Supplementary Methods

This supplement provides additional technical detail complementing the Methods section of the main manuscript. Definitions, cohort constructions, outcome thresholds, and figure/table conventions given in the main manuscript apply throughout and are not repeated here.

Software and Reproducibility

Analyses were conducted in Python 3.13 using scikit-learn, XGBoost, LightGBM, pandas, and numpy. Dependencies are pinned in `pyproject.toml` and `uv.lock`; the full pipeline runs from a single entry point (`uv run main.py`). A fixed random seed of 42 was used throughout for train/test splitting, cross-validation folds, and model estimators that accept a random state.

Machine-Learning Hyperparameters

The following hyperparameters were used for the primary models:

- Logistic regression: L2 penalty, solver `lbfgs`, maximum iterations 1,000.
- Random forest: 300 estimators, fixed random state.
- XGBoost: learning rate 0.05, maximum depth 5, 300 estimators.
- LightGBM: learning rate 0.05, 300 estimators.

The pooled MEC examination weight (2-year MEC examination weight divided by 10 for the pooled 1999–2018 period) was passed to each algorithm as sample weights at fit time. Weights were normalized to a mean of 1 within each training split to prevent weight-scale effects on optimization.

Variable Encoding and Missingness Handling

The family income-to-poverty ratio was discretized into tertiles before encoding. Categorical variables (sex, race and ethnicity, education, usual site of care, self-rated health) were one-hot encoded, and column names were sanitized to remove characters incompatible with gradient-boosting frameworks. The variable capturing healthcare visits during the previous year was excluded from the primary models because of high missingness (approximately 31% of the main cohort) and was included only in the pre-specified sensitivity analysis.

Health insurance status responses coded “Unknown” (29.9% of the main cohort, n = 9,750) were treated as missing and excluded from insurance-stratified analyses. All other key covariates (family income-to-poverty ratio, usual site of care, self-rated health, usual source of care) had no missing values in the analytic cohort.

Estimation Details

Confidence intervals for survey-weighted prevalences used a design-based Taylor-series linearization with the pooled MEC examination weight, primary sampling units, and strata, and t-distribution critical values with degrees of freedom equal to total primary sampling units minus total strata. A fallback to an effective-N binomial approximation was used for subgroups where primary sampling unit / stratum coverage was insufficient for a linearization estimate.

Disparity gaps were expressed as absolute differences (in percentage points) and prevalence ratios relative to Non-Hispanic White adults as the reference group.

Sensitivity Analyses

Three pre-specified sensitivity analyses were performed:

1. **Age restriction (18–45 years).** The main-cohort age ceiling was lowered from 55 to 45 years to reflect the primary-prevention window before cumulative risk accumulates.
2. **Alternative laboratory thresholds.** Two alternative thresholds were evaluated for Model B: (a) HbA1c threshold of 5.7% (pre-diabetes definition, more inclusive) and (b) LDL-C threshold of 160 mg/dL (more stringent).
3. **Healthcare visits during the previous year.** The count of healthcare visits was added to the Model A feature set. Because this variable was missing for approximately 31% of the main cohort, this sensitivity analysis was restricted to the subsample with non-missing responses (N = 22,615).

Supplementary Results

This supplement provides additional numeric detail complementing the Results section of the main manuscript. Point estimates and cohort characteristics reported in the main text are not repeated here.

Cross-Validation Performance

Five-fold stratified cross-validation on the training set (N_train = 26,053 for Model A; N_train = 24,844 for Model B) was consistent with held-out test-set estimates. For Model A, mean survey-weighted AUROC across folds was 0.709 ± 0.007 for XGBoost, 0.713 ± 0.007 for logistic regression, 0.709 ± 0.006 for LightGBM, and 0.664 ± 0.006 for random forest. Fold-level dispersion was small and did not indicate systematic overfitting.

Subgroup Discrimination

Subgroup survey-weighted AUROC values in the held-out test set are given in Table 3C of the main manuscript. For Model A, XGBoost discrimination was highest for the "Other" race and ethnicity group (0.759) and for Non-Hispanic Black adults (0.733), and lowest for Non-Hispanic White adults (0.696). For Model B, subgroup survey-weighted AUROC values ranged from 0.767 to 0.833. Subgroup test-set sizes ranged from 236 (Other, Model B) to 2,690 (Non-Hispanic White, Model A). Smaller subgroup test-set sizes are associated with wider confidence intervals; readers should interpret single-subgroup AUROC values within that context.

Sensitivity Analyses: Detail

- Age restriction to 18–45 years (N_test = 4,656): Model A XGBoost survey-weighted AUROC = 0.739. Subgroup survey-weighted AUROC values were 0.750 (Non-Hispanic Black) and 0.746 (Mexican American).
- Alternative HbA1c threshold of 5.7% (Model B): overall survey-weighted AUROC = 0.825.
- Alternative LDL-C threshold of 160 mg/dL (Model B): overall survey-weighted AUROC = 0.785.
- Adding healthcare visits during the previous year (N_test = 4,523; restricted subsample N = 22,615): overall survey-weighted AUROC = 0.705.

Insurance-Stratified Prevalence

Among uninsured participants, Non-Hispanic Black adults had a survey-weighted prevalence of poor control (main) of 49.8% versus 43.9% among Non-Hispanic White adults (prevalence ratio 1.14; absolute difference +6.0 percentage points). Among insured participants, the absolute disparity for Non-Hispanic Black adults narrowed to −0.7 percentage points; uninsured Non-Hispanic Black participants nonetheless had a higher survey-weighted prevalence than insured Non-Hispanic White participants.

Extended Limitations

The main-manuscript Limitations section lists the eight most salient limitations. The following section expands on those points, adds additional considerations that a reader may find relevant, and does not repeat the main-text paragraph verbatim.

Design and Inference

The analysis is cross-sectional and pools 10 two-year NHANES cycles. Temporal directionality between healthcare-access characteristics and risk-factor control cannot be established from these data. Point estimates and reported associations should not be read as causal effects, and no counterfactual claim is made about the effect of changing any predictor.

Measurement

Blood pressure was ascertained as the mean of multiple oscillometric readings within a single NHANES examination visit. Single-visit blood pressure is subject to regression-to-the-mean and does not correspond to a clinical diagnosis of hypertension, which typically requires repeated measurements or a documented antihypertensive medication history. Similarly, current smoking status was ascertained by self-report and may be subject to social-desirability bias. Laboratory values were measured by NHANES-standardized assays; assay methods have changed over the 1999–2018 period and small measurement drift is possible.

Missing Data

The health insurance status variable had substantial missingness (29.9% of the main cohort responded "Unknown"). This was treated as missing and excluded from insurance-stratified analyses. If the probability of an "Unknown" response is correlated with true insurance status or with the outcome, insurance-stratified estimates may be biased. Healthcare visits during the previous year had approximately 31% missingness in the main cohort and was included only in a pre-specified sensitivity analysis restricted to respondents with a non-missing value; that subsample is not fully representative of the main cohort.

Confounding and Unmeasured Factors

NHANES does not include direct measures of care quality (for example, guideline-concordant medication use for hypertension or dyslipidemia), patient preferences, adherence, structural racism, or neighborhood-level social determinants. Residual confounding by these unmeasured factors may contribute to the observed disparities. The composite outcomes were built from routinely measured variables and do not include medication use, ambulatory blood-pressure monitoring, or clinician diagnosis, so they represent an operational research definition rather than a validated clinical endpoint.

Temporal Heterogeneity

The pooling of 10 two-year cycles across 20 years increases sample size but may obscure temporal trends. Insurance coverage changed materially over this period (notably following implementation of the Affordable Care Act) and the guideline definitions of hypertension were updated in 2017. The main analysis does not model cycle-level heterogeneity beyond weighting; readers should not assume time-invariant associations.

Model Discrimination and Actionability

The moderate absolute survey-weighted AUROC for Model A (0.716) indicates that unmeasured clinical and contextual factors account for substantial residual variation in poor risk-factor control. Even for the higher-discrimination Model B (survey-weighted AUROC 0.822), individual-level positive-predictive value at any given operating threshold is meaningfully constrained by outcome prevalence. Individual-level clinical actionability is limited without independent external validation and prospective evaluation.

External Validity

Prevalence and disparity estimates apply to the U.S. non-institutionalized civilian population aged 18–55 years during 1999–2018. Extrapolation to institutionalized populations, other age bands, populations outside the United States, or the post-2018 period is not directly supported. Within-group heterogeneity by geography, immigration status, or socioeconomic subgroup is not captured by the analytic framework here.

Composite-Outcome Design

The composite outcome definitions ($\geq$ 1 of 2 for main; $\geq$ 2 of 4 for laboratory) were pre-specified before model fitting but are not validated clinical endpoints. Alternative composite thresholds could yield different prevalence estimates and different model discrimination. The sensitivity analyses with alternative HbA1c and LDL-C cut-points partly address this concern but do not exhaust it.