Supplementary Materials: Subgroup Calibration of HbA1c Predictions Across NHANES Cycles: Partition Choice and Temporal Transport

Hussien Alwedyan [1,*], Ghassan Malkawi [2], Ahmed Abdelaziz Elsayed [3], Abdulwehab Ibrahim [1], Ashraf AlShalalfeh [1], Rafiq Manna [1], Mohanad Alata [4] and Mohammad AlWedian [5]

This supplement separates the additional sensitivity analyses (Sections S1–S19, Tables S1–S14, including Table S7a) from the technical provenance and reproducibility evidence retained in Supplementary Technical Section A (Tables A1–A18 and Figures A1–A10). The separate aggregate-data package contains the numerical records used in these summaries.

Abbreviations used throughout are body mass index (BMI), triglycerides (TG), ordinary least squares (OLS), root mean squared error (RMSE), conformalized quantile regression (CQR), primary sampling unit (PSU), mean absolute error (MAE), mean squared error (MSE), fuzzy inference system (FIS), and independent and identically distributed (i.i.d.).

S1. Source reconstruction and cohort closure

The legacy 2,325-row table was reconstructed against the official National Health and Nutrition Examination Survey (NHANES) 2017–2018 J-cycle records using age, BMI, fasting glucose, TG, glycated hemoglobin (HbA1c), and calculated Friedewald low-density lipoprotein (LDL). The reconstruction produced 2,325/2,325 exact unique matches, zero ambiguous matches, and zero unmatched records.

The historical 2,350 count was reproduced after age and core-variable completeness. Exactly 25 otherwise model-variable-complete records had unavailable calculated LDL, all with TG $>$400 mg/dL. The corrected primary analysis removed the LDL-completeness requirement and applied age $\geq$20 years, complete BMI/glucose/TG/HbA1c, fasting duration 8 to $<$24 hours, and valid positive WTSAF2YR. The corrected source cohort was N=2,219: 666/991/562 across ages 20–39/40–64/65+.

The released age variable is top-coded at 80. Age 80 is therefore treated as 80+, not as a known exact age beyond the public code.

Table S1 summarizes the cohort reconstruction and closure checks.

Table S1. Source reconstruction and cohort-closure checks.

Check	Result
Legacy rows uniquely matched	2,325 / 2,325
Ambiguous / unmatched rows	0 / 0
Distinct SEQN identifiers	2,325
Core-variable-complete source	2,350
LDL-only exclusions	25; all TG $>$400 mg/dL
Corrected cohort	2,219
Ages 20–39 / 40–64 / 65+	666 / 991 / 562

S2. Original model evaluation protocol and implementation

The submitted source analysis used five outer folds by five repeats, stratified by the original age groups. Within each outer-development set, 20% was reserved as calibration-only data. Test and calibration outcomes were excluded from model fitting and tuning.

Point models: glucose-only OLS; pooled four-input OLS; hard-age OLS; smooth continuous-age varying-coefficient OLS; pooled LightGBM; hard-age LightGBM.

The smooth varying-coefficient OLS used a cubic age-spline basis with five quantile knots, degree 3, no spline-bias column, and linear extrapolation. BMI, glucose, and TG were standardized within each source-training set. The linear design combined the age-spline basis, the standardized metabolic predictors, and all age-basis $\times$ metabolic-predictor interactions; the fitted OLS included an intercept.

The LightGBM search grid was num_leaves $\in \{7, 15, 31\}$, min_child_samples $\in \{20, 50\}$, and learning_rate $\in \{0.03, 0.07\}$. Fixed settings were n_estimators=300, max_depth=-1, subsample=0.90, colsample_bytree=1.0, reg_alpha=0, and reg_lambda=1.

The additional analyses do not alter the original point-model tuning implementation. The Methods report the supported grid and training-role separation. The original source script was recovered and hash-checked. Selection minimized mean inner-fold RMSE; candidates within 10^{-12} of the minimum were ordered by fewer leaves, larger min_child_samples, and lower learning_rate. subsample_freq=1 made the declared 0.90 row subsampling active. The original implementation used deterministic=True, force_col_wise=True, and n_jobs=1. This code inspection is distinct from a new model-fitting run.

S2a. Frozen temporal pooled-OLS certificate

The submitted frozen temporal pooled OLS used N=1,775 source records for model fitting and N=444 for conformal calibration, with random_state=20260912.

$$\begin{aligned}\widehat{\mathrm{HbA1c}} = {} & 2.44967007534616 + 0.00752952638869337\,\mathrm{Age} \\ & + 0.00755495809294163\,\mathrm{BMI} + 0.0246610043180617\,\mathrm{Glucose} \\ & + 0.0000163474392702398\,\mathrm{TG}.\end{aligned} \tag{S1}$$

Age is in released years (80 denotes 80+), BMI in kg/m^2, glucose and TG in mg/dL, and predicted HbA1c in percent.

The frozen global residual quantiles were $q_{90} = 0.807013028$ and $q_{95} = 1.139892692$, corresponding to nominal 90% and 95% residual intervals; in the main notation these are $q_{1-\alpha}$ for $\alpha = 0.10$ and 0.05. The age-specific q_{90} values were 0.639485528, 0.753998898, and 1.032731396 for ages 20–39, 40–64, and 65+, respectively.

S3. Analysis chronology

Table S2 summarizes the analysis chronology and status.

This chronology is descriptive. The post hoc sensitivity analyses were conducted after initial outcome inspection and are not treated as preregistered tests.

Table S2. Analysis status and chronology. A recorded protocol date is provenance evidence, not independent proof of absence of earlier data inspection.

Analysis	Status
Source reconstruction	Original provenance analysis
Repeated out-of-fold (OOF) models	Original primary source analysis
Global and age-Mondrian intervals	Original primary conformal analysis
Forward 2021–2023 transport	Original frozen temporal evaluation
Internal CQR	Original secondary analysis
Forward CQR	Secondary; implemented after initial linear target evaluation
Backward 2015–2016 test	Secondary; protocol dated 12 September 2026
Calibration curves / precision	Descriptive and retrospective
Fixed-configuration CQR	Post hoc exploratory
200-split calibration partitions	Post hoc sensitivity; 1 October 2026
200-partition OLS refitting	Post hoc sensitivity; 1 October 2026
Fixed-forward paired bootstrap	Post hoc sensitivity; 1 October 2026
Forward survey-design estimates	Post hoc sensitivity; 1 October 2026
Masked-PSU grouped source runs	Post hoc sensitivity; 1 October 2026
Added sex/race-ethnicity	Post hoc sensitivity; 1 October 2026
Alternative age boundaries	Post hoc sensitivity; 1 October 2026
Case-mix standardization	Post hoc sensitivity; 1 October 2026
Continuous-age smoother	Post hoc sensitivity; 1 October 2026

S4. Original internal point-prediction results

Table S3 reports the original repeated-OOF point-model results.

Table S3. Original repeated-OOF point-prediction RMSE. The hard-age OLS contrast is +0.004169, with conditional participant-bootstrap interval [-0.002639, +0.012260].

Model	RMSE
Pooled OLS	0.573252
Hard-age OLS	0.577421
Glucose-only OLS	0.586236
SmoothVC OLS	0.591184
Pooled LightGBM	0.593745
Hard-age LightGBM	0.660133

The participant bootstrap carried all repeated OOF records for a sampled participant together. It did not refit the models and therefore does not represent full training-set uncertainty.

S5. Point-model refitting and cluster-separation sensitivities

A 200-partition refitting sensitivity used new source partitions with common outer test roles for pooled and hard-age OLS.

Table S4 summarizes the 200-partition refitting sensitivity.

These split percentiles are stability summaries conditional on the observed source cohort, not population confidence intervals.

A second sensitivity separated public masked stratum–PSU clusters across outer folds.

Table S5 reports the masked-PSU grouped source sensitivity.

Public masked PSUs are not household identifiers or confidential true sampling units. This sensitivity addresses one form of cluster separation only.

Table S4. Source refitting sensitivity across 200 partitions. Percentiles are empirical split-distribution summaries, not population confidence intervals.

Quantity	Mean	Split 2.5–97.5%
Pooled OLS RMSE	0.573505	0.571175–0.577463
Hard-age OLS RMSE	0.577028	0.571349–0.583754
Hard-minus-pooled RMSE	+0.003523	-0.001388 to +0.009017
Hard-age worse share	0.910	—

Table S5. Masked-PSU grouped source sensitivity. All 200 retained runs were valid; masked PSUs are not household identifiers.

Quantity	Value
Mean hard-minus-pooled RMSE	+0.003935
Split 2.5–97.5%	+0.000561 to +0.008454
Hard-age worse share	0.985
Valid runs	200 / 200

S6. Original residual-conformal results

Table S6 summarizes the original internal coverage and interval widths.

Table S6. Original internal 90% residual intervals. Width is in HbA1c percentage points.

Method	Age	Coverage	Mean width
Global	20–39	0.962162	1.593
Global	40–64	0.892836	1.593
Global	65+	0.856228	1.593
Age-Mondrian	20–39	0.912012	1.219
Age-Mondrian	40–64	0.903734	1.693
Age-Mondrian	65+	0.893594	1.929

The age-Mondrian change is therefore a redistribution of both coverage and width, not a uniform widening or uniform improvement.

S7. 200-split calibration-partition sensitivity

This analysis used 200 source training/calibration allocations, seeds 20261001–20261200. Each split used an 80/20 source allocation stratified by the original age groups. The pooled OLS was fitted on source-training observations only, residual quantiles were calculated on source-calibration observations only, and the 2021–2023 target was evaluated without target fitting or recalibration.

Residual calibration schemes: M0 global residual; M1 age-Mondrian 20–39 / 40–64 / 65+; M2 fasting-glucose Mondrian <100 / 100 to <126 / ≥ 126 mg/dL; M3 predicted-HbA1c bins from source-training prediction tertiles. All schemes were evaluated on the same original age reporting groups.

Table S7 compares the four calibration schemes across the 200 common source splits.

These results do not establish a universal best calibration partition. They show that the conclusion about age-group balance depends materially on the calibration variable and source allocation.

The retained split-level workbook stores overall coverage, minimum age-group coverage, and max–min spread for each method. Because the forward target age-group sizes are fixed (670/1,203/935), the three age-group coverages are uniquely recoverable as integer-covered counts for every one of the 800 split–method rows. Table S7a reports those

Table S7. Forward 90% residual calibration across 200 common source splits. Spread values are percentage points (pp); mean width and interval score are expressed on the HbA1c scale.

Quantity	Global	Age-Mondrian	Glucose-Mondrian	Predicted-HbA1c bins
Mean age spread (pp)	7.40	4.27	2.51	4.18
Median age spread (pp)	7.43	3.99	2.42	3.90
Split 2.5–97.5% (pp)	6.40–8.36	0.51–8.42	1.04–4.27	2.77–7.06
Mean minimum coverage	0.886807	0.891883	0.895768	0.891128
Mean overall coverage	0.909006	0.913727	0.904411	0.908800
Mean width	1.560791	1.670331	1.568452	1.573326
Mean interval score	2.648046	2.644336	2.359701	2.452849

distributions. This is an aggregate reconstruction from retained split outputs, not a new model fit.

Table S7a. Age-group-specific forward 90% coverage across the 200 common source splits. Values are mean coverage with empirical 2.5th–97.5th split percentiles.

Calibration	Age group	Mean coverage	Split 2.5–97.5%
Global residual	20–39	0.9608	0.9403–0.9761
Global residual	40–64	0.8974	0.8753–0.9186
Global residual	65+	0.8868	0.8674–0.9113
Age-Mondrian	20–39	0.8999	0.8612–0.9403
Age-Mondrian	40–64	0.9091	0.8745–0.9360
Age-Mondrian	65+	0.9296	0.8855–0.9595
Glucose-Mondrian	20–39	0.9204	0.8821–0.9493
Glucose-Mondrian	40–64	0.8958	0.8678–0.9169
Glucose-Mondrian	65+	0.9041	0.8759–0.9241
Predicted-HbA1c bins	20–39	0.9022	0.8567–0.9388
Predicted-HbA1c bins	40–64	0.8941	0.8637–0.9211
Predicted-HbA1c bins	65+	0.9324	0.9101–0.9509

For a prediction interval $[l_i, u_i]$ and observed HbA1c y_i, the interval score at miscoverage α is

$$\mathrm{IS}_\alpha(l_i, u_i; y_i) = (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i)\mathbf{1}\{y_i < l_i\} + \frac{2}{\alpha}(y_i - u_i)\mathbf{1}\{y_i > u_i\}. \tag{S2}$$

Here l_i and u_i are the lower and upper interval bounds for record i, $\mathbf{1}$ is an indicator, and $\alpha = 0.10$. Scores are averaged over the evaluation records; lower is better.

S8. Alternative age boundaries and added-predictor sensitivities

Alternative age-Mondrian cutoffs 20–44 / 45–64 / 65+ were evaluated while retaining the original age groups for reporting.

Table S8 reports the alternative age-cutoff sensitivity.

The added-predictor OLS used RIAGENDR for sex and RIDRETH3 for race/Hispanic origin in addition to age, BMI, glucose, and TG, using the same source-split roles. RIAGENDR codes 1=male and 2=female; RIDRETH3 codes 1=Mexican American, 2=Other Hispanic, 3=Non-Hispanic White, 4=Non-Hispanic Black, 6=Non-Hispanic Asian, and 7=Other race including multiracial. For reproducible parameterization, indicators were represented with RIAGENDR=1 and RIDRETH3=1 as reference levels; under an intercept-

Table S8. Alternative age cutoffs 20–44 / 45–64 / 65+, evaluated using the original age reporting groups.

Quantity	Value
Mean age spread (pp)	3.87
Median age spread (pp)	3.61
Split 2.5–97.5% (pp)	0.65–8.21
Mean minimum coverage	0.895485

containing OLS model, alternative reference choices span the same categorical design space [A1].

Table S9 reports the added sex/race-ethnicity sensitivity.

Table S9. Added sex/race-ethnicity sensitivity on the same 200 source allocations.

Quantity	Value
Four-input mean forward RMSE	0.567857
Augmented-model mean forward RMSE	0.559681
Augmented RMSE split 2.5–97.5%	0.557921–0.562022
Mean global age spread (pp)	7.43
Mean minimum age coverage	0.895055

The additional measured demographic variables improved point error modestly but did not remove the age-related coverage spread under one global residual quantile.

S9. Fixed forward coverage bootstrap

Using the submitted frozen OLS and source q90 values, we performed a 10,000-resample paired participant bootstrap. Resampling was stratified by the original target age groups. This analysis is conditional on the fixed model and intervals and does not include full model-training uncertainty.

Table S10 reports fixed-forward coverage and paired bootstrap changes.

Table S10. Fixed-forward participant-bootstrap results. Changes are age-Mondrian minus global. Coverage and changes are proportions; multiply changes by 100 for percentage points.

Statistic	Age	Estimate	Lower 95%	Upper 95%
Spread change	All	-0.026592	-0.061478	0.003244
Minimum coverage change	All	-0.003250	-0.017456	0.023019
Global coverage	20–39	0.965672	0.952239	0.979104
Age-Mondrian coverage	20–39	0.923881	0.902985	0.943284
Global coverage	40–64	0.905237	0.887781	0.921862
Age-Mondrian coverage	40–64	0.891937	0.873649	0.909393
Global coverage	65+	0.895187	0.874866	0.914439
Age-Mondrian coverage	65+	0.935829	0.919786	0.950802

Table S10 gives the group estimates and the paired changes with their bootstrap intervals.

Thus, the fixed forward spread reduction does not demonstrate improved minimum-group coverage.

S10. Temporal survey-design sensitivity

Both temporal population-oriented sensitivities used WTSAF2YR, SDMVSTRA, and SDMVPSU. Taylor-linearized standard errors (SEs) were calculated for weighted age-specific coverage and weighted calibration-in-the-large (CITL). These are survey estimates, not conformal guarantees.

Table S11 reports survey-weighted age coverage and design-based uncertainty for both temporal cohorts.

Table S11. Survey-weighted 90% age coverage in both temporal cohorts. Intervals use the t distribution with 15 design degrees of freedom and Taylor-linearized domain standard errors. Original forward normal-approximate intervals are preserved in the separate data file, not relabelled as t intervals.

Cycle	Method	Age	Coverage	SE	t-based 95%
2015-2016	Global	20-39	0.9560	0.0095	0.9357–0.9762
2015-2016	Global	40-64	0.9083	0.0163	0.8737–0.9430
2015-2016	Global	65+	0.8776	0.0135	0.8489–0.9063
2015-2016	Age-Mondrian	20-39	0.9175	0.0088	0.8988–0.9361
2015-2016	Age-Mondrian	40-64	0.8944	0.0158	0.8608–0.9281
2015-2016	Age-Mondrian	65+	0.9375	0.0087	0.9189–0.9561
2021-2023	Global	20-39	0.9686	0.0056	0.9567–0.9804
2021-2023	Global	40-64	0.9115	0.0101	0.8899–0.9330
2021-2023	Global	65+	0.8962	0.0167	0.8606–0.9317
2021-2023	Age-Mondrian	20-39	0.9239	0.0098	0.9031–0.9447
2021-2023	Age-Mondrian	40-64	0.8978	0.0103	0.8758–0.9198
2021-2023	Age-Mondrian	65+	0.9328	0.0129	0.9053–0.9603

Public-use survey design degrees of freedom in this calculation: 15.

Forward weighted CITL (observed minus predicted): estimate -0.012197; SE 0.015406. The corresponding approximate design-based interval includes zero.

The backward cohort remains a secondary stress test. Its archived predictions and masked design fields were recovered, permitting the same domain variance calculation in both temporal cohorts. Table S11 reports both cycles with t-based intervals; normal-approximate intervals are preserved separately in the aggregate CSV.

S11. Forward case-mix analysis

The frozen submitted OLS was evaluated by observed HbA1c category.

Table S12 summarizes the forward case-mix sensitivity.

Table S12. Forward case-mix sensitivity. Overall RMSE is 0.567696; source-composition standardized RMSE is 0.598805.

Observed HbA1c	N	Forward proportion	Source proportion	RMSE	MAE
$<$5.7%	1707	0.607906	0.539432	0.3983	0.2999
5.7–6.4%	751	0.267450	0.308698	0.3562	0.2814
$\geq$6.5%	350	0.124644	0.151870	1.2408	0.9013

Observed forward overall RMSE: 0.567696. Forward RMSE after descriptively reweighting category-specific mean squared error (MSE) to the source HbA1c-category proportions: 0.598805.

The standardization is descriptive and does not identify a causal explanation for model-ranking change.

S12. Measurement harmonization and recovered raw-TG sensitivity

For the submitted frozen OLS, later-cycle bridged TG reproduced RMSE 0.56769595. Replacing bridged TG with released raw TG while holding the model coefficients fixed gave RMSE 0.56770187. The maximum absolute change in an individual OLS prediction was 0.00081259 HbA1c percentage points.

The recovered historical tree-model predictions permit paired raw-versus-bridged comparisons without a new model fit (Table S13). Pooled LightGBM RMSE was 0.544154 with bridged TG and 0.546519 with raw TG; its ordering relative to pooled OLS persisted. This check does not identify the cause of the model-ranking reversal.

The revised assay ledger additionally records the 2021–2023 HbA1c instrument change from Tosoh G8 to Bio-Rad D-100 and the NHANES conclusion that the released GHB values did not require adjustment after the paired comparison study.

Table S13 reports the paired stored-prediction comparison.

Table S13. Recovered paired historical predictions under raw and bridged later-cycle TG. These are reaggregated stored-output results, not newly refitted historical models.

Model	Bridged-TG RMSE	Raw-TG RMSE	Maximum absolute prediction change
Glucose-only OLS	0.579195	0.579195	0.000000
Pooled OLS	0.567696	0.567702	0.000813
Hard-age OLS	0.566203	0.566155	0.002725
Smooth varying-coefficient OLS	0.570933	0.571296	0.029000
Pooled LightGBM	0.544154	0.546519	0.675551
Hard-age LightGBM	0.621129	0.620449	0.550855

S13. Continuous-age coverage sensitivity

The supplementary analysis data contain descriptive continuous-age coverage estimates for released ages 20–80 for global and age-Mondrian residual intervals. A Gaussian kernel with a four-year bandwidth was used only as a descriptive smoother, with effective local sample size reported. Age 80 is the public top-coded 80+ category and is not interpreted as an exact age above 80.

The continuous display is descriptive and does not establish pointwise conditional coverage.

S14. Archived CQR comparator

The archived secondary CQR implementation used LightGBM quantile regression and is kept separate from the 200-split residual-partition experiment.

The original secondary internal 90% analysis had age-group coverage 0.907207 / 0.897074 / 0.901068, an age spread of 1.01 percentage points, and an interval score of 2.413035.

The original secondary forward 90% analysis had age-group coverage 0.928358 / 0.919368 / 0.921925, an age spread of 0.90 percentage points, and an interval score of 2.283125.

The original fitted quantile-model objects are not part of the available archive. These values are verified from stored interval outputs and are not described as a newly reproduced historical training run.

S15. Backward adjacent-cycle stress test

The 2015–2016 secondary cohort contained N=2,235 participants (712/970/553 by original age group). Submitted pooled-OLS RMSE was 0.657994, MAE 0.397868, prediction-minus-observed bias -0.042936, and prediction-error coefficient of determination (R^2) 0.732433, using the definition in Eq. (A-R2-1).

Submitted 90% global age-group coverage was 0.9579 / 0.8804 / 0.8535, with a spread of 10.43 percentage points. Age-Mondrian coverage was 0.9185 / 0.8619 / 0.9005, with a spread of 5.67 percentage points.

Target HbA1c standard deviation (SD) was approximately 1.2723, compared with approximately 1.0587 in the forward reconstruction. This supports caution when comparing overall RMSE between cycles with different outcome dispersion.

S16. Legacy adaptive-network-based fuzzy inference system (ANFIS) material

The archived ANFIS reconstruction remains supplementary. The original checking observations overlapped training observations, so historical checking error is not treated as independent validation. The archived FIS parameters and stored-model predictions remain valuable for structural reproducibility, but their historical performance is not mixed with the corrected repeated-OOF point-model comparison.

The earlier ANFIS work was an unpublished internal analysis; the archived objects are retained only as a computational provenance case.

S17. Claim-to-evidence boundary

Supported:

- exact source reconstruction and cohort correction.
- no demonstrated point-prediction benefit from hard age partitioning under the tested restricted structure.
- age-related subgroup coverage imbalance under a single global residual quantile.
- subgroup calibration results depend on the calibration partition and source allocation.
- lower max–min spread does not necessarily improve the least-covered group.
- forward overall RMSE depends materially on outcome composition.
- survey-aware uncertainty can be wider than i.i.d.-style interpretations.

Not supported:

- equivalence of pooled and hard-age models.
- universal superiority of age-Mondrian, glucose-Mondrian, CQR, OLS, or LightGBM.
- causal effects of age, sex, race/ethnicity, assay change, or survey design on prediction error.
- exact individual conditional coverage.
- survey-population conformal guarantees.
- independent-health-system validation.
- replacement of laboratory HbA1c or clinical decision benefit.

S18. Machine-readable aggregate outputs

The supplementary aggregate-data package contains clean CSV exports for the source-split analysis, point-model refitting, masked-PSU sensitivity, paired bootstrap, survey estimates, case-mix analysis, and continuous-age display, together with retained aggregate JSON records and reaggregation checks. These files contain aggregate outputs rather than participant-level records.

Participant-level source and target records are not redistributed. The additional aggregate outputs are kept separate from the historical v1.0.2 record.

S19. Recovered-evidence checks and domain-variance implementation

The original point-model script was recovered with SHA-256 `4eba01f4d74cc7c221 11a22968023e5f5b561ed3d2f3faebe2db50221e95ab85`. The paired temporal prediction file contains both raw-TG and bridged-TG predictions for the same 2,808 participants and six models. Table S13 reaggregates those records without retraining. Table S14 separately reports the free calibration intercept and slope, rather than substituting CITL for the intercept. The backward cohort contains the design variables needed for the domain estimates in Table S11.

For domain indicator d_i, value z_i (coverage indicator or observed-minus-predicted error), and positive fasting weight w_i, define

$$\widehat{\mu}_d = \frac{\sum_i d_i w_i z_i}{\sum_i d_i w_i}, \qquad \ell_i = \frac{d_i w_i (z_i - \widehat{\mu}_d)}{\sum_j d_j w_j}. \tag{S3}$$

Thus, $\widehat{\mu}_d$ is the weighted mean within domain d, and ℓ_i is the record-level linearized contribution used in its variance calculation. Records outside the domain have $\ell_i = 0$ and are retained for the design layout. With masked stratum h, masked PSU c, cluster total $L_{hc} = \sum_{i \in (h,c)} \ell_i$, m_h observed PSUs per stratum, and $\overline{L}_h = m_h^{-1} \sum_c L_{hc}$, the with-replacement Taylor variance is

$$\widehat{V}(\widehat{\mu}_d) = \sum_h \frac{m_h}{m_h - 1} \sum_c (L_{hc} - \overline{L}_h)^2. \tag{S4}$$

Both temporal analytic cohorts contain 15 masked strata and 30 masked PSUs, with no singleton stratum. The standard error is $\sqrt{\widehat{V}}$ and the t-based interval is $\widehat{\mu}_d \pm t_{0.975,15}\sqrt{\widehat{V}}$, where $t_{0.975,15}$ is the 97.5th percentile of Student's t distribution with 15 degrees of freedom. The normal-approximate interval instead uses 1.96; it is recorded separately. No finite-population correction is imposed. Complete-case selection and public masking still limit population interpretation.

Weighted backward CITL is -0.038139 (SE 0.022272; t-based 95% interval $[-0.085610, +0.009332]$). Weighted forward CITL is -0.012197 (SE 0.015406; interval $[-0.045035, +0.020641]$). These do not substitute for the unweighted regression diagnostics below.

Table S14. Recomputed linear point-calibration diagnostics from the frozen temporal predictions. CITL is observed minus predicted and differs from the free regression intercept.

Cycle	N	Prediction-error R^2	CITL	Intercept	Slope
2015-2016	2235	0.732433	+0.042936	-0.203274	1.042108
2021-2023	2808	0.712345	-0.018213	-0.291855	1.047288

The source-split workbook was checked by recomputing means, empirical quantiles, paired RMSE differences, and worse-model shares from all 200 retained rows in each relevant sheet. This summary check is not an independent rerun of the generating model fits. Historical CQR coverage and interval scores were reaggregated from retained interval outputs; the original fitted quantile models remain unavailable. These evidence categories are deliberately separated.

Supplementary Technical Section A. Technical provenance and retained reproducibility evidence

This technical section preserves historical analysis tables, model and calibration identities, the ANFIS reconstruction, retrospective precision summaries, and reporting maps as provenance evidence. These retained materials are interpreted under the current claim boundaries in Sections S3 and S17 and are reported as historical provenance, not as evidence of preregistration.

A1. Provenance reconstruction and cohort closure

The legacy 2,325-row table was matched to official National Health and Nutrition Examination Survey (NHANES) 2017–2018 J-cycle records using the six-field retained fingerprint: age, body mass index (BMI), fasting glucose, triglycerides (TG), HbA1c, and calculated Friedewald low-density lipoprotein (LDL). The final matching certificate records 2,325/2,325 exact unique matches, 0 ambiguous matches, 0 unmatched rows, and 2,325 distinct NHANES respondent sequence number (SEQN) values. In the complete J-cycle candidate pool, the matching fingerprint was also unique for the retained analytic records, so the reconstruction did not rely on arbitrary tie breaking.

The public-use age-in-years variable (RIDAGEYR) is top-coded at 80; RIDAGEYR=80 is treated as age 80+ according to the official DEMO_J codebook [A1].

The historical 2,350 count is reproduced after age and core-variable completeness. Exactly 25 records have missing calculated LDL, all with TG $>$ 400 mg/dL, consistent with the source laboratory documentation that calculated Friedewald LDL is not reported above 400 mg/dL [A2]. Requiring nonmissing LDL therefore yields 2,325. This is described as a previously undocumented LDL-completeness restriction rather than an HbA1c feature-selection step.

Table A1 summarizes the row-level provenance and cohort-closure checks.

The 2,325 legacy table includes 131 rows with zero 2-year fasting subsample weight (WTSAF2YR) or fasting duration outside 8 to $<$ 24 hours [A2,A3]. The primary cohort removes those rows and does not require LDL, resulting in N=2,219. The counts by age are 666 (20–39), 991 (40–64), and 562 (65+).

Provenance check	Result
Legacy rows matched to J-cycle	2,325 / 2,325 exact unique
Ambiguous matches	0
Unmatched rows	0
Distinct recovered SEQN	2,325
Model-variable-complete source count	2,350
Rows removed only by LDL completeness	25; all TG $>$ 400 mg/dL
Legacy rows outside valid fasting subsample	131
Corrected primary cohort	2,219

Table A1. SEQN/provenance certificate summary.

A2. Original source analysis protocol

The original primary source analysis was frozen before model comparison. Outer evaluation used 5 folds $\times$ 5 repeats, stratified only by age group and seeded at 20260912. Within each outer-development set, 20% was reserved as calibration-only data using a deterministic age-stratified split. All models shared the same participant roles within a

split. Calibration and outer-test outcomes were excluded from fitting, preprocessing, and hyperparameter selection.

Point models were glucose-only ordinary least squares (OLS); pooled four-input OLS; hard-age OLS; fixed smooth varying-coefficient OLS; pooled LightGBM; and hard-age LightGBM. LightGBM inner tuning used 3-fold cross-validation (CV) on model-training data with num_leaves $\in \{7, 15, 31\}$, min_child_samples $\in \{20, 50\}$, learning_rate $\in \{0.03, 0.07\}$, n_estimators=300, max_depth=-1, subsample=0.90, colsample_bytree=1.0, reg_alpha=0, and reg_lambda=1. The primary metric was root mean square error (RMSE); participant-level paired uncertainty used 10,000 age-stratified bootstrap resamples.

A2a. Exact frozen temporal OLS and conformal calibration certificate

The final temporal pooled OLS and residual-conformal calibration were frozen from one deterministic age-stratified source split before target outcomes were used. Model fitting used 1,775 source rows and conformal calibration used 444 source rows (20% calibration; random_state=20260912). The predictor order was age, BMI, fasting glucose, and TG. The raw-scale fitted equation was:

$$\begin{aligned}\widehat{\mathrm{HbA1c}} = {} & 2.4496701 + 0.00752953\,\mathrm{Age} + 0.00755496\,\mathrm{BMI} \\ & + 0.02466100\,\mathrm{Glucose} + 0.0000163474\,\mathrm{TG}.\end{aligned} \tag{A-OLS1}$$

This equation maps the four measured predictors to the frozen HbA1c point prediction. Each coefficient describes an adjusted predictive association with the other listed predictors held fixed; it is not interpreted as a causal effect.

Table A2 reports the exact frozen split, coefficients, conformal quantiles, and file hashes used for both temporal evaluations.

Frozen item	**Value**
Model-training N	1,775
Conformal-calibration N	444
Split	Age-stratified; calibration=0.20; random_state=20260912
Raw intercept	2.44967007534616
Age coefficient	0.00752952638869337
BMI coefficient	0.00755495809294163
Fasting-glucose coefficient	0.0246610043180617
TG coefficient	0.0000163474392702398
Global q90 / q95	0.807013028 / 1.139892692
Age 20–39 q90 / q95	0.639485528 / 0.919630883
Age 40–64 q90 / q95	0.753998898 / 1.231843407
Age 65+ q90 / q95	1.032731396 / 2.189815895
Frozen JSON SHA-256	`980333416af8db7132897b41c9ee024a0d8a3eb6ec5dfcd639ea3440a8b69f90`
Source role-ledger SHA-256	`e027e808fdef242e1e9784729bcf8259a9c6a9ee76963fb5916498d9d4ed95c9`

Table A2. Exact final source-frozen pooled OLS and residual-conformal certificate used in both temporal evaluations.

A2b. CQR specifications and execution provenance

The archived conformalized quantile regression (CQR) implementation records, source code, tuning ledgers, and row-level predictions are held in verification materials retained

by the authors. The public reproducibility repository provides aggregate CQR metrics and verification records rather than participant-level predictions. Checksums and interval metrics were verified from the stored outputs. Original fitted quantile-model objects are not included; verification therefore checks those outputs rather than retraining the original quantile learners. The internal CQR analysis was designated as secondary; its forward temporal application remains secondary because implementation details were fixed after the first pooled-linear target evaluation. The separate post hoc fixed-configuration analysis is reported as an exploratory sensitivity [A4].

The original learner was LightGBM 4.6.0 quantile regression. Base quantiles were 0.05/0.95 at 90% and 0.025/0.975 at 95%. A 3-fold age-stratified inner CV selected among 12 combinations of num_leaves {7,15,31}, min_child_samples {20,50}, and learning_rate {0.03,0.07} using the mean of lower/upper pinball losses. Fixed settings were 300 estimators, max_depth=-1, subsample=0.90 with subsample_freq=1, colsample_bytree=1, reg_alpha=0, and reg_lambda=1; deterministic single-thread fitting was used. Candidates within 10^{-12} of the best loss were ordered by fewer leaves, larger minimum leaf size, then lower learning rate.

Original source roles match the 5×5 primary splits. With base seed B=20260912 and one-based repeat r and fold f, calibration split seeds were B+1000r+f and inner-CV seeds B+10000r+100f+1. Final lower/upper fitting seeds were B+70000r+700f+m and B+80000r+700f+m, where m is nominal coverage in integer percent. Temporal fitting used the frozen 1,775/444 source split, inner seed B+901+m, and final lower/upper seeds B+90100+m and B+90200+m. Base quantiles were ordered row-wise before scoring; pre-ordering crossings were logged. Exact source code is retained rather than replacing these seed rules by a new implementation. Here, k denotes the finite-sample order statistic and $\widehat{q}$ the conformal correction. Table A3a summarizes the recovered original temporal CQR specification.

Original temporal item	**90%**	**95%**
Base quantiles	0.05 / 0.95	0.025 / 0.975
Selected leaves / min child / learning rate	7 / 20 / 0.03	7 / 20 / 0.03
Source calibration N / order k	444 / 401	444 / 423
Conformal correction $\widehat{q}$	0.14156920289483388	0.19791465799729036
Pre-ordering target crossings	0	0

Table A3a. Recovered original temporal CQR specification. The finite-sample conformity score and interval construction are defined below; no target refitting or recalibration is used.

For calibration observation i, let x_i denote its predictor vector. With lower and upper base quantile predictions $\widehat{q}_L(x_i)$ and $\widehat{q}_U(x_i)$, the conformity score is

$$s_i = \max\{\widehat{q}_L(x_i) - y_i,\ y_i - \widehat{q}_U(x_i)\}. \tag{A-CQR1}$$

For nominal miscoverage α and calibration size n_{cal}, the finite-sample order statistic is

$$k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil, \qquad \widehat{q} = s_{(k)}, \tag{A-CQR2}$$

where $s_{(k)}$ denotes the k-th smallest calibration conformity score, and the conformalized interval is

$$C_\alpha(x) = [\widehat{q}_L(x) - \widehat{q},\ \widehat{q}_U(x) + \widehat{q}]. \tag{A-CQR3}$$

A negative $\widehat{q}$ shrinks the base quantile interval; this is permitted because the nonconformity scores themselves can be negative when the base interval already contains the observed outcome by a margin. Operationally, the quantile models provide lower and upper candidate bounds, and calibration residuals determine the conformal adjustment applied to those bounds.

Tables A3a and A3b distinguish the archived secondary CQR implementation from the separate post hoc exploratory sensitivity.

The separate post hoc fixed-configuration exploratory CQR used LightGBM 4.6.0 (LGBMRegressor) with quantile objectives at 0.05 and 0.95. Hyperparameters were num_leaves=15, min_child_samples=20, learning_rate=0.03, n_estimators=300, max_depth=-1, subsample=0.90, subsample_freq=1, colsample_bytree=1.0, reg_alpha=0, reg_lambda=1, deterministic=True, force_col_wise=True, and n_jobs=1. The same conformity score, finite-sample order statistic, and interval construction in Eqs. (A-CQR1)–(A-CQR3) were used. The final source split used the same 1,775/444 roles and seed 20260912; $\widehat{q} = 0.1914744965$ ($k = 401$). No target retraining or target recalibration was performed.

Table A3b. Separate post hoc fixed-configuration exploratory CQR sensitivity. These model objects and hashes correspond only to the post hoc exploratory analysis.

Post hoc exploratory CQR item	Exact value
Lower / upper quantile	0.05 / 0.95
Learner	lightgbm.LGBMRegressor 4.6.0
scikit-learn version	1.8.0
Final source train / calibration	1,775 / 444; seed 20260912
$\widehat{q}$ (90%)	0.1914744965; k=401
Lower-model SHA-256	`2e134f8fa766413fc648a475a03a1bf6c38bcedcb4edb53aae376e4449cc49a0`
Upper-model SHA-256	`adcb4b52d89c6d523ff8fc322cb108fa836cc1a7c4a0fa67ae29ac0c6535e943`
Analysis-script SHA-256	`b83dd738351a172d67d1ca34d0887477f78df5a862b83c6f4afab8f2a32cce57`
Status	Post hoc exploratory sensitivity; not part of the primary analysis

A3. Internal point-prediction results

Table A4 summarizes repeated out-of-fold point-model performance and participant-level paired contrasts; uncertainty is reported as a 95% confidence interval (CI).

Model	RMSE	Primary paired comparison	Difference	95% CI
Pooled OLS	0.573252	Reference	—	—
Hard-age OLS	0.577421	Hard-age - pooled OLS	+0.004169	[-0.002639, +0.012260]
Glucose-only OLS	0.586236	Glucose-only - pooled OLS	+0.012983	[+0.007637, +0.018213]
SmoothVC OLS	0.591184	SmoothVC - pooled OLS	+0.017931	[+0.004215, +0.037738]
Pooled LightGBM	0.593745	Pooled LGBM - pooled OLS	+0.020493	[+0.002452, +0.039668]
Hard-age LightGBM	0.660133	Hard-age - pooled LGBM	+0.066388	[+0.038151, +0.098418]

Table A4. Repeated out-of-fold point-model performance and participant-bootstrap contrasts.

Interpretation is intentionally model-class specific. The hard-age OLS contrast crosses zero; the hard-age LightGBM contrast does not. The source analysis therefore does not

support hard age partitioning as a general point-prediction improvement. The pooled four-input OLS has a small, internally supported advantage over glucose-only OLS.

A4. Internal 90% conformal uncertainty

Tables A5 and A6 summarize the primary age-specific coverage, interval width, interval score, and between-age coverage spread.

Primary method / group	Coverage	Mean width
Global residual: 20–39	0.962162	1.593
Global residual: 40–64	0.892836	1.593
Global residual: 65+	0.856228	1.593
Age-Mondrian: 20–39	0.912012	1.219
Age-Mondrian: 40–64	0.903734	1.693
Age-Mondrian: 65+	0.893594	1.929

Table A5. Primary internal empirical 90% coverage and width by age.

Primary method	Age spread	Overall interval score
Global residual	10.59 pp	2.687079
Age-Mondrian	1.84 pp	2.638516

Table A6. Primary internal interval-score and age-spread summary.

Tables A7a and A7b document the archived secondary CQR results and the separate post hoc exploratory sensitivity, respectively.

Recalculation from the recovered original interval predictions reproduces the archived secondary estimates in Table A7a. Recovery establishes the provenance of those outputs, not a new independent model-training experiment. No backward CQR result was available in the archived original analysis.

Original secondary analysis	20–39	40–64	65+	Spread, pp	Score
Internal original secondary	0.907207	0.897074	0.901068	1.01	2.413035
Forward original secondary	0.928358	0.919368	0.921925	0.90	2.283125

Table A7a. Original secondary 90% CQR results recomputed from archived interval outputs. All values were recomputed from archived intervals without refitting.

Post hoc exploratory CQR sensitivity produced overall 90% coverage 0.902118, mean width 1.615103, and interval score 2.417463. Age-specific coverage was 0.919520, 0.895459, and 0.893238, for a 2.63-point spread. These values are reported only to document reproducibility of the adaptive-interval analysis; they do not replace or retrospectively redefine the primary residual-conformal analysis.

Setting	Overall	20–39	40–64	65+	Age spread	Score
Internal 2017–2018	0.902118	0.919520	0.895459	0.893238	2.63 pp	2.417463

Setting	Overall	20–39	40–64	65+	Age spread	Score
Backward 2015–2016	0.918568	0.932584	0.914433	0.907776	2.48 pp	2.537669
Forward 2021–2023	0.920940	0.922388	0.916043	0.926203	1.02 pp	2.314123

Table A7b. Post hoc exploratory CQR sensitivity. These are distinct from the original secondary results in Table A7a.

A5. Forward temporal transport: NHANES 2021–2023

Table A8 summarizes forward temporal point and linear-calibration diagnostics; Figure A1 retains the grouped decile view, and Table A9 reports the primary 90% age-specific coverage.

The source-frozen pooled linear model was applied to N=2,808 target participants with no target retraining or recalibration. Primary TG values used the documented assay bridge; the raw-TG sensitivity changed pooled-linear predictions by at most 0.000813 HbA1c percentage points. The point-model result was RMSE 0.567696, mean absolute error (MAE) 0.369897, prediction-error coefficient of determination (R^2) 0.712345 (exact definition in A9), and prediction − observed bias +0.018213; fasting-weighted RMSE was 0.562038.

Table A8. Forward temporal point and continuous-calibration diagnostics. Calibration diagnostics are descriptive and do not update the frozen model.

Metric	Unweighted	Survey-weighted sensitivity
RMSE	0.567696	0.562038
MAE	0.369897	—
R^2	0.712345	—
Prediction − observed bias	+0.018213	—
Calibration intercept	-0.291855	-0.400821
Calibration slope	1.047288	1.067955

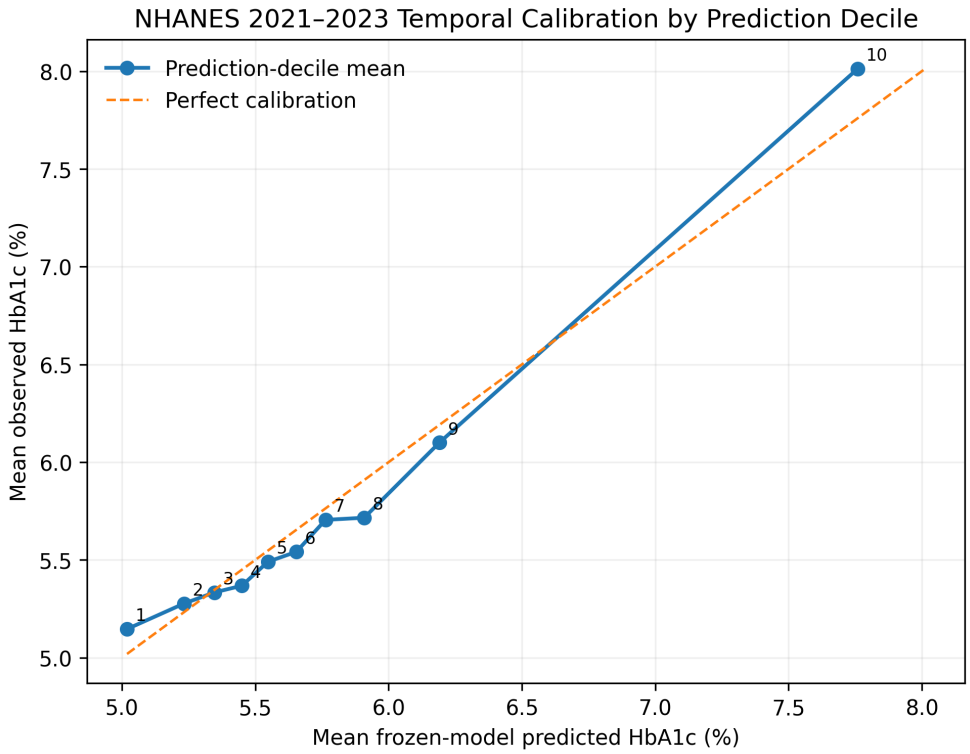


Figure A1. Forward 2021–2023 observed versus predicted HbA1c summarized by frozen-prediction decile. The smooth full-range calibration curve is shown in Supplementary Figure A10; this grouped plot is retained as a secondary descriptive view.

Primary 90% method	20–39 coverage	40–64 coverage	65+ coverage	Age spread
Global residual	0.965672	0.905237	0.895187	7.05 pp
Age-Mondrian	0.923881	0.891937	0.935829	4.39 pp

Table A9. Forward 2021–2023 primary 90% empirical coverage.

A6. Backward adjacent-cycle stress test: NHANES 2015–2016

Table A10 summarizes backward point and calibration metrics; Figures A2 and A3 show the grouped calibration and age-specific coverage views, and Table A11 reports the corresponding conformal coverage values.

The backward protocol was finalized before the 2015–2016 target outcomes were imported. The final cohort was N=2,235 (712 aged 20–39, 970 aged 40–64, and 553 aged 65+). The five XPT components used to construct the backward target cohort had unique SEQN; GLU and TRIGLY fasting weights agreed exactly on the join; retained rows had finite required values, positive WTSAF2YR, and valid age groups.

Metric	2015–2016 result
RMSE	0.657994
MAE	0.397868
Prediction − observed bias	-0.042936
R^2	0.732433
Fasting-weighted RMSE	0.546575
Fasting-weighted MAE	0.350520

Metric	2015–2016 result
Calibration-in-the-large (observed − predicted)	+0.042936
Calibration intercept	-0.203274
Calibration slope	1.042108
Weighted calibration intercept	-0.065667
Weighted calibration slope	1.004809

Table A10. Backward temporal point and continuous-calibration metrics.

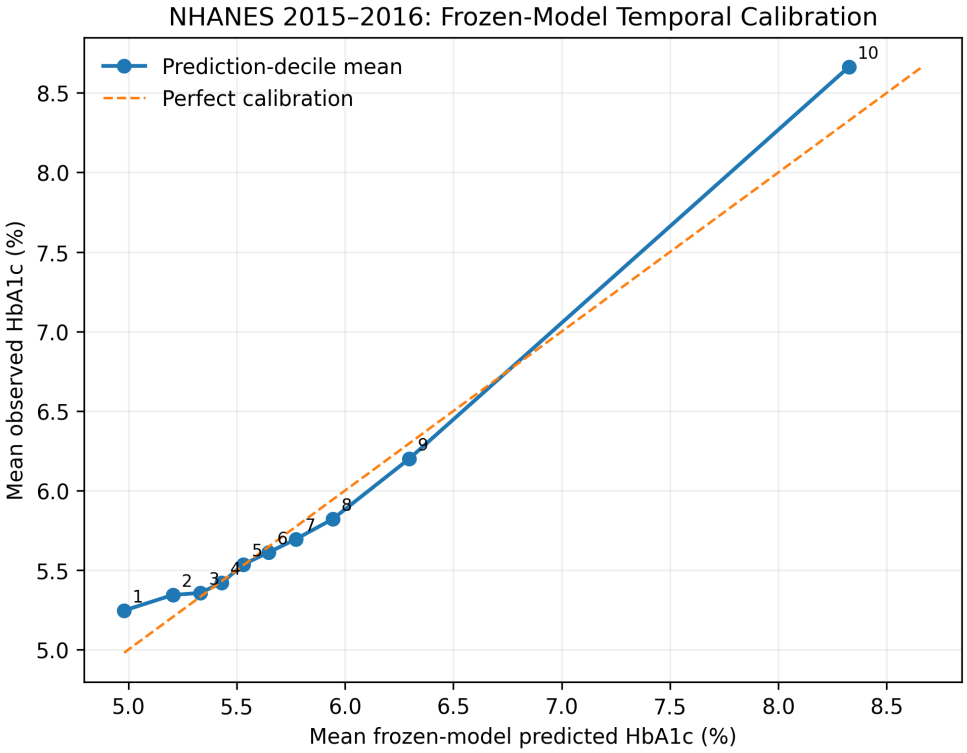


Figure A2. Backward 2015–2016 frozen-model calibration by prediction decile. The smooth full-range calibration curve is shown in Supplementary Figure A10.

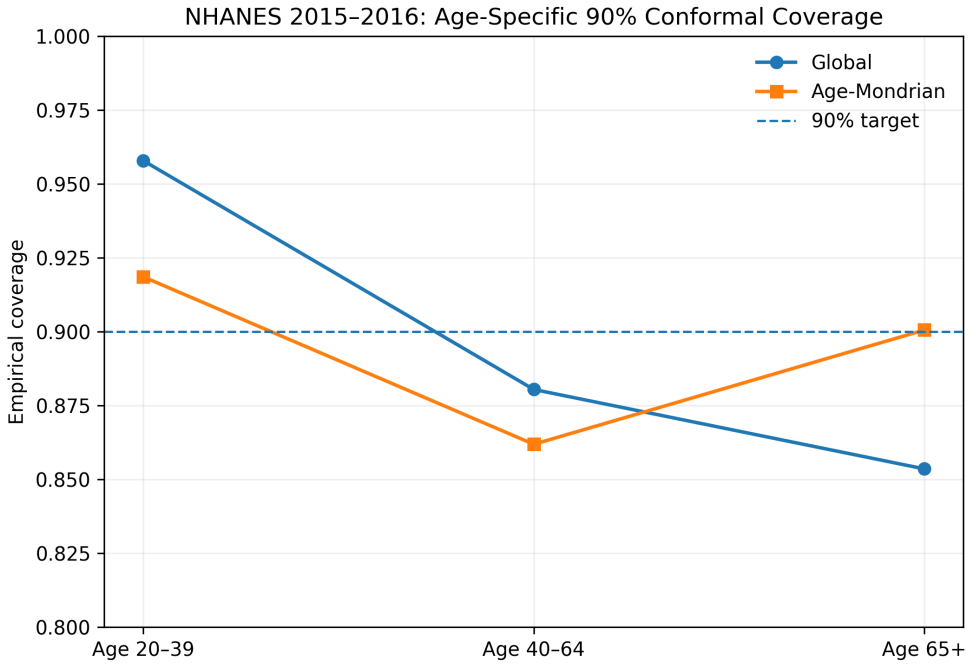


Figure A3. Backward 2015–2016 90% empirical coverage by age under global and age-Mondrian source calibration.

Nominal / method	Overall	20–39	40–64	65+	Age spread
90% global	0.8984	0.9579	0.8804	0.8535	10.43 pp
90% age-Mondrian	0.8895	0.9185	0.8619	0.9005	5.67 pp
95% global	0.9459	0.9789	0.9351	0.9222	5.67 pp
95% age-Mondrian	0.9602	0.9691	0.9412	0.9819	4.07 pp

Table A11. Backward source-calibrated conformal coverage. Age spreads were calculated from full-precision coverage values before rounding. The 95% subgroup values are reproduced from the archived frozen-prediction table using the source-frozen calibration quantiles.

Original unweighted target-source standardized mean differences were small: age -0.065, BMI -0.047, fasting glucose -0.030, TG -0.004, and HbA1c +0.003. These values indicate limited measured adjacent-cycle shift but do not establish exchangeability.

A7. NHANES assay and source-harmonization ledger

Table A12 provides the cycle-by-cycle assay and source-harmonization ledger.

Official NHANES laboratory documentation was reviewed specifically to separate real temporal transport from assay/instrument changes. For 2015–2016 glucose, NHANES documented a C501→C311 transition and forward-adjusted released C501 results to the C311 scale; therefore no additional glucose bridge was applied [A3,A5]. For 2021–2023 TG, the documented instrument/method change was handled by applying the published backward bridge to align the later-cycle measurement with the earlier assay scale [A6]. The CDC bridge paragraph labels the regression measurements as mg/L, whereas the released LBXTLG and earlier-cycle LBXTR variables are documented in their codebooks as mg/dL; the published equation was applied to the released variables on their documented codebook

scale. The full machine-readable ledger with complete SHA-256 values accompanies the archive.

Cycle / component	Variable	Method and bridge / handling	Official source / analysis-input hash prefix
2015–16 GLU_I	LBXGLU	Hexokinase; C501→C311. Released C501 values were already adjusted to the C311 scale; no extra bridge was applied.	CDC GLU_I [A5]; `3a52f1434606...`
2017–18 GLU_J	LBXGLU	Hexokinase; Cobas C311. No adjustment was required versus released 2015–16 values.	CDC GLU_J [A3]; `a985028efacc...`
2021–23 GLU_L	LBXGLU	Hexokinase on the C311 scale; no glucose equipment bridge was used.	CDC GLU_L [A7]; `d231f6b9e651...`
2015–16 TRIGLY_I	LBXTR / LBDLDL	Wahlefeld/Trig-GB; no TG bridge. Friedewald LDL is missing above TG 400 mg/dL.	CDC TRIGLY_I [A8]; `dd5a24f7b3bf...`
2017–18 TRIGLY_J	LBXTR / LBDLDL	Glycerol-blanked reference scale; no method/site change. Friedewald LDL is missing above TG 400 mg/dL.	CDC TRIGLY_J [A2]; `a985028efacc...`
2021–23 TRIGLY_L	LBXTLG	Cobas 6000→8000; non-blanked method. Primary backward bridge: $TG_{\text{old}} = -12.19 + 0.9785\, TG_{\text{new}}$; raw TG retained as sensitivity.	CDC TRIGLY_L [A6]; `d231f6b9e651...`

Table A12. NHANES assay/source ledger. The accompanying machine-readable CSV contains the complete source URLs and analysis-input SHA-256 values.

A8. Extended temporal point-model comparison

Table A13 summarizes the later-cycle model-ranking reversal, and Figure A4 contrasts sample-level with fasting-weighted RMSE for the frozen pooled OLS.

Model	Source repeated-OOF RMSE	2021–2023 RMSE
Glucose-only OLS	0.586236	0.579195
Pooled OLS	0.573252	0.567696
Hard-age OLS	0.577421	0.566203
SmoothVC OLS	0.591184	0.570933
Pooled LightGBM	0.593745	0.544154
Hard-age LightGBM	0.660133	0.621129

Table A13. Model-ranking reversal in the later cycle. The target cycle was never used for model fitting or tuning.

Pooled LightGBM was worse than pooled OLS in source repeated CV but better in 2021–2023. Hard-age LightGBM was worse in both settings. This is retained as evidence against a universal model winner, not as a post hoc selection of LightGBM.

A9. Retrospective continuous-outcome precision summary

Table A14 reports the retrospective sample-level precision summary and explicitly distinguishes the two R^2 definitions used in the study.

The summary is informed by Archer et al. [A9] but is limited to three retrospective sample-level precision quantities. It does not implement all four of their planning criteria, does not use prospectively registered clinical precision margins, and does not justify a sample-size adequacy claim. Normal-approximation intervals use residual-based standard errors for CITL and slope and a Wishart-based approximation for the variance-ratio R^2. They are not design-based survey intervals, simultaneous calibration bands, or confidence intervals for clinical utility.

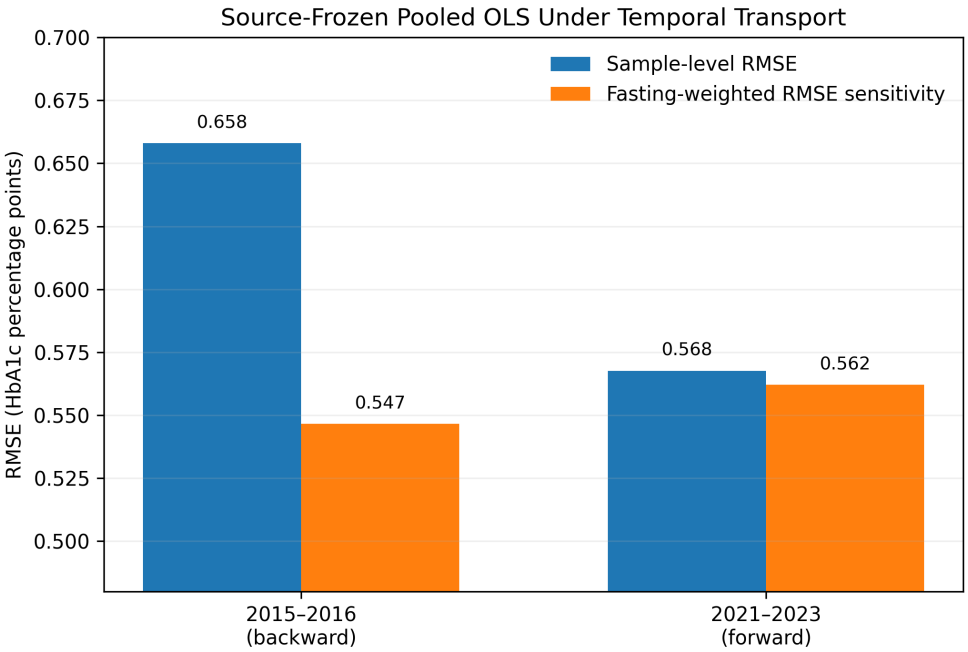


Figure A4. Sample-level and fasting-weighted RMSE for the frozen pooled OLS under backward and forward temporal transport. Weighting changes the estimand; it does not create a population conformal guarantee.

Cycle / N	Variance-ratio R^2 (95% CI)	CITL (95% CI)	Linear slope (95% CI)
2015–2016 / 2,235	0.7336 [0.7147, 0.7525]	+0.0429 [+0.0157, +0.0702]	1.0421 [1.0161, 1.0681]
2021–2023 / 2,808	0.7126 [0.6947, 0.7306]	-0.0182 [-0.0392, +0.0028]	1.0473 [1.0228, 1.0718]

Table A14. Descriptive precision summary under sample-level approximations. The residual-variance-ratio definition is explicit; a complete Archer planning calculation and population-design inference are not claimed.

In the equations below, y_i is observed HbA1c, $\hat{y}_i$ is predicted HbA1c, and $\bar{y}$ is the sample mean of observed HbA1c.

Two R^2 definitions must be distinguished. The prediction-error definition is

$$R^2_{\text{pred}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{A-R2-1}$$

which gives 0.732433 and 0.712345 in the backward and forward cohorts. The variance-ratio definition used in the retrospective precision approximation is

$$R^2_{\text{var}} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}, \tag{A-R2-2}$$

which gives 0.733572 and 0.712641. Here $\text{Var}(\cdot)$ denotes sample variance. Centering the residual removes the mean-bias contribution, so these quantities need not be equal. The approximation in Table A14 pertains to the latter, not to prediction-error R^2.

A9a. Corrected source distributions and out-of-fold prediction display

Figures A5 and A6 provide descriptive source-cohort and out-of-fold prediction displays.

The source graphics use the corrected N=2,219 cohort and the original pooled-OLS out-of-fold predictions. Figure A6 uses only the first predefined repeat to display one held-out prediction per participant; primary metrics continue to use all five repeats. The repeat was not selected for its performance, and predictions were not averaged across repeats to create a new model.

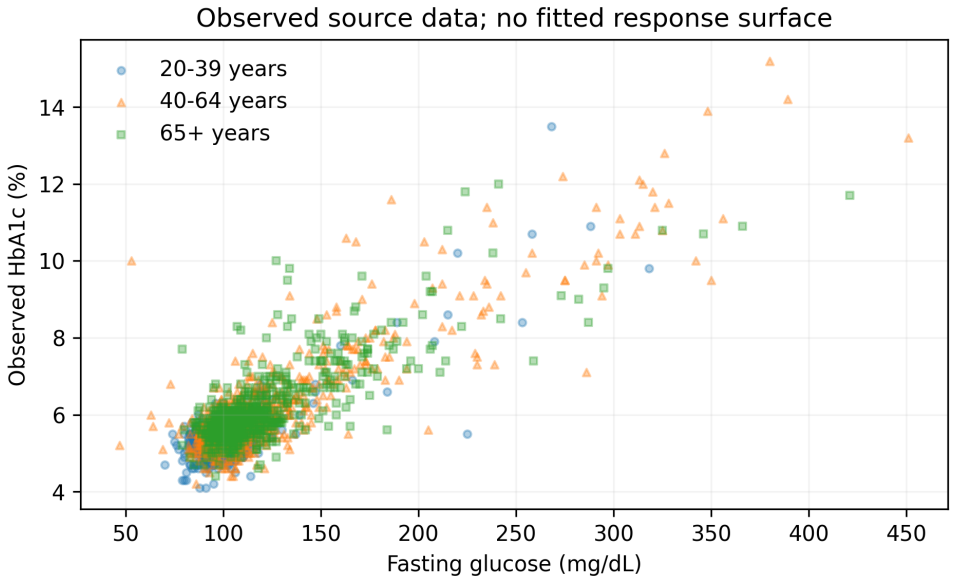


Figure A5. Fasting glucose and measured HbA1c in the corrected source cohort, shown by age. Points are observations; no clinical response function or causal relation is inferred.

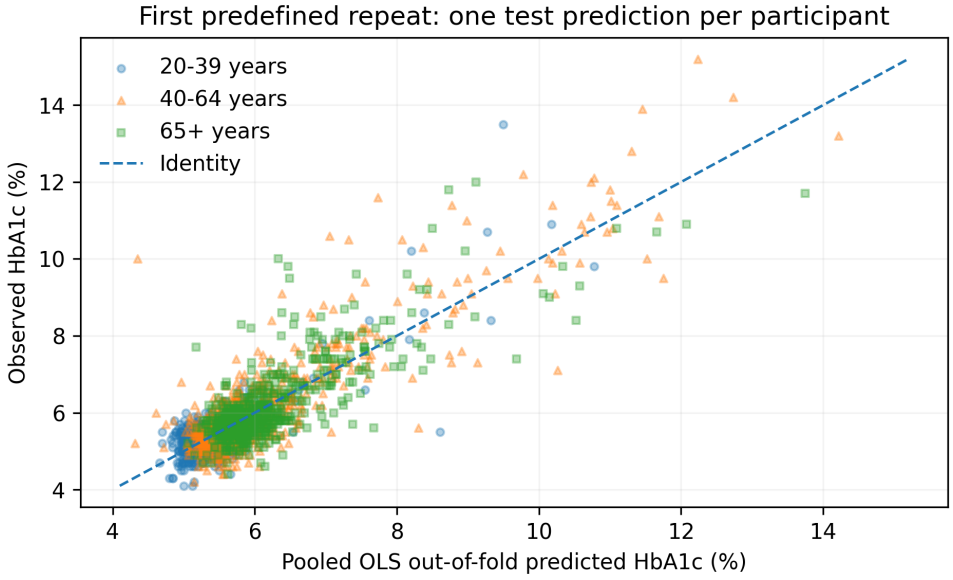


Figure A6. Pooled-OLS out-of-fold predictions in the first predefined repeat. Each of the 2,219 participants appears once. The identity line is a reference, not a fitted calibration curve. All reported primary model comparisons still use five repeats.

A10. Legacy adaptive-network-based fuzzy inference system (ANFIS) reconstruction: retained as a reproducibility case only

Table A15 summarizes the historical and corrected-evaluation ANFIS context.

The historical subtractive-clustering Sugeno fuzzy inference system (FIS) models trained with ANFIS are not part of the primary point-model comparison. Their checking rows were contained in the training data, and their age groups used different clustering radii. The full-data curves therefore characterize optimization and model response, not independent performance. Rule counts cannot be treated as evidence of biological complexity [A10,A11]. Table A15 preserves the separate corrected non-overlapping training/test fixed-split evaluation; those test metrics are not replaced by the training-fit values below.

Age group	Legacy radius / rules	Corrected selected radius / rules	Non-overlapping fixed-split RMSE
20–39	0.35 / 5	0.60 / 3	0.3429
40–64	0.17 / 23 (legacy reported)	0.40 / 4	0.5249
65–80 legacy grouping	0.20 / 13 (legacy reported)	0.25 / 12	0.6489

Table A15. Legacy ANFIS context on the historical N=2,325 table. The original and corrected-evaluation rule counts are not directly comparable as biological complexity measures.

A10a. Mathematical specification and one parameter-complete rule

Table A15a gives one parameter-complete archived premise and its local consequent.

For predictor vector $x = (x_1, x_2, x_3, x_4)$, the coordinates are age (years), BMI (kg/m^2), fasting glucose (mg/dL), and TG (mg/dL). A legacy system has K rules. In rule k, c_{qk} and $\sigma_{qk} > 0$ are the center and spread of the Gaussian premise for input q. Membership degrees determine the rule firing strengths, which act as model weights; neither should be interpreted as probabilities or conformal prediction limits [A10,A11]. The quantity $\mu_{qk}(x_q)$ denotes the Gaussian membership degree of input q in rule k.

$$\mu_{qk}(x_q) = \exp\left[-\frac{(x_q - c_{qk})^2}{2\sigma_{qk}^2}\right], \tag{A-ANFIS1}$$

$$w_k(x) = \prod_{q=1}^{4} \mu_{qk}(x_q), \tag{A-ANFIS2}$$

$$f_k(x) = p_{0k} + \sum_{q=1}^{4} p_{qk} x_q, \tag{A-ANFIS3}$$

$$\widehat{y}(x) = \frac{\sum_{k=1}^{K} w_k(x) f_k(x)}{\sum_{k=1}^{K} w_k(x)}. \tag{A-ANFIS4}$$

Here p_{0k} is the intercept and p_{qk} the coefficient of input q in the local consequent f_k. The quantity $w_k(x)$ in Eq. (A-ANFIS2) is the unnormalized firing strength of rule k; normalization occurs only through the denominator in Eq. (A-ANFIS4). A rule receives more influence when the observed predictors have higher membership in its Gaussian premise region, and the final prediction is the normalized weighted average of the local rule outputs. All archived rules have unit rule weight and identity-linked premise/consequent indices. The 5-, 23-, and 13-rule files contain 65, 299, and 169 premise/consequent numbers, respectively. All 533 match the historical printed parameter tables after rounding to six decimals. Evaluating the unchanged FIS files on the 2,325 stored input rows agrees with saved MATLAB predictions to a maximum absolute difference of 1.47×10^{-12}. This verifies the archived fixed functions, not independent generalization or a new MATLAB training run.

Young-adult rule 2 premise	σ	c
Age	4.024732	32.499575
BMI	8.823817	28.704549
Fasting glucose	23.183674	95.607254
TG	46.247643	62.638571

Table A15a. One parameter-complete archived premise. Its local consequent is $f_2(x) = 4.600227 + 0.006095\,\text{Age} + 0.004644\,\text{BMI} + 0.003109\,\text{Glucose} + 0.000573\,\text{TG}$. *This is one local rule, not the complete model prediction. Full-precision coefficients for all 41 rules and original FIS files are retained in verification materials held by the authors.*

A10b. Training selection, internal scaling, and graphical interpretation

Table A15b records the historical training settings, while Figures A7–A9 document the overlapping checking trace, one membership-function example, and one fully declared response slice.

MATLAB returns the first ANFIS output at the minimum training-error epoch, not necessarily the last epoch [A12]. The archived middle-age run has minimum training RMSE 0.449178907 at epoch 999, versus 0.449480111 at epoch 1000; the saved FIS reproduces the former. Training/checking containment is 140/140, 204/204, and 121/121 rows across the historical groups. No checking trace is interpreted as independent validation.

Historical ANFIS setting	Value
Learning procedure	MATLAB hybrid ANFIS
Epochs	1000
Squash factor	1.25
Acceptance ratio	0.50
Rejection ratio	0.15
Radius, 20–39 years	0.35
Radius, 40–64 years	0.17
Radius, 65–80 years	0.20
Checking selection	Periodic subset (every fifth observation) reused from training

Table A15b. Historical adaptive-network-based fuzzy inference system (ANFIS) settings recovered from the archived MATLAB workflow.

The archived workflow selected a periodic subset of observations as checking data while retaining the complete subgroup in the training set; therefore, the checking observations were not independent of training.

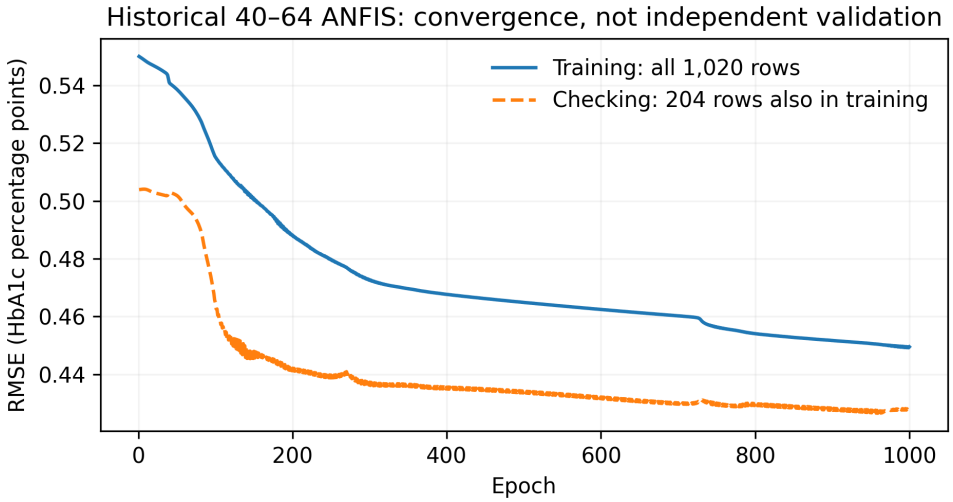


Figure A7. Historical 40–64 training and overlapping-checking traces, regenerated from stored MATLAB arrays. All 204 checking observations were also in the 1,020-row training set. The saved first-output FIS corresponds to minimum training error, reached at epoch 999.

The absence of an external z-score or min–max step in a script does not imply that clustering used raw-unit distances. The archived logs record internal normalization, and subtractive clustering scales data to a unit hyperbox [A13]. The archived age-category numeric arrays contain four inputs and HbA1c, not a stored SEQN column; respondent identifiers were recovered later by source matching. Figure A8 illustrates adaptation of one premise without equating a membership degree with predictive uncertainty.

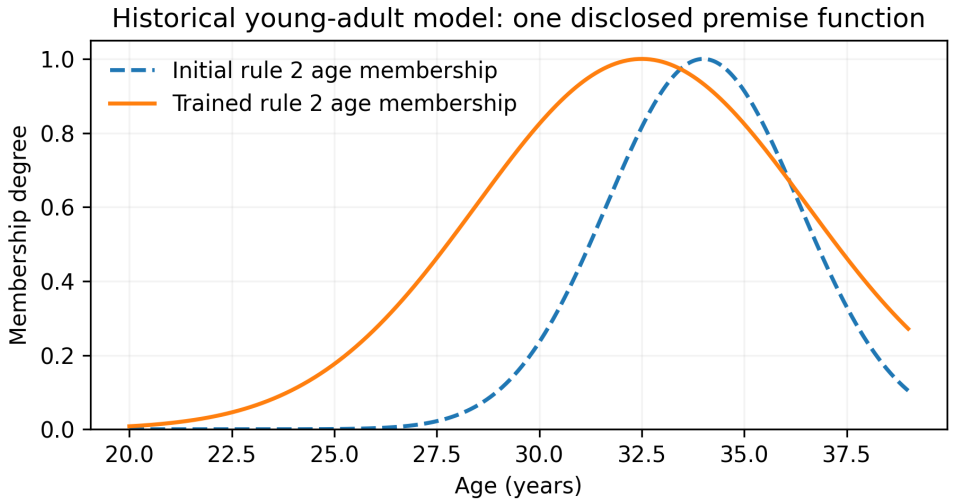


Figure A8. Initial and trained age membership functions for rule 2 of the historical young-adult FIS. Functions are regenerated from the original FIS files; this selected example is a structural illustration, not an uncertainty interval or evidence of age-specific biological mechanisms.

For a documented exploratory diagnostic added after the original analyses, the unchanged middle-age FIS was evaluated on a 101×101 age–TG grid over its stored ranges, holding BMI at 45.15 and glucose at 249. These are midranges, consistent with the documented default reference convention for gensurf [A14], not cohort means. The slice is fully specified here; it is not asserted to reproduce unrecorded historical plotting arguments. Its negative or extreme outputs are fixed-model responses at the declared combinations, not

plausible patient values or prevalence estimates. Marginal input bounds do not establish joint four-dimensional data support.

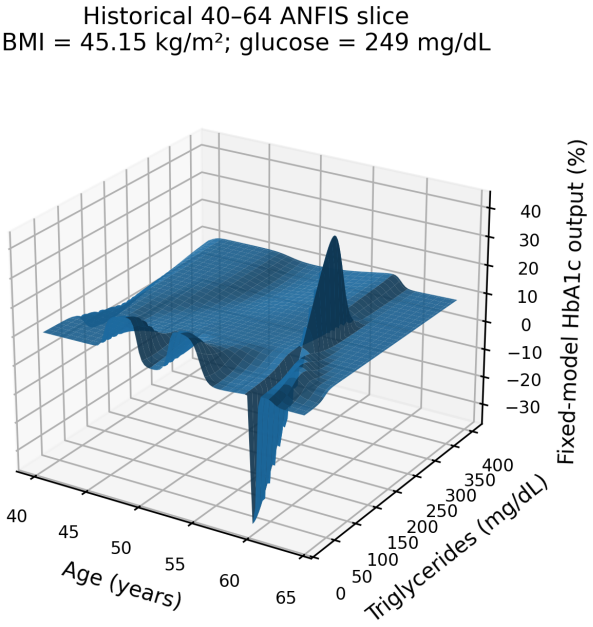


Figure A9. Declared historical-FIS age–TG response slice at BMI=45.15 kg/m^2 and fasting glucose=249 mg/dL. The full computed output range is retained without clipping. The plot describes a fixed model on a rectangular grid, not a causal physiological surface, a representative patient, or an independent performance test.

A10c. Continuity with the earlier unpublished ANFIS analysis

The earlier ANFIS work was not published. It was an internal analysis based on the same legacy dataset and archived ANFIS model files. The present paper changes its inferential role without discarding technically valid structural evidence. Table A15c records what is preserved, what is shown selectively in the Supplement, and which earlier interpretations are no longer supported by the corrected evaluation. This distinction prevents structural reproducibility from being confused with independent predictive validity.

Table A15c. Continuity of the earlier unpublished ANFIS analysis under the corrected reliability study.

Original contribution	Status in the current study
Three age-specific Sugeno ANFIS models	Preserved as archived fixed-model objects and computationally verified.
41 fuzzy rules / 533 stored premise and consequent parameters	All 533 values match the historical printed tables after six-decimal rounding; original FIS files remain authoritative, with one parameter-complete rule displayed here.
Membership-function evolution	Preserved selectively through a representative premise example; the full historical panels remain in verification materials held by the authors.
Response surfaces	Reduced to one fully declared diagnostic slice with fixed reference values and explicit interpretation limits; no physiological mechanism is inferred.
ANFIS convergence behavior	Preserved as a legacy optimization diagnostic, with the training/checking overlap stated explicitly.
Seven-model historical benchmark	Not included in the current numerical comparison because the corrected cohort and evaluation protocol differ; historical values are not mixed with the repeated-OOF evidence.
ANFIS superiority claim	Not retained as a supported claim after the corrected non-overlapping evaluation.
Biological interpretation of rule count	Not retained as a supported interpretation because rule count is strongly affected by clustering radius and selection procedure.

A11. TRIPOD+AI reporting map

Table A16 maps the main reporting domains to their locations in the manuscript and Supplement.

Reporting domain	How addressed	Location
Title/abstract	Prediction-model evaluation and temporal transport summarized; model classes specified in Methods.	Main title; Abstract; Methods 2.2
Background/objectives	Clinical context, reliability gap, and study objectives stated.	Introduction
Data source/setting	NHANES cycles, same-program temporal role, and public-use status reported.	Methods 2.1
Participants	Eligibility, fasting rule, age boundary, and cohort counts reported.	Methods 2.1; Table 1; A1
Outcome	Same-examination measured HbA1c (%) defined; no longitudinal interpretation.	Methods 2.2
Predictors	Age, BMI, fasting glucose, TG; excluded predictors listed.	Methods 2.2
Sample size / precision	Retrospective CITL/slope and explicitly defined variance-ratio R^2 intervals; no complete sample-size certificate claimed.	A9; Table A14
Missing data	Primary cohort requires complete core variables; no target imputation introduced.	Methods 2.1
Data separation	Outer test, calibration-only, and model-training roles explicitly separated.	Methods 2.3; A2–A2a
Model specification	Six point models and LightGBM grid specified; final temporal OLS coefficients and hashes released.	Methods 2.2; A2–A2a
Performance measures	RMSE primary; MAE, R^2, bias, calibration, interval coverage/width/score secondary.	Methods 2.3–2.5
Internal validation	Repeated out-of-fold design and participant bootstrap reported.	Methods 2.3

Reporting domain	How addressed	Location
Temporal evaluation	No target refitting; backward and forward NHANES cycles distinguished from independent external validation.	Methods 2.5; Results 3.4–3.5
Model updating	None in primary temporal tests.	Methods 2.5; A2a
Subgroups	Inherited age groups; subgroup coverage and errors reported.	Results 3.3–3.5
Uncertainty	Global/Mondrian primary; original CQR secondary and distinct post hoc CQR exploratory sensitivity.	Methods 2.4; A2b; A4
Limitations	Data scope, same-program transport, post-primary diagnostics, and conformal boundaries reported.	Limitations
Availability	Public NHANES source and reproducibility archive described.	Data Availability Statement

Table A16. Concise TRIPOD+AI-oriented reporting map. Wording is a study-specific paraphrase rather than a reproduction of the guideline checklist.

A12. PROBAST+AI-oriented self-assessment

Table A17 summarizes the study-specific PROBAST+AI-oriented self-assessment used to bound interpretation.

Domain	Self-assessment conclusion
Participants / data source	Low concern for source identity after exact SEQN reconstruction; applicability limited to NHANES-like adult populations and same-examination prediction.
Predictors	Low concern for availability within the analytic dataset; clinical applicability is limited because only four predictors are modeled and laboratory glucose/TG are required.
Outcome	Measured HbA1c from the same examination; clear and consistently defined. This is concurrent estimation, not future-outcome prognosis.

Domain	Self-assessment conclusion
Analysis	Leakage controls are explicitly implemented in the current analysis: preprocessing/tuning confined to training data, calibration-only role separated, participant-level resampling used. Original CQR metadata and outputs were recovered and verified; its secondary status and the distinct post hoc exploratory status are retained.
Transportability	Moderate limitation: backward and forward tests are within NHANES, not independent health systems.
Survey design	Population estimands remain sensitivity analyses; no survey-population conformal guarantee is claimed.
Overall	The current analysis substantially lowers the principal legacy-workflow risks of leakage and uncertain provenance, but clinical deployment applicability remains limited.

Table A17. PROBAST+AI-oriented self-assessment used to bound claims; this is not an independent risk-of-bias assessment.

A13. Claim-to-evidence boundary

Table A18 states the supported claims and the principal boundaries of inference.

Claim	Status	Evidence / boundary
Legacy data source is NHANES 2017–2018	Supported	2,325/2,325 exact unique row-level reconstruction.
LDL restriction explains 2,350→2,325	Supported	25 missing-LDL rows; all TG $>$ 400 mg/dL.
Hard age stratification improves HbA1c point prediction	Not supported	Hard-age OLS CI crosses zero; hard-age LightGBM is worse.
Reliability calibration differs by age	Supported empirically	Age-specific global interval coverage differs internally and in both target cycles.
Age-Mondrian guarantees nominal subgroup coverage	Not supported	2015–2016 middle group remains below 90%.
Post hoc exploratory CQR proves superior adaptive intervals	Not supported	The post hoc CQR analysis is exploratory and does not alter the primary manuscript conclusions.
Model is externally validated in another health system	Not supported	Both temporal tests are within NHANES.

Claim	Status	Evidence / boundary
Model can replace laboratory HbA1c or guide treatment	Not supported	No clinical utility or decision-impact study.
Survey-population conformal coverage is guaranteed	Not supported	Weighted metrics are sensitivities; no design-based conformal guarantee claimed.

Table A18. Claim-to-evidence map.

A14. Reproducibility artifacts

Additional verification materials retained by the authors include the source-linked cohort and provenance materials, analysis scripts, participant-level predictions, frozen temporal OLS/conformal objects, temporal CQR outputs, and the original FIS/MAT/log files. These participant-level and author-retained legacy materials are not redistributed in the public repository. The public reproducibility repository instead provides source-reconstruction certificates, aggregate derived outputs, frozen model/calibration identity records, the assay/source ledger, aggregate legacy verification summaries, and verification scripts. These materials support the reported reproducibility checks without changing the scientific results.

Figure A10 provides the retained full-range smooth calibration curves for the two temporal cohorts.

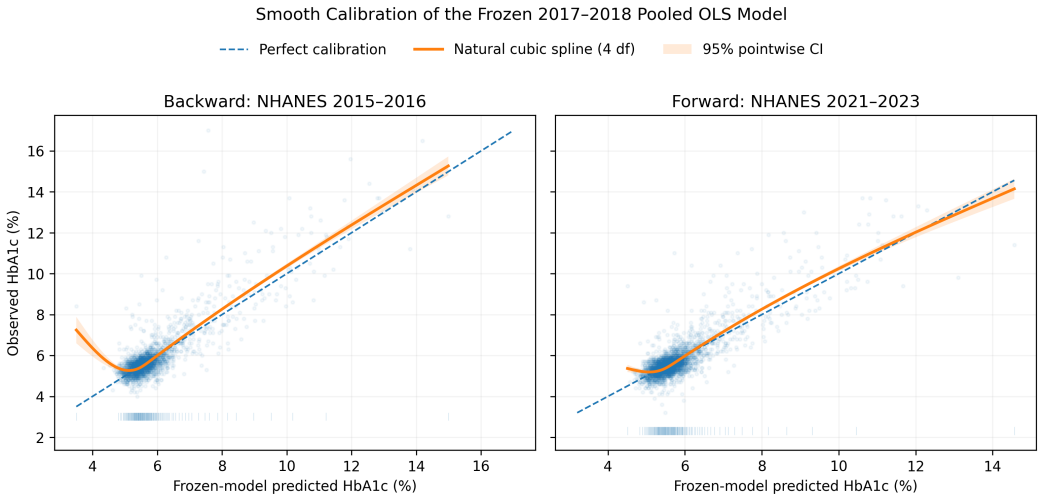


Figure A10. Retained full-range smooth calibration curves for frozen pooled OLS in the two NHANES temporal cohorts. These descriptive curves do not update the frozen model.

References for Supplementary Technical Section A

A1. National Center for Health Statistics. NHANES 2017–2018 Data Documentation, Codebook, and Frequencies: Demographic Variables and Sample Weights (DEMO_J). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DEMO_J.htm. Accessed 13 September 2026.

A2. National Center for Health Statistics. NHANES 2017–2018 Data Documentation, Codebook, and Frequencies: Cholesterol - Low-Density Lipoprotein (LDL) & Triglycerides (TRIGLY_J). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/TRIGLY_J.htm. Accessed 13 September 2026.

A3. National Center for Health Statistics. NHANES 2017–2018 Data Documentation, Codebook, and Frequencies: Plasma Fasting Glucose (GLU_J). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/GLU_J.htm. Accessed 13 September 2026.

A4. Romano Y, Patterson E, Candès EJ. Conformalized Quantile Regression. Advances in Neural Information Processing Systems. 2019;32:3543–3553.

A5. National Center for Health Statistics. NHANES 2015–2016 Data Documentation, Codebook, and Frequencies: Plasma Fasting Glucose (GLU_I). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2015/DataFiles/GLU_I.htm. Accessed 13 September 2026.

A6. National Center for Health Statistics. NHANES August 2021–August 2023 Data Documentation, Codebook, and Frequencies: Cholesterol - Low-Density Lipoproteins (LDL) & Triglycerides (TRIGLY_L). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/TRIGLY_L.htm. Accessed 13 September 2026.

A7. National Center for Health Statistics. NHANES August 2021–August 2023 Data Documentation, Codebook, and Frequencies: Plasma Fasting Glucose (GLU_L). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/GLU_L.htm. Accessed 13 September 2026.

A8. National Center for Health Statistics. NHANES 2015–2016 Data Documentation, Codebook, and Frequencies: Cholesterol - Low-Density Lipoprotein (LDL) & Triglycerides (TRIGLY_I). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2015/DataFiles/TRIGLY_I.htm. Accessed 13 September 2026.

A9. Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. Statistics in Medicine. 2021;40(1):133–146. doi:10.1002/sim.8766.

A10. Jang JSR. ANFIS: Adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics. 1993;23(3):665–685. doi:10.1109/21.256541.

A11. Chiu SL. Fuzzy model identification based on cluster estimation. Journal of Intelligent & Fuzzy Systems. 1994;2(3):267–278. doi:10.3233/IFS-1994-2306.

A12. MathWorks. anfis: output arguments, trained fuzzy inference system. https://www.mathworks.com/help/fuzzy/anfis.html. Accessed 13 September 2026.

A13. MathWorks. subclust: data scaling for subtractive clustering. https://www.mathworks.com/help/fuzzy/subclust.html. Accessed 13 September 2026.

A14. MathWorks. gensurf: reference values for non-displayed inputs. https://www.mathworks.com/help/fuzzy/gensurf.html. Accessed 13 September 2026.