

CONFIDENCE WITHOUT CERTAINTY

Quantifying uncertainty in NHANES prediabetes prevalence

Technical report

Planimeter Kft.

Version 1.0.0

24 September 2026

Methodological demonstration only. This analysis uses public NHANES data to illustrate bootstrap and Bayesian uncertainty communication. It is not a survey-design-corrected national prevalence estimate and is not a clinical decision tool.

Executive summary

Objective. To show how an apparently simple prevalence estimate can be reframed as a distribution of plausible values, and how threshold probabilities can connect uncertainty to decision-oriented reporting.

Data. The released analysis contains 3,459 NHANES 2017-2018 participants aged 40 years or older with non-missing HbA1c and BMI. Prediabetes was defined for this demonstration as HbA1c $\geq 5.7\%$ and $< 6.5\%$; 1,304 participants met that definition.

Key result. The unweighted sample prevalence was 37.7%. The participant-level bootstrap 95% interval was 36.1%-39.3%. A Beta(1,1)-prior beta-binomial model produced a posterior median of 37.7% and a 95% credible interval of 36.1%-39.3%.

Interpretation. The two approaches tell the same broad story from different perspectives: the estimate is not a fixed fact but the center of an uncertainty distribution. In the Bayesian illustration, the posterior probability of prevalence exceeding 35% was 99.94%.

1. Data and analytical framework

1.1 Source data

The development workflow used public NHANES 2017-2018 component files accessed through the nhanesA R package: DEMO_J (demographics), GHB_J (glycohemoglobin) and BMX_J (body measures). The compact released CSV is sufficient to reproduce the published sample-based uncertainty analysis without repeated downloads.

1.2 Analysis population and endpoint

Participants were retained when age was at least 40 years and both HbA1c and BMI were available. Prediabetes was defined as HbA1c $\geq 5.7\%$ and $< 6.5\%$. The resulting analysis set contains 3,459 unique NHANES participant sequence numbers.

Characteristic	Value
Analysis sample	3,459 participants
Prediabetes cases	1,304
Median age	61 years
Median BMI	28.9 kg/m ²

1.3 Reproducibility

The public package contains the analysis-ready participant dataset, reference bootstrap and posterior draws, a reference summary, two R scripts, validation outputs and a data dictionary. All paths in the public scripts are relative. NHANES survey-design variables are retained in the released dataset for transparency even though the original Insight analysis is unweighted.

1.4 Statistical analysis

- Point estimate: the proportion of retained participants meeting the HbA1c-based prediabetes definition.
- Bootstrap: 3,000 participant-level nonparametric resamples, with the 2.5th and 97.5th percentiles used as an uncertainty interval.
- Bayesian illustration: a Beta(1,1) prior combined with y prediabetes cases among n participants, generating 10,000 draws from Beta(1+y, 1+n-y).
- Decision probabilities: posterior probabilities that prevalence exceeds 25%, 30% and 35%.
- Predictive visualisation: 400 bootstrap logistic models of prediabetes on BMI and age were used to show uncertainty around an illustrative fitted risk curve.

1.5 NHANES survey-design note

NHANES is a complex multistage probability sample. The released data retain WTMEC2YR, SDMVPSU and SDMVSTRA, but the original Insight deliberately used a simple unweighted analysis to make uncertainty concepts visually accessible. The reported prevalence therefore describes the analysed sample and should not be presented as a formal U.S. population prevalence estimate.

2. Results

2.1 One estimate becomes a distribution

Quantity	Result
Unweighted prevalence	37.7%
Bootstrap 95% interval	36.1%-39.3%
Posterior median	37.7%
Bayesian 95% credible interval	36.1%-39.3%

The bootstrap and beta-binomial posterior are not the same inferential procedure, yet in this large analysed sample they produce nearly identical central values and interval widths. The practical message is not the numerical coincidence itself; it is that uncertainty can be treated as an object of analysis rather than as a small annotation around a headline number.

2.2 From an estimate to threshold probabilities

Posterior question	Probability
Prevalence > 25%	100.0%
Prevalence > 30%	100.0%
Prevalence > 35%	99.94%

Threshold probabilities translate the posterior distribution into decision-oriented language. Their meaning is explicitly model-based: each probability is conditional on the beta-binomial model, the Beta(1,1) prior and the analysed sample.

2.3 Uncertainty around a fitted relationship

The final visualisation repeatedly refits a logistic regression of prediabetes on BMI and age in bootstrap samples. Instead of a single deterministic curve, the result is a family of plausible curves. This extends the same communication principle from a prevalence estimate to a fitted model: prediction itself carries uncertainty.

3. Interpretation

The central lesson is that confidence and certainty are different concepts. A point estimate can be precise enough to support a narrow uncertainty interval while still remaining an estimate. Bootstrap resampling expresses how the answer changes under repeated sampling from the observed data; the Bayesian model expresses a distribution over plausible prevalence values conditional on a prior and likelihood. Both make the uncertainty visible rather than implicit.

For decision support, the most useful question may therefore move from “What is the prevalence?” to “Which values remain plausible?” or “What is the model-based probability that a decision threshold is exceeded?” This shift can improve communication without claiming that uncertainty has disappeared.

4. Limitations

- The analysis is unweighted and does not incorporate the NHANES complex survey design. The 37.7% result is an analysed-sample prevalence, not a design-based national estimate.
- The complete-case rule requires non-missing HbA1c and BMI. The example does not model missingness or adjust weights for this restriction.
- The beta-binomial illustration assumes independent Bernoulli observations and does not represent NHANES clustering or stratification.
- Prediabetes is operationalised only through HbA1c in this example. Alternative clinical definitions can use additional measurements and would define a different estimand.
- Posterior probabilities depend on the chosen prior and likelihood. They should not be interpreted as model-free probabilities that a population fact is true.
- The BMI-age logistic model is an uncertainty visualisation, not a validated clinical prediction model, causal analysis or treatment decision rule.

5. Reproducibility package

The public package separates compact source-derived data, frozen reference outputs and executable analysis code. This makes it possible to reproduce the narrative without redistributing large files and also makes the provenance of the published figures explicit.

File	Purpose
nhanes_prediabetes_analysis.csv	Participant-level analysis data and retained NHANES design variables.
bootstrap_prevalence.csv	3,000 development-run bootstrap prevalence estimates.
posterior_prevalence.csv	10,000 development-run beta-binomial posterior draws.
confidence_summary.csv	Frozen numerical summary used for validation.
confidence_analysis_planimeter_v1.0.0.R	Main public analysis and static figure-generation script.
confidence_data_preparation_planimeter_v1.0.0.R	NHANES download and compact dataset construction.
validation_checks.csv	Independent recalculation of every published summary value.

6. License and AI use

The Planimeter R source code is released under the GNU General Public License v3.0 or later (GPL-3.0-or-later). Generative AI was used as a support tool during drafting, code development, review and validation. The analytical design, interpretation, final materials and responsibility for the release remain with Planimeter Kft.

Sources and methodological references

National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES), 2017-2018 public-use data.

Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC; 1993.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. CRC Press; 2013.

The public code and data dictionary supplied with this release define the exact computational workflow used for the reproducibility package.