RESEARCH ARTICLE

Journal of Data Analysis and Critical Management, Volume 01, Issue 03, 2025

Computational Phenotyping of Unrecognized Prediabetes: A Reproducible Data-Analytics Development and Temporal Validation Protocol

Mahadi Hasan Khan[1*], Shima Ali Sadia[2], Eusha Mohtasim[3]

[1]MEng – Software Engineering for Embedded Systems, RPTU Kaiserslautern-Landau, Germany

[2]King Graduate School, Monroe University, United States

[3]Department of Computer Science, University of Liberal Arts Bangladesh, Bangladesh

Abstract

This protocol establishes a repeatable protocol for computational processing of public data to test the hypothesis that multiday wrist-accelerometry features can be used to distinguish prediabetes (based on multiple biomarker criteria) in adults not professionally diagnosed. Only deidentified NHANES 2011-2014 public use files on or before July 31, 2025, will be utilized in the study. These principal solutions are data analytic: deterministic wearable QA, predetermined circadian feature engineering, development-only preprocessing, regularized probability modeling, explicit leakage barriers, and frozen temporal validation. NHANES 2011-2012 is the development cycle and NHANES 2013-2014 is the untouched temporal-validation cycle. The Circadian Dysfunction Score (CDS) is the primary digital exposure, which is a combination of the lower relative amplitude, lower interdaily stability, higher intradaily variability and higher log transformed L5/rest-period activity. The main model is the ridge logistic regression model. Validation thresholds are not deployment thresholds, the primary incremental-validation estimand is the difference in specificity at 90% sensitivity between the nonlaboratory baseline model and the baseline-plus-CDS model at cycle 2013-2014. Separately, thresholds for cycle 90% and 95% sensitivity will be set at the development stage and then frozen and used without change across the rest of the cycle to measure the achieved sensitivity, specificity and referral burden. Secondary performance measures must include calibration intercept, calibration slope, Brier score, AUROC and precision-recall AUC. A separate survey-weighted epidemiologic layer will calculate prevalence and association for pooled 2011-2014 data based on the NHANES design variables and fasting-subsample weights. Prediabetes is defined by the thresholds set by the ADA: HbA1c and FPG, and prior recognition is defined by DIQ160. The wearable model is only a risk prioritization tool for biochemical screening and not a diagnostic substitute. There are no outcome estimates due to the fact that this manuscript is a protocol and a statistical analysis plan.

Keywords: Data analytics; digital phenotyping; wearable time series; feature engineering; temporal validation; calibration; prediabetes; NHANES; public-use data

Journal of Data Analysis and Critical Management (2025); DOI: 10.64235/4938nr36

Introduction

Prediabetes is characterized by blood sugar levels that are higher than normal but not high enough to be diagnosed as diabetes. According to the American Diabetes Association standards of care in diabetes-2025 the diagnosis of prediabetes in nonpregnant adults is based on HbA1c levels between 5.7% and 6.4%, fasting plasma glucose levels between 100 and 125 mg/dL, or 2-hour plasma glucose results between 140 and 199 mg/dL after an oral glucose tolerance test (OGTT) of 75 g (American Diabetes Association Professional Practice

Corresponding Author: Mahadi Hasan Khan, MEng – Software Engineering for Embedded Systems, e-mail: ankurmahadi@outlook.com

How to cite this article: Khan, M.H., Sadia, S.A., Mohtasim, E. (2025). Computational Phenotyping of Unrecognized Prediabetes: A Reproducible Data-Analytics Development and Temporal Validation Protocol. Journal of Data Analysis and Critical Management, 01(3):76-90.

Source of support: Nil

Conflict of interest: None

Committee, 2025). The thresholds are not used for autonomous diagnosis in the present protocol.

The clinical target is deliberately narrow to include only those who answer No to NHANES item DIQ160, who are participants who have undiagnosed prediabetes, as defined by biomarkers. DIQ160 indicates if a doctor or other health professional ever told the participant that they had prediabetes, impaired fasting glucose (IFG), impaired glucose tolerance (IGT), borderline diabetes, or blood glucose that is above normal but not in the diabetes range. A No response is considered as no prior reported recognition, not as lack of abnormal glycemic value being ever observed. Recognition may be a sign of care by some health professional, screening opportunity, recall, terms used, and previous glycemia changes as well as current metabolic control.

Analytic opportunity is due to the high dimensional temporal structure of wrist-monitor data. The minute level activity and ambient-light records can be converted into interpretable features of amplitude, regularity, fragmentation, phase and light-activity alignment. Rest-activity rhythmicity, amplitude, timing, and light-activity alignment were previously linked to glycemic markers or abnormal glucose metabolism in NHANES 2011-2014 (Xu et al., 2022; Xiao et al., 2022, 2023). A lower relative amplitude has also been linked to increased HbA1c in adults with prediabetes (Paewponsong et al., 2023). These studies provide biological and epidemiologic plausibility and do not provide the more limited computational question of incremental screening value in an unaltered temporal validation design.

This protocol therefore treats unrecognized prediabetes primarily as a data-analytics benchmark. The central tasks are to construct a deterministic wearable feature pipeline, prevent information leakage, compare nested probability models, freeze all learned transformations before temporal validation, quantify calibration as well as discrimination, and measure screening efficiency at high-sensitivity operating points. Pooled survey-weighted inference remains important but serves a supporting epidemiologic role rather than controlling the model-development workflow.

The primary predictive hypothesis is that a baseline-plus-CDS model will have higher specificity at 90% sensitivity than the prespecified nonlaboratory baseline model in NHANES 2013-2014. The primary association hypothesis is that higher CDS is associated with greater survey-weighted odds of unrecognized biomarker-defined prediabetes in pooled NHANES 2011-2014. All other model classes, individual wearable features, light-activity variables, subgroup analyses, continuous biomarkers, and alternative glycemic definitions are secondary or sensitivity analyses.

Methods

Study Design, Analytic Emphasis, and Evidence Cutoff

This is a retrospective secondary analysis protocol using only public-use NHANES 2011-2012 and 2013-2014 files. No restricted-use NCHS files, Research Data Center files, external electronic health records, proprietary wearable datasets, or controlled-access cohorts are permitted. The computational prediction stream and the survey-weighted epidemiologic stream are intentionally separated so that population inference does not silently alter model development or validation.

All literature, clinical guidance, data documentation, and methodological standards used to define the protocol must have been published or publicly available on or before July 31, 2025. Sources first published or materially updated after that date are ineligible to modify the protocol-defined design. For mutable web documentation, the analysis record will identify the dated document/version relied upon; later webpage changes will not be used to revise the documented evidence base.

The final report will follow STROBE for the observational epidemiologic component and TRIPOD+AI for the prediction-model component (von Elm et al., 2007; Collins et al., 2024). The protocol itself is designed as an auditable statistical analysis plan: every primary target, feature, model, metric, threshold rule, data split, and leakage barrier is specified below before outcome-linked model comparison.

Public Data Architecture and Provenance

NHANES uses a complex multistage probability design to represent the civilian, noninstitutionalized U.S. population. The 2011-2014 physical activity monitor component used wrist-worn ActiGraph GT3X+ devices and released public minute-level movement, ambient light, monitor-status, and quality information (National Center for Health Statistics, 2022a, 2022b). Public demographic, questionnaire, examination, and laboratory files are linked by respondent sequence number.

The analysis will maintain a machine-readable provenance manifest containing the NHANES cycle, public file name, source URL, retrieval date, file

checksum, variable names used, and transformation code version. Source files will be treated as immutable inputs. No manual spreadsheet editing of participant-level data is permitted. All cohort derivation and feature generation will be performed by scripted code from the original public files.

Cohort Construction and Target Definition

The primary analytic population comprises adults aged 20 years or older who attended the mobile examination center, belong to the applicable fasting laboratory subsample, have DIQ160 available, and meet wearable-validity requirements. Exclusions are pregnancy; self-reported established diabetes other than gestational diabetes; glucose-lowering treatment indicating established diabetes; diabetes-range HbA1c or fasting plasma glucose for the principal prediabetes endpoint; missing outcome or recognition status; and insufficient or irrecoverably invalid wearable data. Participant counts at every exclusion step will be reported separately for 2011-2012 and 2013-2014 before any model performance estimate is interpreted.

The primary prediction population is restricted to DIQ160 = No. Within this population, the positive class is biomarker-defined prediabetes and the negative class is normoglycemia under the same laboratory definition. This formulation prevents the model from being described as recognition-specific while actually comparing all prediabetes with all normoglycemia. Participants who are normoglycemic but report prior recognition are excluded from the primary prediction task and summarized separately.

Prediabetes will be defined as HbA1c 5.7%-6.4% and/or fasting plasma glucose 100-125 mg/dL. HbA1c >=6.5% or fasting plasma glucose >=126 mg/dL will be considered diabetes-range and excluded from the primary endpoint. Because NHANES laboratory measurements are single-occasion measurements, all labels are epidemiologic classifications. A secondary sensitivity endpoint will broaden the target to unrecognized dysglycemia, defined as prediabetes-range or diabetes-range biochemistry among participants without self-reported established diabetes; this sensitivity analysis addresses the practical screening concern that diabetes-range cases may also require referral.

Deterministic Wearable Quality-Control Pipeline

Quality control is part of the model specification, not a discretionary preprocessing step. The same code and rules will be applied to both cycles before any glycemic label is passed to the modeling layer.

Input granularity

Retain minute records from the public physical-activity-monitor files and the associated public monitor-status and quality fields. Original activity and light values will never be overwritten.

Valid minute

A minute is eligible only when public monitor-status/quality information indicates an analyzable recording and the required movement value is present. Minutes classified as invalid or nonwear by the public-use status

Table 1: Public-use data components and computational roles

Public file(s)	*Information*	*Role*
DEMO_G / DEMO_H	age, sex, race/ethnicity, education, poverty-income ratio, weights, strata, masked PSU	cohort definition; survey design; subgroup analysis
DIQ_G / DIQ_H	diabetes history; DIQ160; treatment indicators	recognition status; established-diabetes exclusions
GHB_G / GHB_H	HbA1c	primary and sensitivity outcome definitions
GLU_G / GLU_H	fasting plasma glucose; fasting-subsample weights	primary outcome; population weighting
BMX_G / BMX_H	BMI; waist circumference	baseline prediction; covariate adjustment
BPX_G / BPX_H; BPQ_G / BPQ_H	systolic blood pressure; antihypertensive treatment	baseline prediction; covariate adjustment
SMQ_G / SMQ_H	smoking variables	baseline prediction; covariate adjustment
PAXHD_G / PAXHD_H; PAXMIN_G / PAXMIN_H	device metadata; minute movement; ambient light; monitor status; quality fields	quality control; feature engineering

Note. Suffix G denotes NHANES 2011-2012 and suffix H denotes NHANES 2013-2014. Exact variable-level mappings will be documented in the version-controlled data dictionary used for the reported analysis.

variables are excluded. The primary pipeline will not interpolate or synthesize minute-level activity across invalid/nonwear intervals.

Valid day

A 24-hour day must contain at least 1,200 valid minutes (20 hours). The primary participant-level requirement is at least four valid days. Sensitivity analyses require at least five and at least six valid days.

Rolling-window coverage

M10 and L5 windows are evaluated only when at least 90% of the minutes within the candidate window are valid. If a valid day contains no qualifying 10-hour or 5-hour window, that day does not contribute to the relevant feature. A participant must retain at least four contributing days for the primary M10/L5-derived feature set.

Temporal alignment

Calculations use the clock-time and day structure supplied in the public minute file. No post hoc time-zone reconstruction from restricted geography is permitted. Clock-time variables are analyzed as circular quantities rather than ordinary linear values.

Quality diagnostics

The pipeline will output per participant the number of valid days, valid minutes/day, fraction of invalid minutes, number of usable M10/L5 windows, and feature-specific missingness. These diagnostics will be summarized before model fitting and retained in the audit log.

Feature Engineering and Circadian Representation

Feature construction is label-independent within the present workflow: glycemic labels are not available to the feature-extraction functions, and no feature is added or removed because of observed outcome association. Candidate features remain literature-informed and are therefore not claimed to be historically outcome-naive.

$$RA = (M10 - L5) / (M10 + L5)$$

The core Circadian Dysfunction Score (CDS) has exactly four components: -RA, -IS, +IV, and +log1p(L5 magnitude). Component orientation is fixed so that larger values represent less favorable temporal organization. In each training operation, component means and standard deviations are estimated from the training data only, standardized values are calculated, and the four standardized components are averaged with equal weights. No outcome-derived weights are allowed.

For the final temporal-validation model, transformation parameters and standardization constants are estimated once from the complete 2011-2012 development cycle after model tuning is finalized and are then applied unchanged to 2013-2014.

Table 2: Predefined primary and secondary wearable features

Feature	*Definition*	*Role*	*Predefined handling*
Relative amplitude (RA)	(M10-L5)/(M10+L5)	Primary CDS component; lower = adverse	No outcome-based tuning
Interdaily stability (IS)	day-to-day regularity of 24-h pattern	Primary CDS component; lower = adverse	Standard algorithm on valid observations
Intradaily variability (IV)	fragmentation of sequential activity pattern	Primary CDS component; higher = adverse	Standard algorithm on valid consecutive observations
L5 magnitude	mean activity in least-active valid 5-h window	Primary CDS component; higher = adverse	log1p before standardization
M10 magnitude	mean activity in most-active valid 10-h window	Secondary	log1p if right-skewed by prespecified rule
M10 midpoint / L5 midpoint	clock-time midpoints	Secondary timing	sine/cosine encoding
Cosinor mesor/amplitude/ acrophase	single-component 24-h cosinor	Secondary parametric	cosinor is the only primary parametric form; no model switching
Light-activity alignment	phase difference, phasor magnitude/angle	Secondary	excluded from primary CDS

Note. Feature formula implementations, edge-case rules, and unit tests will be stored with the analysis code. The extended-cosine alternative is removed from the primary plan to eliminate post hoc model switching.

Validation-cycle means or standard deviations are never used to rescale the CDS. This preserves the numerical meaning of the predictor across the development-validation boundary.

Timing variables are converted to radians and represented by paired sine and cosine terms. The primary parametric rhythm sensitivity analysis uses a conventional 24-hour cosinor only. Principal-components analysis, when reported, is secondary: centering, scaling, component number, and loadings are estimated in development data only and applied unchanged to temporal-validation data. No PCA component will replace the prespecified CDS because of apparent outcome performance.

Before outcome modeling, the pipeline will report distributions, missingness, pairwise correlations among CDS components, and the prevalence of extreme feature values. A leave-one-component-out CDS analysis and individual-component models will be sensitivity analyses. Feature values will not be winsorized in the primary pipeline; impossible values caused by computational failure will be treated as missing and traced to the corresponding quality-control log.

Prediction Task and Predefined Model Hierarchy

The computational estimand is individual-level probability of current unrecognized biomarker-defined prediabetes among eligible participants with DIQ160 = No. Prediction-model fitting is therefore distinct from the nationally representative prevalence estimand. The primary model comparison is nested and fixed in advance.

M1 includes both BMI and waist circumference because they provide complementary operational measures of adiposity; ridge regularization is used specifically to stabilize correlated predictors. A prespecified sensitivity analysis will compare BMI-only, waist-only, and BMI-plus-waist versions of M1. Coefficients are not interpreted causally.

Model Development, Tuning, and Data-Leakage Control

The primary algorithm is ridge logistic regression (L2-penalized logistic regression). Ridge, rather than feature-selecting lasso, is chosen because the primary objective is stable probability estimation under correlated nonlaboratory and circadian predictors. Elastic-net logistic regression is a secondary algorithmic sensitivity analysis; tree ensembles are not part of the prespecified primary protocol and will not be introduced because of favorable apparent performance.

NHANES 2011-2012 is the only development cycle. Hyperparameter tuning uses five-fold grouped cross-validation, with masked PSU as the grouping unit so participants from the same public survey cluster are not split across training and assessment folds. PSU groups will be assigned to folds using a deterministic seed and a greedy balancing algorithm based only on fold size and outcome count; the fold assignment itself will be stored. If the public design prevents a stable five-fold grouped split, the number of folds will be reduced to the largest value >=3 for which every assessment fold contains both outcome classes; this contingency is determined by class support, not by model performance.

The ridge penalty is selected by minimum mean cross-validated log loss. A second, stronger-regularization value within one standard error of the minimum will be reported as a stability sensitivity analysis, but it cannot replace the minimum-log-loss model as the primary specification. Continuous predictors are standardized inside each training fold; binary indicators are not standardized. Predictor imputation is learned inside each training fold as described below. Every learned preprocessing quantity is therefore estimated without access to the corresponding held-out observations.

Table 3: Prespecified prediction hierarchy

Model	*Predictors*	*Purpose*	*Status*
M0: Minimal	age + sex	Sanity/reference model	Secondary
M1: Nonlaboratory baseline	age + sex + BMI + waist circumference + systolic BP + antihypertensive treatment + smoking	Primary comparator	Primary
M2: Digital CDS	M1 + CDS	Primary digital model	Primary
M3: Expanded digital	M2 + prespecified timing / M10 / light features	Incremental feature sensitivity	Secondary

Note. Laboratory lipids and glycemic laboratory values are never predictors in the primary screening models. Race/ethnicity, education, and poverty-income ratio are used for descriptive/subgroup analyses and in epidemiologic adjustment; a sociodemographic-enriched prediction sensitivity model may add them but cannot replace M1 as the primary comparator.

After the penalty is selected, cross-validated out-of-fold development predictions are regenerated using the fixed primary penalty. These out-of-fold predictions are used to derive the development operating thresholds corresponding to at least 90% and at least 95% sensitivity. For each target, the selected threshold is the highest probability cutoff achieving the target sensitivity; ties are resolved in favor of the higher threshold. The numerical thresholds are recorded before 2013-2014 outcomes are evaluated.

The final ridge model is then refitted once on all eligible 2011-2012 development observations using the predefined preprocessing rules and selected penalty. All final imputation values, scaling constants, feature transformations, coefficients, intercept, and development-derived thresholds are serialized to a model object. Temporal validation loads this object; it is prohibited from estimating any preprocessing parameter or coefficient from 2013-2014.

Temporal Validation and Primary Performance Estimand

NHANES 2013-2014 is the untouched temporal-validation cycle. The model object produced from 2011-2012 will generate participant-level probabilities without refitting. The primary comparison is M2 versus M1.

The primary validation estimand is the paired difference in specificity at 90% sensitivity, ΔSpecificity90 = Specificity90(M2) - Specificity90(M1), estimated from the 2013-2014 validation ROC curves. This is a curve-based performance estimand: the validation outcomes are used only to estimate the operating characteristic at a common 90% sensitivity and not to create a deployable probability threshold. A positive value favors the CDS model. A 95% confidence interval will be obtained from 1,000 paired validation resamples that preserve the masked survey clustering by resampling public PSUs within strata where feasible; if a stratum-level resample is not computationally estimable, a PSU-cluster bootstrap that keeps all members of each selected PSU together will be used and labeled explicitly.

Mandatory secondary validation measures are Brier score, calibration intercept, calibration slope, AUROC, area under the precision-recall curve, sensitivity, specificity, positive predictive value, negative predictive value, and referral proportion. Differences between M2 and M1 will be estimated on the same validation observations. Calibration plots will use prespecified probability bins for display plus a smooth calibration curve; inference will rely on calibration intercept and slope rather than visual appearance.

A separate operational analysis will apply the frozen development-derived 90% and 95% sensitivity thresholds to 2013-2014 without modification. For each model and threshold, the final report will give achieved validation sensitivity, specificity, referral proportion, false positives per true positive, and number of biochemical tests per detected case. Because validation sensitivity may differ from the development target, these frozen-threshold metrics will not be mislabeled as specificity at exactly 90% or 95% sensitivity.

The primary predictive performance analysis is participant-level and unweighted because the model estimates conditional risk in the eligible analytic sample. A secondary policy-oriented analysis will recalculate key temporal-validation operating characteristics with the appropriate 2013-2014 fasting-subsample survey weights to describe expected performance under the later-cycle U.S. population distribution. Weighted PPV, NPV, calibration, and referral burden will be clearly distinguished from unweighted computational-validation metrics.

Table 4: Explicit leakage barriers

Potential leakage source	*Rule*	*Purpose*
Outcome label	Unavailable to feature-extraction functions	Prevents outcome-driven feature engineering
2013-2014 data	Not used for feature selection, scaling, imputation, tuning, coefficient estimation, or threshold selection	Preserves temporal validation
Scaling	Learned within training fold; final constants from 2011-2012 only	Prevents validation-distribution leakage
Imputation	Learned within training fold; final values from 2011-2012 only	Prevents validation-data leakage
PCA sensitivity	Loadings learned in development only	Prevents unsupervised leakage
Thresholds	Development OOF predictions only for frozen operational thresholds	Prevents outcome-guided thresholding
Model redesign	Primary CDS, M1, M2, ridge algorithm, and primary metric cannot be changed after outcome inspection	Prevents result-dependent specification

Missing Data Strategy

Outcome, DIQ160 recognition status, and failure to meet the wearable-validity criterion are not imputed. A participant lacking the primary outcome or primary CDS is excluded from the corresponding primary model.

For prediction, missing baseline predictor values are handled with a deterministic training-only pipeline: continuous predictors are median-imputed and categorical predictors are mode-imputed within each training fold. Final imputation values are estimated from the complete 2011-2012 development cycle and applied unchanged to 2013-2014. A secondary sensitivity analysis adds predictor-specific missingness indicators when a predictor has at least 1% missingness in development data. This simple strategy is chosen for deployability and to avoid introducing outcome-dependent multiple-imputation models into the prediction engine.

For survey-weighted association models, missing adjustment covariates will be handled by multiple imputation by chained equations as the primary approach. Fifty imputations will be generated using models that include the outcome, CDS, all adjustment variables, public strata, masked PSU identifiers where technically supported, and the sampling weight or an appropriate function of it. Survey-design estimates will be calculated within each imputed dataset and combined using Rubin rules. Complete-case estimates will be reported as a sensitivity analysis. The outcome, recognition status, and wearable exposure will not be imputed.

Sample Size and Model-Complexity Control

The dataset is fixed; therefore feasibility is determined by the eligible participant and event counts after the predefined cohort and wearable rules are applied. Before any performance comparison, the analysis will report development and validation sample size, number of positive outcomes, event fraction, number of candidate model parameters, and the distribution of events across masked PSUs.

The primary ridge models M1 and M2 have fixed predictor sets and remain the priority analyses. The adequacy of the expanded M3 and elastic-net sensitivity models will be evaluated against contemporary prediction-model sample-size principles rather than an unqualified events-per-variable rule (Riley et al., 2019). If the available information is inadequate for a secondary model, that model will be omitted and the omission documented; the primary ridge comparison will not be expanded to compensate for a weak result.

Survey-Weighted Epidemiologic Layer

The epidemiologic layer is intentionally separate from prediction development. Pooled NHANES 2011-2014 analyses will use the appropriate combined-cycle fasting-subsample weights, public strata, and masked PSU variables according to NCHS estimation guidance (Chen et al., 2018). Descriptive tables will report unweighted counts together with survey-weighted percentages or means and design-based 95% confidence intervals. Domain analyses will preserve the full survey design rather than delete non-domain observations before variance estimation.

The primary association model is survey-weighted logistic regression among participants with DIQ160 = No, comparing unrecognized biomarker-defined prediabetes with normoglycemia. The exposure is the pooled CDS standardized in the pooled analytic cohort and clearly labeled as distinct from the development-anchored CDS used for prediction. Model A adjusts for age, sex, and race/ethnicity. Model B is the predefined primary confounder-adjusted model and additionally includes education, poverty-income ratio, smoking, and alcohol use. Model C is an attenuation model adding BMI, waist circumference, systolic blood pressure, and antihypertensive treatment. Total activity and sleep-duration proxies are secondary attenuation variables, not part of the primary confounder model.

Nonlinearity of the CDS association will be examined with restricted cubic splines using knots at the 10th, 50th, and 90th percentiles of the pooled CDS distribution. Secondary biomarker analyses will examine continuous HbA1c, fasting glucose, fasting insulin, and HOMA-IR when fasting conditions support the calculation. HOMA-IR will be calculated as fasting insulin (microIU/mL) x fasting glucose (mg/dL) / 405 (Matthews et al., 1985).

HOMA-IR = fasting insulin (microIU/mL) x fasting glucose (mg/dL) / 405

Subgroup Performance, Error Analysis, and Robustness

Age, sex, and race/ethnicity subgroup analyses are secondary error analyses. For each subgroup, the final report will show sample size, event count, AUROC, Brier score, calibration intercept/slope where estimable, sensitivity, specificity, and false-negative rate. No subgroup comparison will be interpreted inferentially when either outcome class has fewer than 20 observations; such results will be labeled descriptive and uncertainty intervals will be emphasized.

Prespecified robustness analyses include: at least five and at least six valid wearable days; HbA1c-only and fasting-glucose-only outcomes; the broader unrecognized-dysglycemia endpoint; age >=35 years; weekday-only and weekend-only features; BMI-only and waist-only baseline variants; exclusion of ambient-light features; adjustment for total activity volume; individual CDS components; leave-one-component-out CDS; PCA sensitivity with development-only loadings; extreme timing-value checks; complete-case survey models; and weighted temporal-validation metrics.

The analysis will include structured error inspection without changing the primary model: distributions of predicted risk in false-negative and false-positive cases, data-quality summaries by error class, and comparison of wearable-validity characteristics between correctly and incorrectly classified participants. These analyses are diagnostic only and cannot be used to modify the predefined primary CDS or refit the validation model.

Reproducibility, Software Governance, and Version Control

All analysis stages will be scripted in R and/or Python: download verification, file linkage, cohort derivation, wearable quality control, feature extraction, model fitting, validation, survey analysis, and figure/table generation. Random seeds, package versions, operating-system information, and input-file checksums will be recorded. Derived data objects will be reproducible from public source files and code; manual correction of participant values is prohibited.

The protocol, code implementing the predefined cohort/outcome/CDS definitions, feature-unit tests, prediction model hierarchy, cross-validation folds, hyperparameter grid, primary metric, threshold-selection rule, missing-data rules, and validation script will be maintained under version control. The exact analytic version used to generate the reported results will be archived and identified in the final manuscript so that the analysis can be reconstructed from public source files. Any code or specification change that affects results will be documented in a deviation log with the original and revised behavior, reason, date, and whether the change occurred before or after outcome inspection. Changes made after outcome inspection that alter a primary definition, model, or metric will be labeled exploratory rather than silently incorporated into the primary analysis.

Automated assertions will verify that validation-cycle rows cannot enter training objects, validation-derived scaling constants are absent, model coefficients are unchanged during temporal validation, and the 2013-2014 outcome is never read by preprocessing functions. The validation script will terminate with an error if any prohibited refitting operation is detected. These checks convert data-leakage prevention from a narrative promise into an executable property of the pipeline.

Ethics, Privacy, and Data Governance

The analysis will use only deidentified NHANES public-use files released by the U.S. National Center for Health Statistics. The investigators will not receive names,

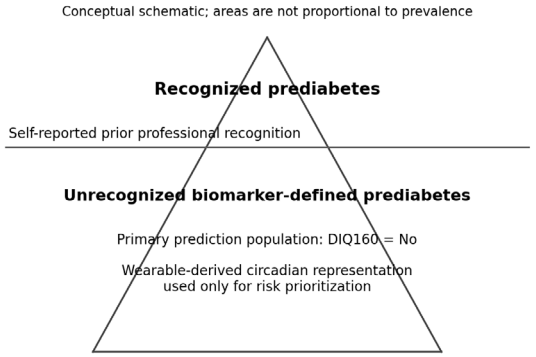
Table 5. Predefined analysis specification summary

Element	*Predefined specification*
Data	NHANES 2011-2012 development; NHANES 2013-2014 temporal validation; public-use only
Evidence cutoff	July 31, 2025; no post-cutoff source may change definitions or interpretation
Primary population	Adults >=20 years, fasting subsample, DIQ160 = No, valid wearable data
Primary outcome	HbA1c 5.7%-6.4% and/or FPG 100-125 mg/dL; diabetes-range excluded
Primary wearable exposure	Equal-weight CDS = standardized -RA, -IS, +IV, +log1p(L5)
Primary baseline	Age, sex, BMI, waist, systolic BP, antihypertensive treatment, smoking
Primary model	Ridge logistic regression
Primary validation comparison	M2 vs M1 difference in specificity at 90% sensitivity in 2013-2014
Operational thresholds	90% and 95% sensitivity thresholds derived from development OOF predictions; frozen
Primary weighting	Unweighted participant-level prediction; weighted policy metrics secondary
Association model	Survey-weighted pooled 2011-2014; Model B is primary confounder model
Primary-design change policy	Any change to CDS, M1, M2, ridge algorithm, primary metric, or temporal split after outcome inspection must be disclosed and treated as exploratory

addresses, direct identifiers, restricted geographic variables, linkage keys, or other restricted-use records; will not contact participants; and will not attempt reidentification or linkage to identifiable external information. Use of public-use files will follow the NCHS public-use data agreement applicable within the July 31, 2025 evidence cutoff (National Center for Health Statistics, 2024).

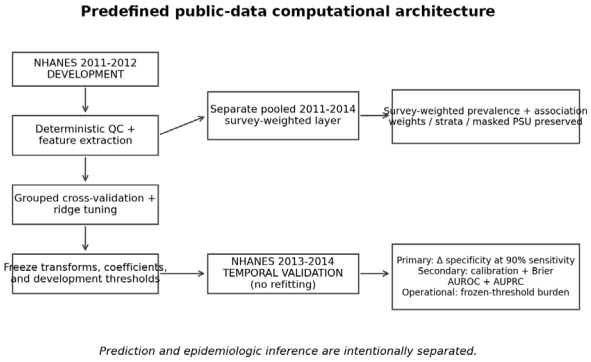
The original NHANES data collection was performed by NCHS under its participant-protection and informed-consent procedures. Publicly archived consent materials document the original participant-consent framework for the 2011-2012 and 2013-2014 cycles (National Center for Health Statistics, 2011, 2013a). The present investigators conduct secondary analysis only and have no participant interaction.

Under U.S. Office for Human Research Protections guidance, research does not involve human subjects under 45 CFR 46.102(e) when investigators do not obtain information through intervention or interaction and do not obtain, use, or generate identifiable private information that can readily be linked to individuals (Office for Human Research Protections, 2018). This protocol is designed to satisfy those conditions by using only public-use deidentified files. The investigators will nevertheless comply with any determination required by their responsible institutions. No institutional exemption, waiver, or approval will be claimed unless it has actually been issued, and the final manuscript will report the factual institutional determination and reference number if one is required.



Note. The primary prediction population is restricted to participants without self-reported prior professional recognition. The wearable representation is a prescreening signal, not a diagnosis.

Figure 1: Recognition-aware target definition



Note. The computational and survey-inference streams are separated. The 2013-2014 cycle is never used to learn preprocessing parameters or model coefficients.

Figure 2: Predefined development, leakage-control, temporal-validation, and epidemiologic architecture

Discussion

The protocol is designed to be heavily focused on data analysis and not on etiologic epidemiology. It is not its primary purpose to present another cross-sectional association between activity rhythms and glycemia since previous NHANES studies have shown that association. The main contribution consists of a repeatable computational experiment where the representation of the features in the data, the baseline model, the regularization method, the data split, the boundaries used in the preprocessing, the leakage checks, the metric used for validation and the rules used in the operating classification are fixed prior to comparison of the outcomes.

The temporal split is not a random train-only vs. test-only partitioning, but rather when all the final model learning takes place in the 2011-2012 NHANES cycle and the testing takes place in a later NHANES cycle. It remains a within-program validation: Sensing platform; survey context; processing environment of public data, within cycles. Successful temporal validation would then help with stability across calendar time within a common measurement ecosystem, rather than transportability to consumer watches or different devices, or clinical health systems.

The main prediction indicator, specificity at a 90% sensitivity, is selected because it is a problem that may be faced in a prescreening situation where the cost of a missed prediction is high but too many biochemical referrals have a detrimental effect on the efficiency. This distinction between the frozen operation thresholds and this curve based validation estimand is reflected in the manuscript. This separation prevents one of the typical ambiguities: making an estimate of the operating characteristic on the validation data does not mean the same as claiming to be able to use the threshold derived

from the validation data in a subsequent prospective setting. The second question is explicitly addressed in the separate frozen-threshold analysis, without refitting.

Calibration is not done for show-it-off purposes. If the score increases AUROC but the probabilities are unstable or the score is not very well calibrated over time, it is not yet ready for screening implementation. Similarly, the analyses of errors in sub-groups are designed to identify error instability, and not to give biological interpretation to the sub-groups. No reconstruction of the primary CDS will be reported if there are results that are null or imprecise or heterogeneous.

The role of the epidemiologic layer is more limited. In contrast, nationally representative prevalence and association estimates rely on survey weights, and secondarily, policy metrics are reported that reflect the individual-level conditional risk for the development of the prediction. This explicit separation helps to avoid confusing survey inference and predictive modeling as one and the same. A clinical interpretation is still conservative: the result will be a single-occasion biochemical classification; DIQ160 will be based on self-reported recognition; and actigraphy will be a behavioral index of temporal organization and not a true measure of the endogenous circadian phase.

Important limitations remain. Requires, and may have limited sample size, fasting laboratory participation, recognition information, and sufficient wrist-monitor data, so that there is selection that weighting does not fully correct. The two-cycle split can lead to sparse development data for complex algorithms, which is why the primary model is a low-variance ridge logistic model with a small CDS instead of a machine-learning search of the unrestricted model space. The CDS is transparent but a bit designed; individual-component and leave-one-component-out analyses will be performed to check for the domination of one component. Third, prospective clinical utility, causal effect and transportability are not possible within NHANES due to temporal validation.

The protocol is designed to be informative for multiple outcomes. If a positive ΔSpecificity90 and acceptable calibration is obtained, it might warrant a future prospective study of biochemical testing triggered by wearables. Circadian phenotyping would be an epidemiologic marker, but not a screening instrument when stable association, but negligible incremental prediction is found. The poor temporal calibration, or frozen-threshold sensitivity that is sensitive to change, or concentration of subgroups would be against deployment. If the primary outcome is null, then the null result is kept as the main outcome and is not used to re-engineer the feature score in a post hoc manner.

Conclusion

This public data only protocol outlines a pre-specified computational experiment to determine if a compact circadian representation helps screen-oriented prediction of unrecognized biomarkers defined as prediabetes. It is data-analytic: deterministic minute-to-minute quality control, outcome-independent feature construction, ridge probability modelling, grouped development resampling, explicit leakage barriers, frozen 2011-2012 preprocessing, and untouched 2013-2014 temporal validation. A survey-weighted epidemiologic inference is maintained outside of the predictive pipeline, and is used to give the population context. The design keeps the passive risk prioritization and biochemical diagnosis distinct and is reproducible based on only NHANES public-use files.

Data and Code Availability: Data and code are available upon request.

All participant-level inputs will be from the NHANES 2011-2012 and 2013-2014 Public-Use Files, available from the U.S. National Center for Health Statistics. There is no data with restricted use or controlled access that is required. The authors will publish analysis code, the variable dictionary, input-file checksums, feature-unit tests, and an archived analysis version identifier and scripts that can be used to recreate all tables and figures from the public source files at final publication. When redistribution is inconsistent with NCHS terms, files will not be repackaged from the source files provided in NHANES, but rather will be made available for deterministic download and verification.

References

Chen, T.-C., Parker, J. D., Clark, J., Shin, H.-C., Rammon, J. R., & Burt, V. L. (2018). National Health and Nutrition Examination Survey: Estimation procedures, 2011-2014. Vital and Health Statistics, Series 2(177), 1-26.

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

Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F., & Turner, R. C. (1985). Homeostasis model assessment: Insulin resistance and beta-cell function

from fasting plasma glucose and insulin concentrations in man. Diabetologia, 28(7), 412-419. https://doi.org/10.1007/BF00280883

National Center for Health Statistics. (2011). NHANES 2011-2012: Brochures and consent documents. https://wwwn.cdc.gov/Nchs/Nhanes/ContinuousNhanes/Documents.aspx?BeginYear=2011

National Center for Health Statistics. (2013a). NHANES 2013-2014: Brochures and consent documents. https://wwwn.cdc.gov/Nchs/Nhanes/ContinuousNhanes/Documents.aspx?BeginYear=2013

National Center for Health Statistics. (2013b). NHANES 2011-2012: Diabetes questionnaire (DIQ_G) [Data documentation]. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2011/DataFiles/DIQ_G.htm

National Center for Health Statistics. (2015). NHANES 2013-2014: Diabetes questionnaire (DIQ_H) [Data documentation]. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2013/DataFiles/DIQ_H.htm

National Center for Health Statistics. (2022a). NHANES 2011-2012: Physical activity monitor-minute (PAXMIN_G) [Data documentation]. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2011/DataFiles/PAXMIN_G.htm

National Center for Health Statistics. (2022b). NHANES 2013-2014: Physical activity monitor-minute (PAXMIN_H) [Data documentation]. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2013/DataFiles/PAXMIN_H.htm

National Center for Health Statistics. (2024). Data User Agreement for public-use NCHS data. https://www.cdc.gov/nchs/policy/data-user-agreement.html

Office for Human Research Protections. (2018). Coded private information or biospecimens used in research: Guidance. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/coded-private-information-or-biospecimens-used-research.html

Paewponsong, J., Gerber, B. S., Anothaisintawee, T., Chirakalwasan, N., Saetung, S., & Reutrakul, S. (2023). The association between rest-activity parameters and hemoglobin A1c in patients with prediabetes. Chronobiology International, 40(6), 834-839. https://doi.org/10.1080/07420528.2023.2215337

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. Statistics in Medicine, 38(7), 1276-1296. https://doi.org/10.1002/sim.7992

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. PLoS Medicine, 4(10), e296. https://doi.org/10.1371/journal.pmed.0040296

Xiao, Q., Durbin, J., Bauer, C., Yeung, C. H. C., & Figueiro, M. G. (2023). Alignment between 24-h light-dark and activity-rest rhythms is associated with diabetes and glucose metabolism in a nationally representative sample of American adults. Diabetes Care, 46(12), 2171-2179. https://doi.org/10.2337/dc23-1034

Xiao, Q., Matthews, C. E., Playdon, M., & Bauer, C. (2022). The association between rest-activity rhythms and glycemic markers: The U.S. National Health and Nutrition Examination Survey, 2011-2014. Sleep, 45(2), zsab291. https://doi.org/10.1093/sleep/zsab291

Xu, Y., Su, S., McCall, W. V., Isales, C., Snieder, H., & Wang, X. (2022). Rest-activity circadian rhythm and impaired glucose tolerance in adults: An analysis of NHANES 2011-2014. BMJ Open Diabetes Research & Care, 10(2), e002632. https://doi.org/10.1136/bmjdrc-2021-002632

Adepoju, S. A., & Adepoju, M. A. (2024). From Portals to Case Graphs: A Reference Architecture and Benchmark for Safety Investigation Operations with Agentic Orchestration.

Awodire, M. (2024). Zero-trust and least-privilege IAM design in federal/regulated cloud deployments. *ADHYAYAN: A JOURNAL OF MANAGEMENT SCIENCES*, *14*(02), 84-93.

Khan, H. A. (2024). DATA-DRIVEN EPIDEMIOLOGICAL MODELING USING MACHINE LEARNING FOR DISEASE SPREAD FORECASTING AND PUBLIC HEALTH DECISION SUPPORT IN THE UNITED STATES. *International Journal of Applied Mathematics*, *37*(6s), 178-192.

Areghan, E., & Ndibe, O. S. (2024). Explainable AI for autonomous threat detection in critical infrastructure systems. *Journal of Computational Analysis and Applications*, *33*(8), 6841-6857.

Moetiara, E. (2020). Risk assessment of airborne PM2. 5 exposure of forest and land fire haze to petrol-pump officers in Pekanbaru. *Acta Scientific Medical Sciences*, *4*(9), 152-157.

Adekoya, A. S. (2024). Enterprise Risk Compliance Architecture in Systemically Important Banks: Integrating Stress Testing, Capital Adequacy, and FX Exposure Modeling. *ADHYAYAN: A JOURNAL OF MANAGEMENT SCIENCES*, *14*(02), 66-74.

Anifowose, K. (2024). Development and Validation of an HPLC-Based Analytical Method for Simultaneous Quantification of Biomarkers in Human Biological Samples. *Well Testing Journal*, *33*(S2), 788-803.

Adekoya, A. S. (2023). Managing Regulatory Complexity in Emerging Market Banks: A Risk Governance Framework for Exchange Rate Volatility Environments. *ADHYAYAN: A JOURNAL OF MANAGEMENT SCIENCES*, *13*(02), 70-76.

Moetiara, E. (2023). Effectiveness of Integrated Occupational Health Protection Programs During Transboundary Haze Events: A Multi-Site Evaluation in the Oil and Gas Sector. *SRMS JOURNAL OF MEDICAL SCIENCE*, *8*(02), 161-166.

Taiwo, S. O., Aramide, O. O., & Tiamiyu, O. R. (2024). Explainable AI models for ensuring transparency in CPG markets pricing and promotions. *Journal of Computational Analysis and Applications*, *33*(8), 6858-6873.

Ojuri, M. A. (2024). Cybersecurity Vulnerability Testing as a Core Component of Quality Assurance. *Pioneer Research Journal of Computing Science*, *1*(3), 122-138.

Kola, J. N. (2024). Longitudinal Cohort Intelligence for Self-Insured Employer Groups: A Predictive Framework for Healthcare Cost Trajectory Modeling and Proactive Risk Intervention. *Journal of Information Systems Engineering & Management*, *10*.

Mehta, S. (2024). Architecting the Hub-and-Spoke Governance Model: Balancing Centralized Guardrails with Federated Domain Agility. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *16*(04), 206-215.

Ezeagwuna, D. (2024). Enterprise Workflow Automation and Workforce Productivity: Evaluating the Economic Benefits of Service Now Adoption Across Industries. *International Journal of Technology, Management and Humanities*, *10*(04), 299-313.

Ojuri, M. A. (2024). Integrating Cybersecurity Standards into Software Quality Assurance Frameworks: A Holistic Approach. *Journal of Computer Science and Technology Studies*, *6*(1), 258-271.

Areghan, E., & Onwuegbuchi, O. (2024). Predictive Cyber Threat Analysis in Cloud Platforms Using Artificial Intelligence and Machine Learning Algorithms. *Applied Science, Computing, and Energy*, *1*(1), 197-205.

Kola, J. N. (2024). Longitudinal Cohort Intelligence for Self-Insured Employer Groups: A Predictive Framework for Healthcare Cost Trajectory Modeling and Proactive Risk Intervention. *Journal of Information Systems Engineering & Management*, *10*.

Supplementary Appendix A

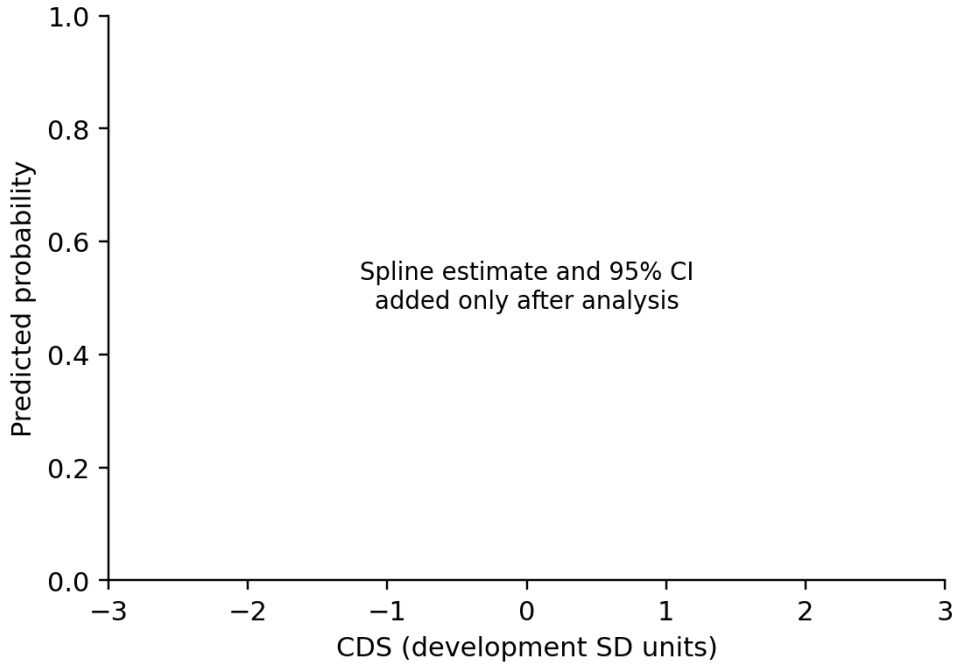
Computational Analysis Contract

A1. Development-cycle-only learning. Every parameter that can be learned from participant data - imputation values, scaling constants, ridge penalty, PCA loadings in sensitivity analyses, final coefficients, and frozen operating thresholds - must originate from NHANES 2011-2012.
A2. Validation immutability. NHANES 2013-2014 may be used only to create predictions and compute prespecified validation statistics. Any code path that calls fit, transform-fit, imputer.fit, scaler.fit, PCA.fit, or model.fit on a validation object is a protocol violation.
A3. Primary metric. The primary prediction comparison is M2 versus M1 difference in validation specificity at a common 90% sensitivity. This curve-based estimand is distinct from the separate frozen-threshold operating analysis.
A4. Model permanence. The primary ridge algorithm, M1 predictor list, four-component CDS, temporal split, and primary metric cannot be replaced because another analysis performs better.
A5. Result-independent feature QA. Feature debugging may correct coding errors only when the correction is justified by raw-data logic, unit tests, or documentation independent of glycemic outcome. Every such correction is versioned and logged.
A6. Primary association model. In pooled survey data, Model B (age, sex, race/ethnicity, education, poverty-income ratio, smoking, alcohol) is the primary confounder-adjusted association model. Model C adds adiposity and blood-pressure variables as attenuation analysis.
A7. Reporting of negative evidence. Null associations, no incremental specificity, poor calibration, or unstable subgroup results remain reportable outcomes and do not authorize post hoc reconstruction of the CDS.

Supplementary Appendix B

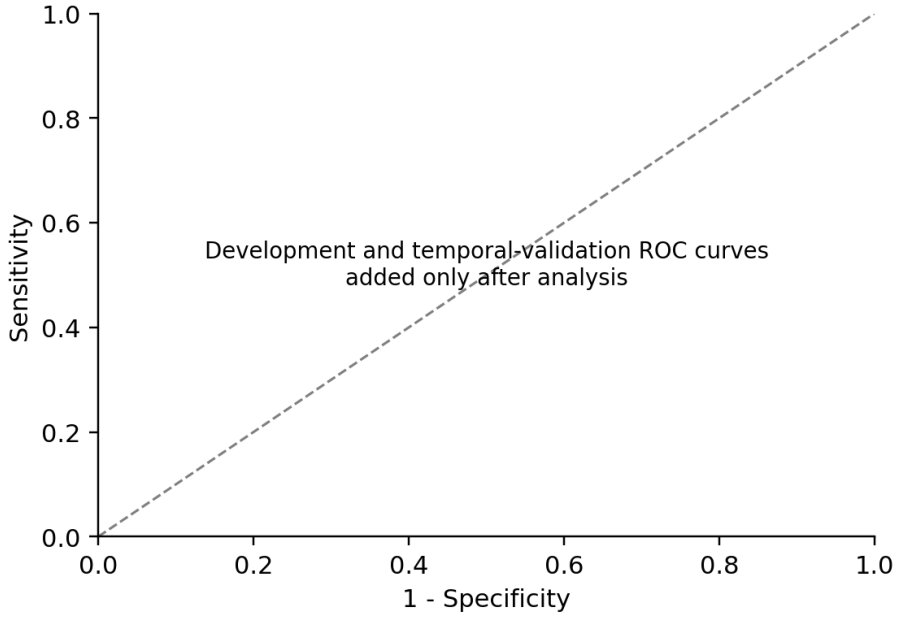
Planned Output Set

Output	*Prespecified content*
Participant flow	Unweighted counts by cycle and exclusion reason
Feature QA	Coverage, missingness, distributions, component correlations, QC failures
Development model	Coefficients, ridge penalty, development OOF metrics, frozen thresholds
Primary validation	Δ specificity at 90% sensitivity with 95% CI
Calibration	Intercept, slope, Brier score, calibration plot
Discrimination	AUROC and AUPRC with paired model differences
Operational screening	Frozen-threshold achieved sensitivity, specificity, referral fraction, FP/TP, tests per detected case
Survey epidemiology	Weighted prevalence and OR per 1-SD pooled CDS
Robustness	Alternative wear thresholds, glycemic definitions, baseline variants, CDS component analyses
Subgroup error analysis	Counts, calibration/discrimination where estimable, false-negative rates



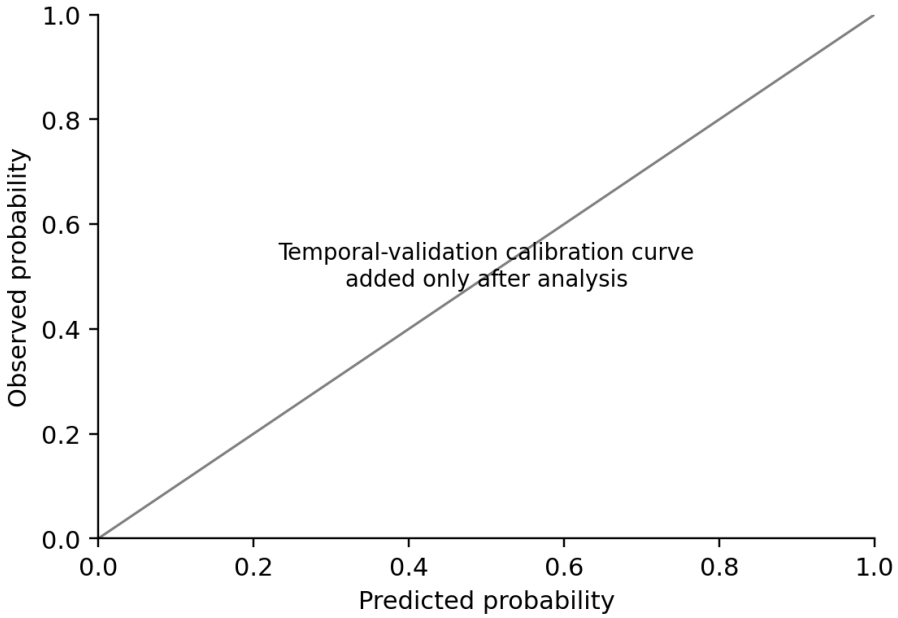
Note. Data-free display specification. The fitted association and confidence interval will be added only after the predefined model is executed.

Figure S2: Circadian Dysfunction Score spline display specification



Note. Data-free display specification. Development and temporal-validation curves will be added only after analysis.

Figure S3: Receiver operating characteristic display specification



Note. Data-free display specification. The diagonal represents perfect calibration; observed calibration will be added after validation.

Figure S4: Temporal-validation calibration display specification