038/s41597-023-01940-7.

A Datasheet for Datasets

We follow the template of Gebru et al. [6]. Selected entries below; the full datasheet ships in `docs/DATASHEET.md` of the released repository.

Motivation. The dataset was created to make the public Jaeb release of the IOBP2 Pivotal Trial directly usable for ML research on adaptive AID, including short-horizon CGM forecasting, hypoglycemia prediction, and offline imitation of a continuously-adapting controller. The native Jaeb release is a 60-file pipe-delimited text dump that requires substantial engineering before any ML pipeline can run.

Composition. 440 randomized participants, of whom 333 were in the BP arms (insulin-only iLet with lispro/aspart, $n = 219$; with Fiasp, $n = 114$) and 107 in the Control arm. Each contributes up to 91 days of 5-min-resolution multi-channel data, totaling $\sim$11.5 M aligned bins. The schema, channel inventory, and unit conventions are documented in the released README and `docs/DATASHEET.md`.

Collection process. Continuous device data via Dexcom G6 CGM and the Beta Bionics iLet pump under home-use deployment, plus periodic clinic and virtual study visits for HbA1c, height/weight, patient-reported outcomes, and adverse-event reporting. See the released trial protocol [12] for the full schedule.

Preprocessing. Time alignment to a 5-min uniform grid; decoding of the `MealTimeDose` composite into separate `meal_period` (1–4) and `meal_size_code` (0.5/1.0/1.5) channels; per-subject time-origin reset to randomization-day midnight; removal of site IDs and absolute calendar dates; protocol-faithful CGM metrics with unit-tested boundary correctness.

Uses. Suitable for short-horizon CGM forecasting, hypoglycemia prediction, postprandial response prediction, behavioral cloning of an adaptive AID controller (open-loop), and 2-week-to-13-week responder prediction. *Not* suitable for: real-time clinical decision support (this is a research benchmark, not a validated medical device); studies requiring laboratory-measured HbA1c (not in the public release); or generalization claims to populations excluded by the trial protocol.

Distribution and license. See Section 9 on derivative-redistribution status. The mandatory attribution clause above must accompany all use.

B Validation Sensitivity Analysis

When restricting our pipeline to days 14–90 (matching the Russell et al. [12] censoring of the run-in), Control-arm mean glucose drops from 180 to 174 mg/dL and Control TIR rises from 54.3% to 55.1%, narrowing the gap to the published numbers ($\sim$169 mg/dL and $\sim 54\%$). The remaining $\sim$5 mg/dL discrepancy is plausibly attributable to (a) NEJM's per-protocol exclusion criteria that we do not implement (we report intent-to-treat metrics on the full coverage-passing cohort), and (b) calibration differences between the local CGM stream and the trial's central-lab calibration adjustments.

C Bootstrap Confidence Intervals for T5

T5 test-set bootstrap (subject-level percentile method, $B{=}1000$, `seed` $=$ 20240301, identical scheme to T1–T4 in `scripts/07_bootstrap_baselines.py`). Reproduced by `scripts/11_bootstrap_t5_baselines.py`; full per-metric CIs in `data/processed/baselines_t5_ci.json`.

- Linear (median-imputed): MAE $\in [5.09, 7.30]\%$ TIR; T5c AUROC $\in [0.71, 0.92]$.
- HistGBR: MAE $\in [5.41, 7.64]$; T5c AUROC $\in [0.69, 0.90]$.
- Arm $\times$ age lookup: MAE $\in [5.83, 8.95]$; T5c AUROC $\in [0.54, 0.81]$.
- Constant: MAE $\in [7.91, 11.51]$; AUROC degenerate (0.50) by construction (no class signal).

Resamples whose binarized target is degenerate (all responders or all non-responders) are excluded from the AUROC tally; with the released seed this drops 1 of 1000 resamples for the Linear and HistGBR baselines. Confidence intervals are wider than the values reported in an earlier draft of this manuscript; the present numbers are the reproducible output of `scripts/11_bootstrap_t5_baselines.py` and should be cited in preference to any earlier-draft figures.

D Compute Resources and Reproducibility

The full pipeline (data engineering, alignment, splits, T1–T5 baseline training, evaluation) runs in approximately 12 minutes on a single M-series MacBook Pro (16 GB RAM, no GPU). The largest in-memory working set is the iLet decision log at $\sim$1.7 GB. All scripts use deterministic seeds (`seed=20240301` for the canonical split, `random_state=0` for sklearn models). The 27 unit tests run in under 5 seconds.