Hellenic Complex Systems Laboratory

Probability and uncertainty estimation for a reference concentration threshold

Technical Report XXXI

Theodora Chatzimichail, Rallou A. Chatzimichail and Aristides T. Hatjimihail
2026

Probability and uncertainty estimation for a reference concentration threshold

Theodora Chatzimichail, MRCS [a], Rallou A. Chatzimichail, MEng, MSci, PhD [b], Aristides T. Hatjimihail, MD, PhD [c]

[a] tc@hcsl.com, [b] rc@hcsl.com, [c] ath@hcsl.com

[a, b, c] *Hellenic Complex Systems Laboratory*

DOI: 10.5281/zenodo.24445993

Abstract

Background: Quantitative diagnostic probabilities depend on uncertain population distributions and imperfect measurements. We developed *BDPU-Threshold* in the Wolfram Language to estimate these probabilities from paired test and reference measurements. A separate simulation uses the parametric bootstrap to estimate sampling uncertainty. The notebooks include reference data, worked examples and optional analyses by age.

Results: NHANES 2013–2014 yielded 1,702 eligible adults aged 20–79 without a previous diabetes diagnosis, of whom 68 met the measured reference threshold. We fitted separate two- and three-dimensional mixtures to the same 1,361 participants and retained 6,000 posterior draws for each model. At an observed fasting plasma glucose of 126 mg/dL, the model without age gave a probability of 0.2228 that analytical-error-free two-hour plasma glucose at the same visit was at least 200 mg/dL. The parameter, new-result analytical measurement and total standard uncertainties were 0.0306, 0.0336 and 0.0455, with a 95% parameter credible interval of 0.1680–0.2860. At age 50 years, the age-conditioned probability was 0.1366 and total standard uncertainty 0.0395. Among 341 held-out participants, 14 met the measured reference threshold. Brier scores for this observed outcome were 0.03044 without age and 0.03088 with age; the paired difference was 0.00044 (95% confidence interval −0.00198 to 0.00361). Mean predicted probabilities were 0.0295 and 0.0255, respectively, compared with an observed frequency of 4.11%; calibration intercepts were 0.42 and 0.59 and slopes 1.25 and 1.43. Age conditioning did not demonstrate improved calibration in this exploratory internal comparison.

Conclusions: *BDPU-Threshold* uses the saved posterior draws to estimate reference-positive probability and quantify uncertainty from model parameters and analytical error in the new test result. The model with age underpredicted measured reference positivity more than the model without age. With only 14 positive outcomes, this comparison gives limited evidence about calibration. Population-wide calibration, clinical benefit and validity for the latent reference endpoint require further evaluation.

Keywords

Bayesian inference, medical diagnosis, measurement uncertainty, posterior predictive probability, NHANES, fasting plasma glucose, oral glucose tolerance test, Wolfram Language, age conditioning, parametric bootstrap, software reproducibility

1. Background

BDPU-Threshold estimates probabilities conditional on an observed quantitative test result. For paired measurements, the 'Disease' and 'Nondisease' choices refer to an analytical-error-free reference concentration at or above the threshold and below it, respectively. In simulations, the same choices refer to the specified disease and nondisease classes. The software combines a posterior model for paired test and reference measurements with a separate simulation workflow for examining finite-sample uncertainty. Its 'Tables' and 'Plots' interface presents the evaluated inputs, probabilities and uncertainty summaries while retaining detailed model and age analyses.

Related diagnostic software has examined uncertainty in quantitative-result probabilities, threshold-conditioned predictive values and fitted measurement distributions (1–3). For real data, the present revision jointly models log-transformed test and reference concentrations using a two-component Gaussian mixture, with explicit parameter priors and analytical measurement-error assumptions for both measurements. Here, Q denotes the conditional reference-positive probability, which depends on the uncertain model parameters and the analytical-error-free test concentration underlying a new observed result. The law of total variance decomposes its variance into posterior parameter and new-result analytical measurement components. Parameter credible intervals summarize posterior variation in the probability after averaging over uncertainty in the underlying test concentration; total-Q credible intervals additionally reflect that concentration uncertainty.

The separate simulation workflow supports Normal, Lognormal, Gamma, Student t and Uniform class distributions, with independently specified prevalence and parametric-bootstrap refitting. Its pointwise confidence intervals describe sampling variation under the selected scenario assumptions. Appendix A gives the mathematical derivations. Appendix B describes the notebook functions, their inputs and the results they return. The contribution demonstrated here is a reproducible implementation and numerical verification of these methods, together with a comparison of documented functions against the cited predecessor programs; comparative methodological or clinical superiority is not established.

2. Implementation

2.1.Data source and case-study population

The source retrieval notebook identified six NHANES cycles, 2005–2016, and demographic, fasting-glucose, OGTT, diabetes-questionnaire, reproductive-health, and glycated-hemoglobin components. All 36 corresponding official XPT files and codebooks were retrieved, with source URLs and SHA-256 hashes recorded. The primary analysis uses 2013–2014 to avoid cross-cycle instrument harmonization and unquantified bridge uncertainty (4). Data from the other cycles are preserved on their published measurement scales and are not silently pooled into the primary posterior.

The other five cycles form an archived retrieval and cohort-construction record. They were not used in either frozen model or in the reported holdout evaluation and therefore do not constitute external validation.

Adults aged 20–79 were selected. Age 80 represents a top-coded group, not a true upper age of 80 (5). No previous diagnosed diabetes was operationalized as DIQ010 $= 2$, meaning "No"; "Borderline," unknown, refused, and missing responses were excluded (6). Eligibility required positive paired FPG and OGTT values, no recorded pregnancy in RIDEXPRG or RHD143, a complete OGTT with the full glucose challenge, a specimen collected 105–135 min after the challenge, and a positive OGTT subsample weight. This denotes no known pregnancy; unrecorded pregnancy is not ruled out. HbA1c completeness was not required.

The OGTT protocol comprises a glucose challenge, generally 75 g, with a second venipuncture approximately 2 h later (4). The observed reference classification was two-hour plasma glucose of at least 200 mg/dL. FPG was not included in that definition. In the measurement model, the latent endpoint is the

analytical-error-free, same-visit OGTT concentration at or above that threshold. This endpoint is narrower than confirmed clinical diabetes, for which repeat confirmation may be needed in the absence of unequivocal hyperglycemia (7).

A reproducible stratified random split, seed 31014, reserved approximately 20% within each observed reference class. Training comprised 1,361 participants with 54 positives; evaluation comprised 341 with 14 positives. The split used reference status to preserve the rare positive class, and the model was fitted without holdout measurements. Subsequent model checks and sensitivity analyses are exploratory; this analysis was not prospectively registered. Table 1 summarizes cohort selection.

Table 1. Selection of the primary NHANES 2013–2014 cohort

Selection step	Remaining participants
Demographic participants	10,175
Age 20–79	5,417
Reported no previous diabetes	4,604
Positive paired FPG and OGTT	1,709
No recorded pregnancy	1,706
Complete full OGTT challenge	1,706
OGTT collection 105–135 min; positive OGTT weight	1,702
Observed reference-positive subset	68
Training / holdout	1,361 / 341

2.2.Generative model and parameter priors

Let F and G denote latent log FPG and log OGTT, respectively, with concentrations expressed relative to 1 mg/dL before taking natural logarithms, denoted by log. Let W be the observed pair of log measurements and θ the mixture weights, component means, and observed covariance matrices. A two-component Gaussian mixture allows skewness and heterogeneous dependence. The components describe distributional heterogeneity and are not diabetic and nondiabetic classes.

The index i identifies a training participant, and $k = 1{,}2$ identifies a mixture component. The vectors W_i, Z_i and ε_i are the observed pair, latent pair and additive analytical error, respectively; R is the fixed analytical-error covariance. Component k has weight π_k, mean μ_k, latent covariance Σ_k and observed covariance T_k. The relation $\Sigma_k \succ 0$ means positive definiteness. The notation N_2 denotes a bivariate normal distribution parameterized by its mean and covariance, φ_2 its density, and p a probability density; a vertical bar denotes conditioning.

$$W_i = Z_i + \varepsilon_i,\ \varepsilon_i \sim \mathrm{N}_2(0, R),\ T_k = \Sigma_k + R,\ \Sigma_k \succ 0$$

$$p(W_i \mid \theta) = \sum_{k=1}^{2} \pi_k \varphi_2(W_i;\, \mu_k, T_k),\ \sum_{k=1}^{2} \pi_k = 1$$

The Dirichlet distribution is parameterized by its concentrations. The notation IW_2 denotes a two-dimensional inverse-Wishart distribution with degrees of freedom ν_0 and scale matrix Ψ_0; m_0 is the prior mean vector and κ_0 the relative precision of the component mean. The operator diag forms a diagonal matrix from its arguments.

$$(\pi_1, \pi_2) \sim \mathrm{Dirichlet}(1, 1)$$

$$T_k \sim \mathrm{IW}_2(\nu_0, \Psi_0) \mid (T_k - R \succ 0),\ \mu_k \mid T_k \sim \mathrm{N}_2(m_0, \frac{T_k}{\kappa_0})$$

The prior hyperparameters were $m_0 = (\log 100, \log 110)$, $\kappa_0 = 0.1, \nu_0 = 4,$ and $\Psi_0 = \mathrm{diag}(0.04, 0.25)$. A component covariance prior is conditioned on a positive definite latent covariance $T_k - R$. The two components have identical exchangeable priors. These hyperparameters are fixed assumptions, not estimates from the case-study FPG distribution.

The prior is assigned to the total observed covariance. Consequently, varying the fixed analytical-error covariance R also changes the induced prior for the latent covariance $\Sigma_k = T_k - R$ and the associated conditional mean prior. The CV sensitivity scenarios do not hold a biological-covariance prior fixed.

For component k, let n_k be the assigned-observation count, $\bar{w}_k$ their observed-vector mean, and S_k the sum of their centered outer products. The updated relative mean precision, degrees of freedom, mean vector and scale matrix are κ_k, ν_k, m_k and Ψ_k, respectively. A superscript T denotes transpose. The conjugate updates are:

$$\kappa_k = \kappa_0 + n_k,\ \nu_k = \nu_0 + n_k,\ m_k = \frac{\kappa_0 m_0 + n_k \bar{w}_k}{\kappa_0 + n_k}$$

$$\Psi_k = \Psi_0 + S_k + \frac{\kappa_0 n_k}{\kappa_0 + n_k}(\bar{w}_k - m_0)(\bar{w}_k - m_0)^{\mathrm{T}}$$

Gibbs sampling alternates latent allocation updates, Dirichlet weight updates through normalized independent gamma draws, and component normal–inverse-Wishart updates. Covariance draws outside $T_k - R \succ 0$ are rejected. The resulting chain targets the stated restricted posterior. Finite-chain summaries are Monte Carlo estimates, not exact independent posterior samples. Independent seeds 4101–4104 were used for 6,500 iterations per chain; 2,000 were discarded and every third subsequent draw retained, yielding 1,500 per chain.

Numerical covariance acceptance additionally requires the smallest eigenvalue of the latent covariance to exceed 10^{-10} on the modeled log and standardized scales. A component update fails after 10,001 unsuccessful proposals, corresponding to an initial proposal and up to 10,000 redraws; failure is propagated rather than returning a partial fit.

2.3.Analytical measurement model

The 2013–2014 fasting-glucose laboratory manual reports QC lot means, standard deviations, and replicate counts (8). For QC lot l, let n_l, m_l and s_l denote its replicate count, mean and sample standard deviation, respectively. Within-lot relative variances were pooled separately for low and high concentrations using weights $n_l - 1$. Pooling the rounded tabulated values gave FPG CV 0.01332098 (approximately 1.33%) and high-concentration CV 0.01023046 (approximately 1.02%), with 125 pooled within-lot degrees of freedom in each set.

The pooled degrees of freedom document the CV calculation; they do not parameterize a posterior distribution for the fixed analytical-error covariance.

Let c_{pool} denote the pooled coefficient of variation, and c_F and c_G the fixed CV inputs assigned to FPG and OGTT, respectively.

$$c_{\mathrm{pool}}{}^2 = \frac{\sum_l (n_l - 1)\left(\frac{s_l}{m_l}\right)^2}{\sum_l (n_l - 1)},$$
$$R = \mathrm{diag}\big(\log(1 + c_F{}^2), \log(1 + c_G{}^2)\big)$$

The low-concentration estimate informed FPG error. The high-concentration fasting-glucose QC estimate was used as an explicitly assumed proxy for OGTT analytical imprecision. It was not estimated from participant OGTT dispersion. The CVs were fixed conditional inputs; uncertainty in their estimation was assessed through scenarios rather than a fitted CV posterior. Analytical errors were assumed independent across the paired assays and zero-mean on the log scale, implying median-unbiased concentrations. Common calibration bias, specimen effects, and day-to-day biological variation were not

identified by the single paired measurements. The conditional uncertainty total below does not include these unmodeled sources.

2.4.Posterior probability of reference positivity given FPG

Let t be the observed FPG concentration and $y = \log t$ its log value, using the stated concentration convention. The reference-positive indicator is $D = \mathbf{1}_{\{G \geq \log 200\}}$. It is one when the condition holds and zero otherwise. Let $\mathcal{D}$ denote the training data. The new result updates parameter draws through its predictive density; omitting this weighting would produce a modular conditional average rather than the posterior predictive ratio of the joint model.

$$p(\theta \mid \mathcal{D}, y) = \frac{p(y \mid \theta)p(\theta \mid \mathcal{D})}{\int p(y \mid \theta)p(\theta \mid \mathcal{D})\mathrm{d}\theta}$$

For a possible latent log-FPG value f, $q(\theta, f)$ is the reference-positive probability conditional on f and θ; $\bar{q}_\theta(y)$ averages it over F conditional on y and θ. The operators Pr , E and Var denote probability, expectation and variance; moment subscripts identify the averaging distribution.

$$q(\theta, f) = \Pr(G \geq \log 200 \mid F = f, \theta),\ \bar{q}_\theta(y) = \mathrm{E}_{F|y,\theta}[q(\theta, F)]$$

The posterior reference-positive probability is $p_D(t) = \Pr(D = 1|\mathcal{D}, t)$.

$$p_D(t) = \mathrm{E}_{\theta \mid \mathcal{D}, y}[\bar{q}_\theta(y)]$$

Coordinate subscripts select mean and covariance entries: $\mu_{F,k}$ is the FPG entry of μ_k, while $\Sigma_{FF,k}$, $T_{FF,k}$ and R_{FF} are its latent, observed and analytical-error variances.

For fixed θ and component k, latent log FPG F given observed log FPG y is normal, with mean $\mu_{F,k} + \frac{\Sigma_{FF,k}(y - \mu_{F,k})}{T_{FF,k}}$ and variance $\Sigma_{FF,k}R_{FF}/T_{FF,k}$. Component weights given y are proportional to π_k times the component density of observed log FPG at y. At $F = f$, $q(\theta, f)$ averages the component OGTT upper-tail probabilities conditional on f, with weights proportional to π_k times the component density of latent log FPG at f. Averaging $q(\theta, F)$ over F conditional on y and θ gives $\bar{q}_\theta(y)$, whose closed-form normal-tail mixture expression supplies the implemented first moment directly.

2.5.Sampling measurement and total uncertainty

Define $Q = q(\theta, F)$, the conditional reference-positive probability when both θ and the analytical-error-free FPG are uncertain. Let u_s, u_m and u_total denote the parameter, new-result measurement and total standard uncertainties of Q, respectively. The exact law of total variance, conditional on training data and the new observed FPG, gives the specified components:

$$u_\mathrm{s}{}^2 = \mathrm{Var}_{\theta|\mathcal{D},y}[\bar{q}_\theta(y)]$$

$$u_\mathrm{m}{}^2 = \mathrm{E}_{\theta|\mathcal{D},y}[\mathrm{Var}_{F|y,\theta}(Q)]$$

$$u_\mathrm{total}{}^2 = u_\mathrm{s}{}^2 + u_\mathrm{m}{}^2,\ u_\mathrm{total} = \sqrt{u_\mathrm{s}{}^2 + u_\mathrm{m}{}^2}$$

Parameter uncertainty u_s is posterior uncertainty over model parameters, conditional on the training-data measurement model. New-result measurement uncertainty u_m is uncertainty of Q due to the unobserved latent FPG of the individual being evaluated. OGTT and FPG analytical error in training data already affect parameter inference, so the first term is not a measurement-free sampling contribution. These components do not isolate every measurement source in the training sample. The total is uncertainty of the conditional probability Q and excludes Bernoulli variability of a future binary outcome.

A parameter credible interval is computed from $\bar{q}_\theta(y)$ under $p(\theta|\mathcal{D}, y)$. A total-Q credible interval is computed from draws of Q under the joint parameter and latent-FPG posterior. These intervals answer different questions. They are weighted posterior quantile intervals and are not constructed by clipping symmetric expanded-uncertainty intervals. Frequentist interval coverage is not asserted. Complementary reference-positive and reference-negative probabilities have the same standard uncertainty under identical conditioning.

2.6.Numerical verification and model evaluation

In the archived case-study checks, independent numerical implementations were compared on identical saved posterior draws. Probability and all three standard uncertainties agreed within 10^{-12} on the checked subset. The new-result measurement term became zero with zero FPG analytical error, and variance closure was verified. Thirty-two-node Gauss–Hermite integration agreed with 64 nodes at 126 mg/dL to numerical precision. The archived native sampler was tested on a bounded cohort. The Wolfram notebooks retain the saved data, reference posterior draws and numerical inputs. These historical checks concern the archived inputs and are distinguished from verification of the present notebook revision.

Version 1.0.0 evaluates latent-test variance as squared deviations from the analytic conditional mean. Checked Gauss–Hermite rules are supplemented by transition-aware adaptive integration in Wolfram Language; failed numerical checks return an explicit failure. The archived verification record reports agreement of the all-draw default calculation within 9.93×10^{-16} and a maximum absolute variance error of 2.78×10^{-14} in 15 analytic cases. These retained historical results are not represented as newly regenerated chains or a new clinical validation study. Appendix A specifies the native numerical method.

Current reevaluation of all 6,000 embedded age-free reference draws at FPG 126 mg/dL returned probability 0.22278194, parameter SD 0.03061092, new-result measurement SD 0.03363828 and total SD 0.04548145. With the native seed-555 prediction inputs, the total-Q interval was 0.14434134–0.32042303. Native evaluation of the embedded age-conditioned model at age 50 returned probability 0.13658467. These are reevaluations of saved posteriors, not new model fits.

Convergence was evaluated for label-invariant probabilities at 90, 110, 126, 140, and 160 mg/dL, latent reference-positive prevalence, and observed FPG tail probability. Rank-normalized split and folded potential scale-reduction diagnostics, denoted $\hat{R}$, were calculated (9). Monte Carlo error of the likelihood-weighted probability ratio was evaluated using its centered influence sequence and chain autocorrelation. Posterior predictive replicated cohorts examined FPG skewness and kurtosis, measured reference positivity, and reference positivity within FPG 120–140 mg/dL. Fixed-error, stronger-prior, and one-component scenarios assessed sensitivity; these exploratory scenarios used single chains and are not exhaustive robustness evidence.

Internal evaluation compared both frozen models on the same original holdout participants, using each participant's measured FPG and, for the age-conditioned model, recorded age. The endpoint was measured OGTT at least 200 mg/dL. Predictions used the total observed-reference covariance, including analytical measurement variance, and likelihood weighting across mixture components and posterior draws. This scoring endpoint differs from the analytical-error-free endpoint used in the uncertainty example. Neither model was refitted, tuned or recalibrated for scoring on the holdout; training-row identity and absence of train–holdout overlap were verified.

The measured-reference calculation conditions the observed Gaussian mixture directly. Let U be measured log OGTT and let x contain observed log FPG, together with standardized age when used. The block label O selects these observed conditioning coordinates. For M retained draws and K mixture components, indexed by h and k, let Ω_{hk} be the total observed covariance and μ_{hk} its mean; d_O is the number of conditioning coordinates. The normalized weights $\alpha_{hk}(x)$ sum to one over all draws and components; r and j index the same draws and components in their normalization:

$$\alpha_{hk}(x) = \pi_{hk}\varphi_{d_O}(x; \mu_{O,hk}, \Omega_{OO,hk}) / \sum_{r=1}^{M}\sum_{j=1}^{K}\pi_{rj}\varphi_{d_O}(x; \mu_{O,rj}, \Omega_{OO,rj}).$$

The component conditional mean $m_{hk}(x)$ and variance v_{hk} of measured log OGTT are

$$m_{hk}(x) = \mu_{U,hk} + \Omega_{UO,hk}\Omega_{OO,hk}^{-1}(x - \mu_{O,hk}),$$

$$v_{hk} = \Omega_{UU,hk} - \Omega_{UO,hk}\Omega_{OO,hk}^{-1}\Omega_{OU,hk}.$$

With M retained draws, K mixture components and standard normal cumulative distribution function Φ, the holdout probability is

$$p_{\mathrm{obs}}(x) = \sum_{h=1}^{M}\sum_{k=1}^{K}\alpha_{hk}(x)\Phi((m_{hk}(x) - \log 200)/\sqrt{v_{hk}}).$$

The total reference variance is retained in this calculation. Sequential conditioning on age and then measured FPG gives the same result. Integrating a correctly specified measured-reference tail over the latent test value is also equivalent, but is unnecessary for this scoring endpoint.

The logit transformation is the natural logarithm of the odds of a probability strictly between zero and one. Calibration summaries included observed and mean predicted positivity, their difference, the observed/expected event ratio, a logistic calibration intercept with logit(prediction) as an offset (ideal zero), and a calibration slope estimated with a free intercept (ideal one) (10). Brier score and logarithmic loss summarized overall probability prediction, while AUROC summarized discrimination. Fixed risk bins and logistic calibration curves were descriptive. The Brier comparator was the constant training-positive fraction.

The age comparison was exploratory. A dated analysis plan preceded computation of the age-model holdout predictions. Percentile 95% confidence intervals used 5,000 ordinary paired participant bootstrap resamples (seed 20261018), with identical resampled participants for both models and differences defined as age minus no age. These unweighted intervals are conditional on the saved fits and exclude model-training and survey-design uncertainty. Calibration-fit failures were recorded; none occurred. Fourteen positives limit precision and do not support reliable flexible or age-subgroup calibration assessment.

2.7.Survey design and inferential scope

Paired FPG/OGTT survey summaries use WTSOG2YR rather than fasting-only weights, together with SDMVSTRA and SDMVPSU(4,11). A descriptive weighted reference-positive fraction and Taylor-linearized standard error were calculated, preserving zero domain contributions for excluded participants and the 15 strata and 30 primary sampling units. Additional protocol and complete-pair selections were not weight-recalibrated. The primary likelihood is unweighted and conditionally iid; its posterior describes the selected cohort and does not establish United States population uncertainty. Weighted descriptive estimates cannot convert that posterior into a survey-design posterior.

2.8.Software architecture and use

BDPU-Threshold is implemented in the Wolfram Language. *BDPU-Threshold.nb* contains the working analysis and interactive calculator. *BDPU-ThresholdHelper.nb* provides methods, callable functions, and examples.

After selecting ‘Start analysis’, users review the data source, outcome, observed result, and interval level (as a percentage) before selecting ‘Calculate probabilities’. ‘More settings’ controls the reference threshold, display precision, plot range, interval type and optional age conditioning. ‘Data and model’ contains the fitting and simulation controls.

Changing the training data, analytical coefficients of variation (CVs), priors or number of mixture components requires an explicit refit. The ‘Import CSV and fit’ action runs one exploratory chain. ‘Restore reference posteriors’ reloads the saved fits. To evaluate another observed result or interval level, enter the new value and select ‘Calculate probabilities’ using the existing fit.

The headline gives the probability of the selected outcome, followed by the applied measurement and interval level. ‘Tables’ reports the probability, interval and standard uncertainty. ‘Plots’ shows how probability and uncertainty vary over the test range, alongside the measurement distributions. For real data, the parameter interval describes variation across posterior parameter draws. The total-Q interval also includes uncertainty in the underlying test concentration. The simulation reports a pointwise percentile-bootstrap interval. Point estimates and interval endpoints remain absolute when relative uncertainty is selected. Age comparisons, profiles and surfaces have separate controls. Export actions use the retained result and its applied settings. Applying the workflow to another condition requires a suitable continuous reference endpoint, compatible measurement-error information, defensible priors and evaluation.

2.9.Optional age conditioning

With age enabled, *a* three-dimensional joint Gaussian mixture models log test, log reference and recorded age standardized about 50 years with scale 20 years. Age is treated as observed without measurement or rounding error. The analytical-error covariance has the two assay variances and *a* zero age variance. Default priors retain mean precision 0.1, use covariance degrees of freedom 5 and add age prior mean 0 and scale 1 on the standardized coordinate. The age-disabled model prior has covariance degrees of freedom 4. These are separate fitted models, not an age label attached to the two-dimensional posterior.

The age model uses the same 1,361 training measurement pairs, in identical order, with complete recorded ages 20–79. The 341 holdout participants were not used. The embedded reference posterior was originally obtained from four chains, seeds 6101–6104, each used 6,500 iterations, burn-in 2,000 and thinning 3, producing 6,000 retained draws. Six targeted label-invariant split rank/folded $\hat{R}$ checks at ages 25, 50 and 75 and FPG 110 and 126 mg/dL were below 1.00143. These limited checks do not certify comprehensive convergence or predictive performance. The Gaussian age density is not truncated at the training range, but the application refuses predictions outside the observed age range.

Conditioning first updates the component weights by their age densities and obtains conditional means and covariances through *a* Schur complement. Posterior parameter draws must also be weighted by their age likelihood; the subsequent test-result likelihood completes the posterior weighting. Omitting the parameter-draw age factor would change the estimand. The existing probability integration and variance decomposition then apply to the age-conditioned model. Uncertainty in the fitted age distribution and its associations is included through parameter draws; uncertainty in the recorded age itself is excluded.

2.10. Simulated data and uncertainty

The simulation engine specifies independent disease and nondisease measurement distributions from Normal, Lognormal, Gamma, Student *t* or Uniform families, using original-scale means and SDs. Population prevalence supplies the prior disease probability. The user enters it as a decimal independently of the class counts. Select 'Generate simulation and bootstrap' before calculating probabilities and repeat this step after changing the generating settings. The engine reports the known generating probability and the probability estimated from the original simulated samples by moment fitting. Parametric bootstrap samples are drawn from the fitted class distributions and moment-refitted; their probability variation estimates sampling SD and pointwise percentile intervals. Class SDs are re-estimated within each bootstrap. Fixed prevalence, family choice, Student *t* degrees of freedom and any supplied age-effect coefficients do not acquire an uncertainty distribution.

A reported simulation interval and SD require at least 100 replicates and complete valid support across all replicates at that test value. If an estimate or bootstrap probability is undefined, the application counts the failures and retains *a* gap rather than discarding failed replicates to calculate *a* conditional interval. These sampling intervals are not the real-data parameter or total-Q credible intervals. No additional analytical-error variance component is introduced for simulated measurements.

2.11. Comparison views and implementation contracts

The real-data comparison evaluates the age-disabled two-dimensional fit and the separate age-conditioned three-dimensional fit on the same positive test grid, threshold and interval level. Matching training measurement rows, fitted CVs, measurement labels and units are required. 'Fit paired comparison models' reads a three-column CSV containing the test measurement, reference measurement and age. It replaces the two fits only after both succeed. The curves need not coincide at age 50 because the models and dimensional priors are distinct. Their difference is not *a* causal age effect, and the intervals attached to each curve do not define an interval for their difference.

Simulation comparison contrasts anchor settings with age effects disabled against *a* selected-age scenario. Class means change linearly by specified original-scale amounts per decade; prevalence changes on the log-odds scale. Class SDs, families and degrees of freedom are held equal across the two generating scenarios. The same simulation and bootstrap seeds and class counts are used, and effects are applied once before generation and refitting. This convention supports reproducibility but does not guarantee

identical random variates across different libraries or distributions. The baseline is an anchor scenario, not *a* probability marginalized over population age.

Both simulated curves use the union of their support-aware grids, preserving independent gaps and narrow Uniform supports. Zero age slopes or selected age equal to the anchor reproduce the same scenario exactly. Relative SD is calculated separately using each curve's selected disease or nondisease probability and is undefined at zero probability. Separate applied snapshots preserve previous successful comparisons when pending inputs change or an action fails. Appendix B documents the module and reproducibility contracts.

2.12. Probability surfaces and profiles over age

The three-dimensional display evaluates probability directly on an explicit grid of test measurements and ages. For real data it uses the age-conditioned posterior and latent reference threshold. For simulations it can display the known generating probability or the probability fitted to samples generated at each age under the specified scenario. It displays probability only; no uncertainty surface or simultaneous probability region is inferred.

The age profile instead holds the user's test measurement exactly fixed and evaluates the selected ages directly, without extracting an interpolated section from the three-dimensional grid. Its uncertainty action calculates the real-data parameter, new-result measurement and total standard uncertainties, with parameter or total-Q credible intervals. In simulation mode it calculates the sample-fitted probability, bootstrap sampling SD and pointwise percentile confidence interval, and shows the generating probability separately. The age profile takes the interval level as a fraction, with a default of 0.95. Changing only this level changes interval endpoints, not the probability or standard uncertainty. Unsupported values remain gaps; Uniform-family views use unconnected points and interval bounds to avoid implying interpolation across discontinuous support.

3. Results

3.1. Posterior probabilities and uncertainty

Table 2 presents the primary mixture results. At an observed FPG of 126 mg/dL, the estimated reference-positive probability was 0.2228, with sampling parameter SD 0.0306, new-FPG analytical SD 0.0336, and total SD 0.0455. For the archived case-study prediction inputs, the parameter 95% credible interval was 0.1680–0.2860; the total-Q interval was 0.1440–0.3223. Both intervals use the same point estimate. The wider interval also reflects uncertainty in the analytical-error-free FPG underlying the new measurement.

Table 2. Reference-positive probability and conditional standard uncertainty by observed FPG

FPG mg/dL	Point estimate	Sampling parameter SD	New FPG analytical SD	Total SD	95% parameter credible interval
90	0.0057	0.0011	0.0007	0.0012	0.0039–0.0080
100	0.0162	0.0020	0.0023	0.0031	0.0125–0.0203
110	0.0447	0.0050	0.0067	0.0083	0.0354–0.0552
120	0.1230	0.0159	0.0207	0.0261	0.0949–0.1567
126	0.2228	0.0306	0.0336	0.0455	0.1680–0.2860
140	0.4641	0.0676	0.0211	0.0708	0.3434–0.6143
160	0.6229	0.0791	0.0140	0.0803	0.4762–0.7922

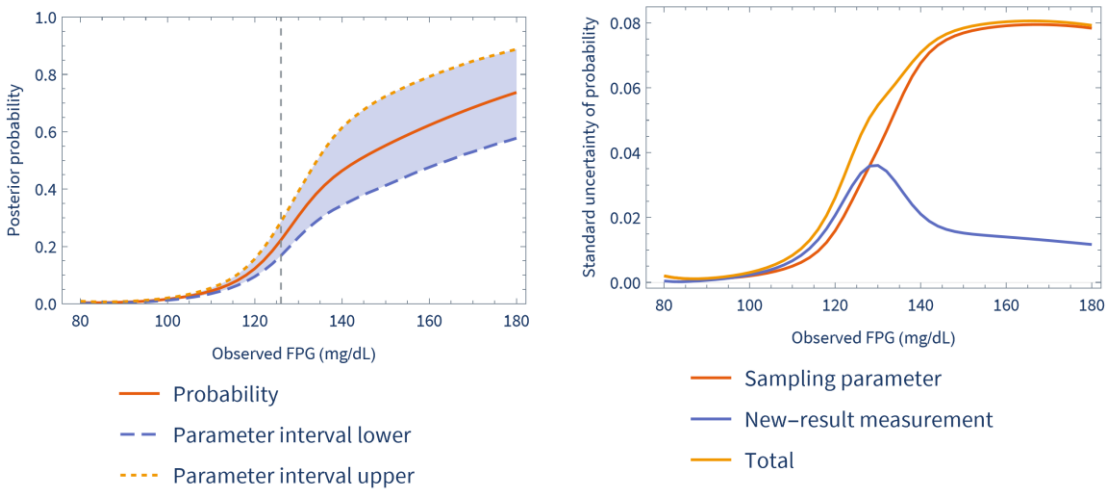


Figure 1. Posterior reference-positive probability and uncertainty by observed FPG. The left panel shows the integrated probability and its parameter credible interval. The right panel shows parameter, new-result analytical measurement and total standard uncertainties. The vertical line marks 126 mg/dL for orientation; the probability curve conditions on the quantitative result.

3.2.Convergence and model checks

All checked rank-normalized split and folded $\hat{R}$ values were below 1.0015. At FPG 126 mg/dL, new-result likelihood weighting retained importance effective sample size (ESS) 5,933 of 6,000 draws. The probability-ratio autocorrelation ESS was approximately 2,817 and Monte Carlo standard error 0.00058. These checks support numerical stability for the examined estimands, conditional on the selected model; they do not establish its clinical validity.

The posterior expected measured-reference-positive fraction was 0.0372, compared with the training observed fraction 0.0397. Replicated-cohort 95% intervals included the observed reference fraction and reference positivity in the FPG 120–140 mg/dL band. FPG-tail fit remained imperfect: observed log-FPG excess kurtosis 13.07 exceeded the replicated 95% upper limit 12.56, with replicated upper-tail probability 0.024. Observed skewness 2.05 was near the replicated upper tail. Figure 2 provides marginal fit displays; marginal visual agreement does not establish conditional calibration.

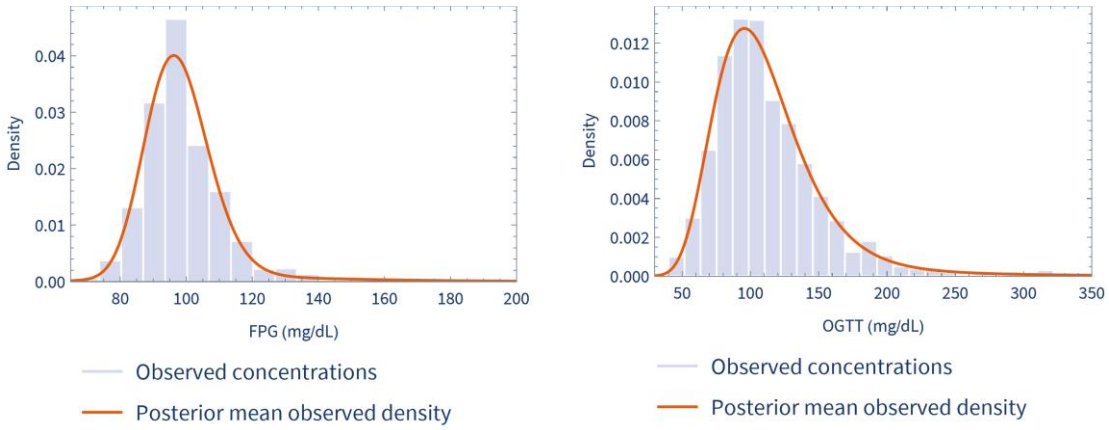


Figure 2. Training measurement distributions and posterior mixture densities. Histograms show observed concentrations; curves show posterior mean observed-density estimates. Display limits emphasize the main data range. Extreme tails remain a source of model inadequacy.

3.3.Sensitivity and internal evaluation

Table 3 shows probability sensitivity at 126 mg/dL. Removing analytical error gave probability 0.2242 and eliminated the new-FPG measurement term. Setting both CVs to 2% gave probability 0.2238 and total SD 0.0599. Stronger component priors gave probability 0.2130. A single Gaussian gave 0.1533,

demonstrating material dependence on distributional assumptions; its smaller interval is not evidence of greater accuracy.

The scenario labeled Stronger prior used $\kappa_0 = 1$ and $\Psi_0 = \mathrm{diag}(0.16,1)$, with unchanged m_0 and $\nu_0 = 4$. It therefore changes the covariance prior scale as well as the conditional mean precision; it is not solely an increase in prior concentration.

Table 3. Exploratory model and analytical-error scenarios at FPG 126 mg/dL

Scenario	Point estimate	Sampling parameter SD	New FPG analytical SD	Total SD
Primary two-component mixture	0.2228	0.0306	0.0336	0.0455
Zero analytical error	0.2242	0.0316	0.0000	0.0316
CV 2% each	0.2238	0.0317	0.0509	0.0599
Stronger prior	0.2130	0.0297	0.0318	0.0435
Single Gaussian	0.1533	0.0146	0.0143	0.0205

The age-free holdout Brier score was 0.03044 versus 0.03937 for the constant training-fraction predictor, a descriptive relative reduction of 22.7%. On the same 341 holdout participants, the age-conditioned Brier score was 0.03088; the paired difference was 0.00044 (95% confidence interval −0.00198 to 0.00361). The 14 measured reference positives gave an observed frequency of 4.11%, compared with mean predicted probabilities of 0.0295 without age and 0.0255 with age. Corresponding observed/expected event ratios were 1.39 and 1.61. Among the 14 holdout participants with FPG at least 120 mg/dL, seven were positive, compared with 4.218 expected by the age-free model. The paired evaluation is reported below.

Calibration intercepts were 0.42 without age and 0.59 with age; slopes were 1.25 and 1.43. Both models underpredicted the observed positive frequency, with greater underprediction after age conditioning. The difference in absolute mean calibration error was 0.00392 (95% confidence interval −0.00370 to 0.00691). Logarithmic loss decreased from 0.12057 to 0.11531 and AUROC increased from 0.8617 to 0.9198, but their paired differences also had intervals spanning zero: −0.00525 (−0.01923 to 0.00864) and 0.0581 (−0.0165 to 0.1543), respectively. These findings do not demonstrate improved calibration, predictive superiority or equivalence.

The complete-cohort observed positive fraction was $68/1{,}702 = 4.00\%$. Its OGTT-weighted descriptive estimate was 3.2529%, with Taylor standard error 0.4565 percentage points. The difference reflects the selected sample composition and weights; the weighted figure is presented with the additional-selection limitation stated above.

3.4.Age and implementation comparisons

Table 4 contrasts the age-disabled two-dimensional fit with the age-conditioned three-dimensional fit at the same FPG result and latent OGTT threshold. At age 50 years, the probability is 0.1366, sampling parameter SD 0.0325, new-result measurement SD 0.0225 and total SD 0.0395; the 95% parameter credible interval is 0.0830–0.2091. Compared with the age-disabled probability 0.2228, this is a difference between separate fitted models on the same measurement rows. It does not demonstrate improved clinical prediction. A separate measured-reference holdout comparison follows the interval results. Both models use a latent OGTT threshold of 200 mg/dL, 6,000 retained posterior draws and shared prediction inputs.

Table 4. Real-data probability and uncertainty at FPG 126 mg/dL, with and without age

Model and age	Point estimate	Parameter SD	New-test SD	Total SD	95% parameter interval
2D without age	0.2228	0.0306	0.0336	0.0455	0.1680–0.2860
3D age 25 years	0.0425	0.0137	0.0062	0.0151	0.0237–0.0721
3D age 50 years	0.1366	0.0325	0.0225	0.0395	0.0830–0.2091
3D age 75 years	0.3494	0.0602	0.0378	0.0711	0.2368–0.4701

The archived total-Q interval reported for the original case study was 0.1440–0.3223. With the same age-free posterior and the current seed-555 prediction inputs, the calculator gives 0.1443–0.3204 (Table 5). A controlled historical substitution reproduced both intervals without changing the probability, SDs, parameter interval or importance ESS. The difference reflects finite Monte Carlo prediction inputs, not an age effect or a new fit. The archived age-50 shared-input total-Q interval is 0.0775–0.2303.

Table 5 presents the interval-limits results corresponding to the age-free and age-50 uncertainty summaries. The table gives both complementary outcomes for each fit, with separate panels for parameter and total-Q limits. In the calculator, 'Tables' shows the selected outcome. The reference-positive outcome denotes analytical-error-free, same-visit OGTT at or above 200 mg/dL; reference-negative denotes its complement. These labels do not establish a confirmed clinical diagnosis.

Table 5. Probability point estimates and 95% credible limits at FPG 126 mg/dL

(A) PARAMETER CREDIBLE

Measurand	Point estimate	Lower limit	Upper limit
2D without age: reference-positive probability	0.2228	0.1680	0.2860
2D without age: reference-negative probability	0.7772	0.7140	0.8320
3D age 50 years: reference-positive probability	0.1366	0.0830	0.2091
3D age 50 years: reference-negative probability	0.8634	0.7909	0.9170

(B) TOTAL-Q CREDIBLE

Measurand	Point estimate	Lower limit	Upper limit
2D without age: reference-positive probability	0.2228	0.1443	0.3204
2D without age: reference-negative probability	0.7772	0.6796	0.8557
3D age 50 years: reference-positive probability	0.1366	0.0775	0.2303
3D age 50 years: reference-negative probability	0.8634	0.7697	0.9225

Observed FPG is 126 mg/dL (approximately 7.0 mmol/L); the latent OGTT threshold is 200 mg/dL (approximately 11.1 mmol/L). Each model uses 6,000 saved posterior draws. Parameter credible limits describe the posterior distribution of the conditional mean; total-Q limits also propagate uncertainty in the new latent test result. The point estimate is the same analytic posterior mean for both interval types. Reference-negative estimates equal $1 - p$, with bounds $[1 - U, 1 - L]$ for reference-positive bounds $[L, U]$. Total-Q limits use one latent-test draw per posterior draw from the current seed-555 inputs for the age-free fit and the matching saved prediction tape for the age-50 fit; their reported digits identify that realization rather than Monte Carlo endpoint precision. No model was refitted. Table 6 reports the paired measured-reference holdout evaluation.

Table 6. Paired held-out prediction of measured OGTT positivity

Measure	Without age (95% CI)	With age (95% CI)
Mean predicted probability	0.0295 (0.0228 to 0.0376)	0.0255 (0.0196 to 0.0334)
Calibration intercept	0.42 (−0.30 to 0.91)	0.59 (−0.11 to 1.08)
Calibration slope	1.25 (0.85 to 1.97)	1.43 (1.01 to 2.26)
Brier score	0.03044 (0.01539 to 0.04729)	0.03088 (0.01561 to 0.04848)
Logarithmic loss	0.12057 (0.06614 to 0.18220)	0.11531 (0.06381 to 0.17436)
AUROC	0.8617 (0.7226 to 0.9732)	0.9198 (0.8532 to 0.9737)

The same 341 holdout participants included 14 measured reference positives (4.11%). Confidence intervals are percentile intervals from 5,000 paired participant bootstrap resamples, conditional on the frozen fits. The calibration intercept uses a fixed slope of one; the calibration slope uses a freely estimated intercept. Their ideal values are zero and one, respectively. Paired differences and uncertainty are reported in the text. These measures do not validate the analytical-error-free endpoint.

Historical verification on saved real-data comparison inputs recorded maximum probability discrepancy below 5.6×10^{-17} and interval-endpoint discrepancy below 1.2×10^{-14}. For the holdout comparison, a separately implemented numerical calculation reproduced all 341 predictions from each model within 5×10^{-15} and the point metrics within 3.5×10^{-13}. The holdout bootstrap intervals are retained archived summaries; the present notebook revision does not rerun the 5,000-resample analysis. These checks support numerical agreement for the specified inputs, not independent clinical validation or interval coverage. Table 7 summarizes the applied Wolfram architecture.

Table 7. Wolfram architecture and calculation contracts

Responsibility	**Wolfram component**	**Defined contract**
Interface	Main diagnostic notebook	Selected outcome probability, applied inputs, tables and plots
Real inference	BayesianDiseaseFit and BayesianDiseaseProbability	Restricted mixture posterior and checked probability moments
Age conditioning	Native age fit and conditioning functions	Component and parameter-draw age weights; fitted-range guard
Simulation	Native simulation and bootstrap functions	Five class families; fixed prevalence; bootstrap refitting
Age views	Comparison, profile and surface functions	Matched real fits or supplied scenarios; explicit support gaps
Exports	Native CSV, JSON and figure actions	Applied results and numerical context
Reference data	Embedded measurements and posterior draws	Saved inputs retain their historical provenance
Methods	Methods helper notebook	Equations, function contracts and worked examples

Table 8 compares the published functions of the cited predecessors with the present implementation. The comparison concerns documented methods and software scope, not comparative diagnostic performance.

Table 8. Comparison with cited predecessor software

Software and source	**Published functionality**	**Relation to *BDPU-Threshold* current**
Bayesian Diagnostic Uncertainty (2024) (1)	Normal, Lognormal and Gamma class distributions; sampling and measurement uncertainty propagated by a first-order Taylor approximation, with approximate confidence limits.	Real-data inference fits a joint latent test/reference mixture. Posterior integration separates parameter and new-result measurement variance and provides parameter and total-Q credible limits.
Bayesian Diagnostic Insights (2024) (2)	Quantitative-result probabilities and threshold-conditioned positive and negative predictive values, with first-order uncertainty propagation in a computational notebook.	*BDPU-Threshold* focuses on the observed quantitative result, adds optional age conditioning, and supplies a native Wolfram notebook interface and executable methods.
Bayesian Diagnosis (2023) (3)	Parametric and kernel-density models for posterior disease probabilities from one or two diagnostic tests.	*BDPU-Threshold*'s real-data workflow models a test paired with a continuous reference endpoint. Its separate simulation workflow uses parametric-bootstrap refitting; it does not reproduce the general two-test kernel-density comparison.

Predecessor entries describe the published versions in references (1–3), not all subsequent revisions. *BDPU-Threshold* entries are supported by the implementation and verification records described here. The comparison does not establish equivalence of estimands, numerical accuracy, interval coverage, computational efficiency or clinical benefit.

Figures 3 and 4 use the saved real-data models and show the probability and total-uncertainty comparisons at age 50 years. Figures 5 and 6 illustrate the archived simulation comparison, not a second NHANES analysis. At anchor age 50 years, Normal disease and nondisease means are 130 and 100, SDs 20 and 15, and prevalence 0.1. At age 70, mean changes of 5 and 1 per decade give means 140 and 102; a prevalence log-odds change of 0.3 per decade gives prevalence 0.1684. There are 500 measurements in each class and 200 bootstrap replicates, with seeds 20261010 and 20261011. Sampling variation is conditional on these fixed assumptions.

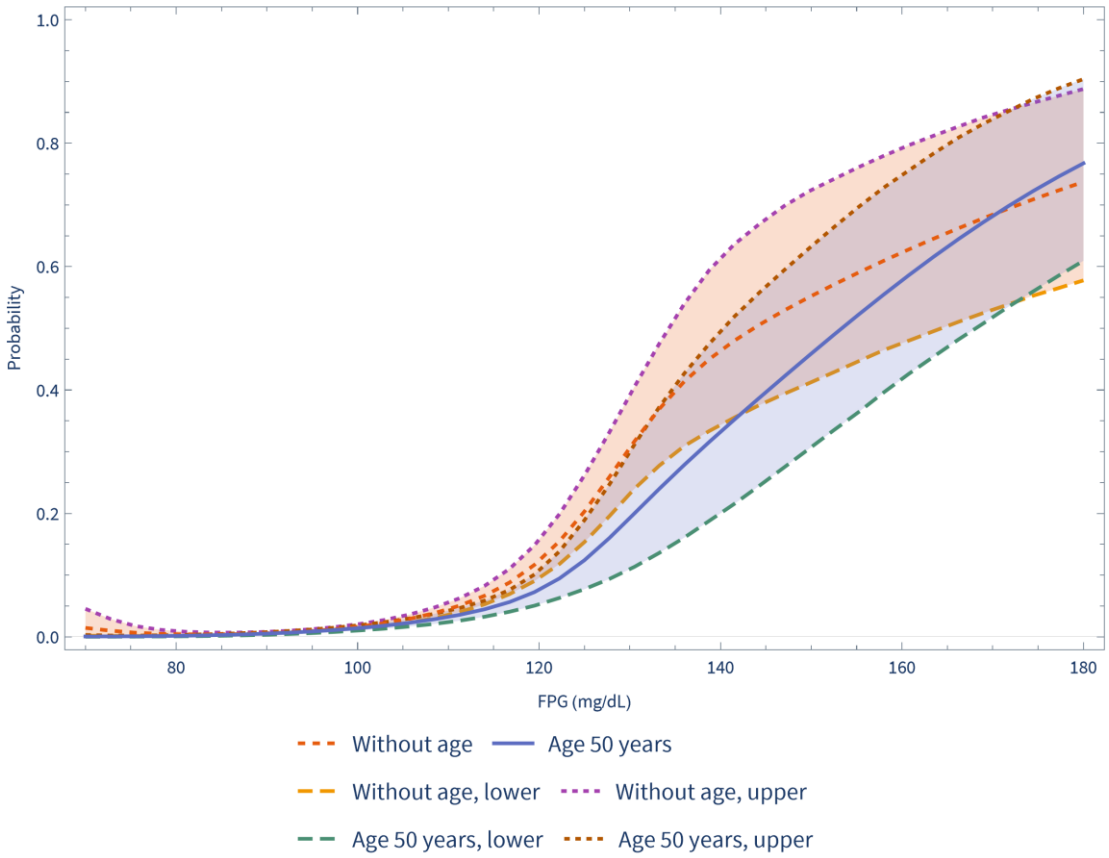


Figure 3. Real-data probability comparison. Orange dashed: the age-disabled 2D fit. Blue solid: the 3D fit conditioned on age 50. Shading denotes separate 95% parameter credible intervals; the latent OGTT threshold is 200 mg/dL.

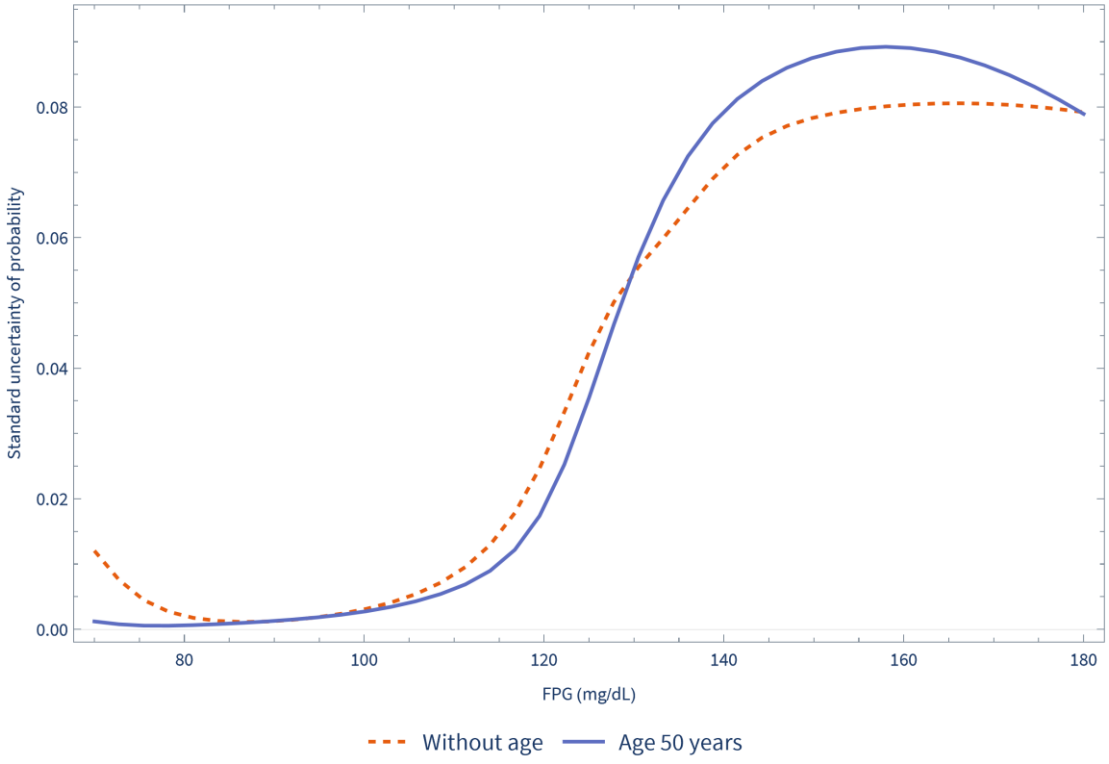


Figure 4. Real-data total standard uncertainty for the same models and grid as Figure 3. Table 4 also reports parameter and new-result measurement components. A smaller uncertainty at a test value does not establish better predictive accuracy.

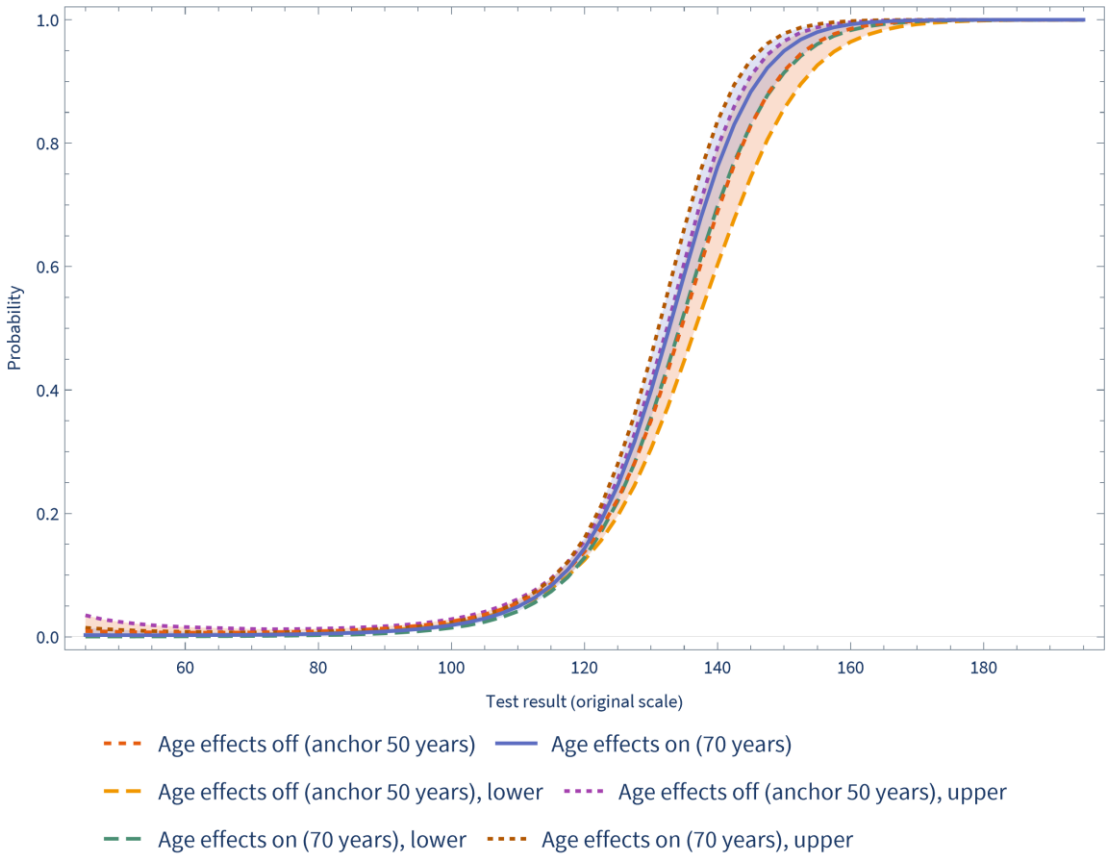


Figure 5. Simulated-data probability comparison. Orange dashed: anchor settings at 50 years, with age effects disabled. Blue solid: the selected-age scenario at 70 years. Shading denotes separate 95% pointwise bootstrap percentile intervals; the settings and seeds are stated in the text.

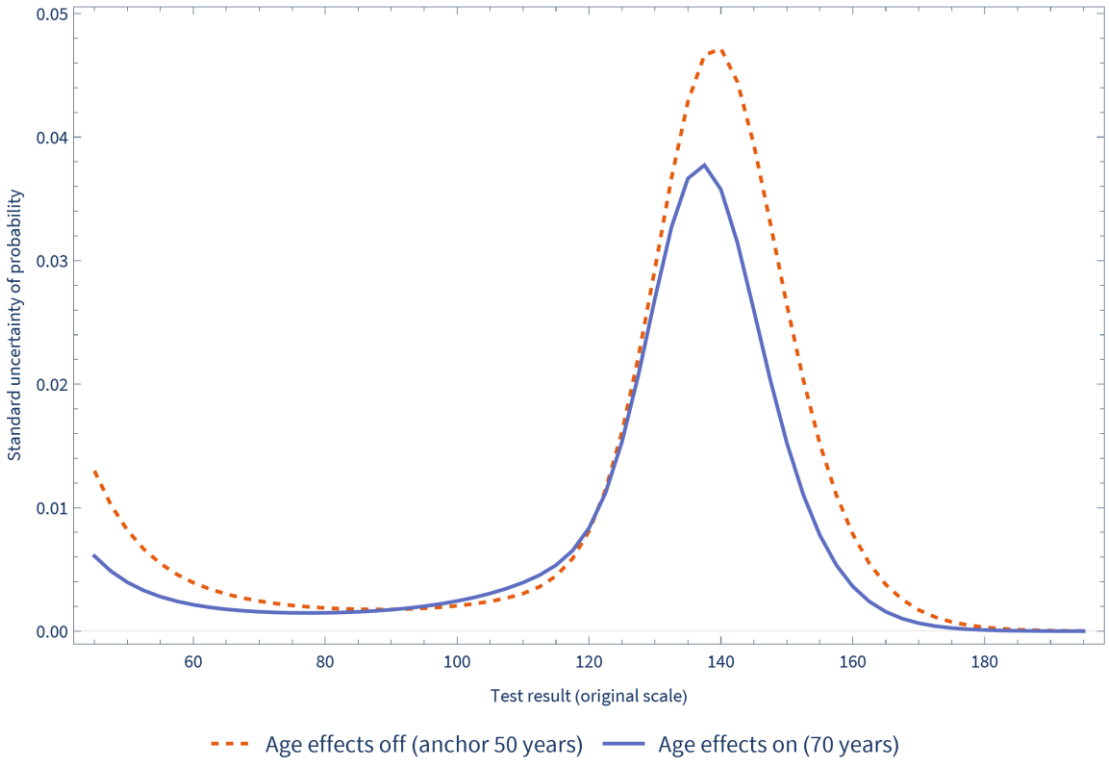


Figure 6. Bootstrap sampling standard uncertainty for the two simulation scenarios in Figure 5. SDs are re-estimated in bootstrap samples. Uncertainty in fixed prevalence, family, degrees of freedom and age slopes is excluded.

3.5.Age profiles and probability surfaces

Figures 7 and 8 evaluate the saved real-data age model at a fixed FPG measurement across the selected ages. Figures 9 and 10 show the corresponding fitted simulation probability and its sampling uncertainty under the stated Normal scenario. Figures 11 and 12 display probability over both test measurement and age. The figures show conditional calculations using the archived fits and simulation settings. Table 6 assesses predictive calibration against held-out measurements.



Figure 7. Real-data probability over age at exactly FPG 126 mg/dL and latent OGTT threshold 200 mg/dL. The saved three-dimensional posterior is evaluated at 15 ages from 25 to 75 years. The orange line shows the integrated probability. The shaded pointwise 95% parameter credible interval has a blue dashed lower bound and a gold dotted upper bound. The profile is evaluated directly at the fixed measurement, without interpolation from a surface.

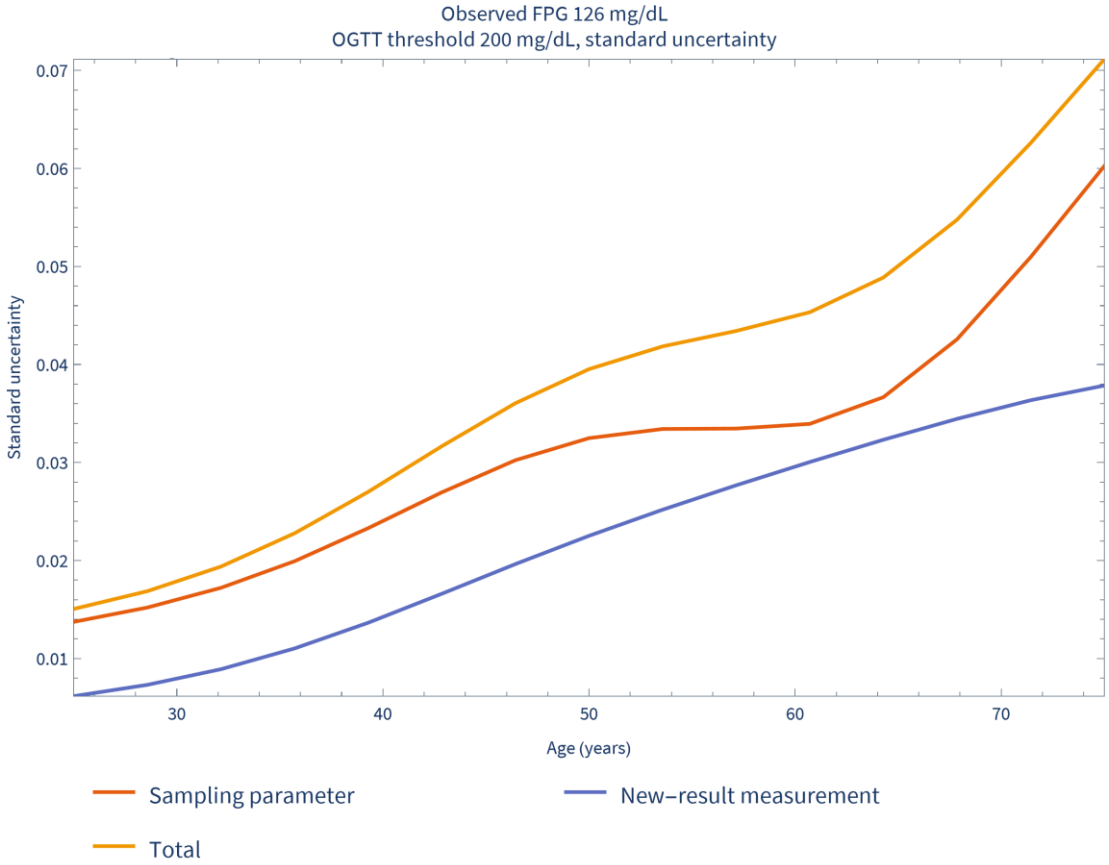


Figure 8. Standard uncertainty of the conditional probability over age for the same fixed measurement, threshold, posterior and age grid as Figure 7. Sampling parameter uncertainty is orange, new-result measurement uncertainty is blue, and total uncertainty is gold. The three quantities retain the variance definitions in Appendix A and exclude future binary-outcome variability.

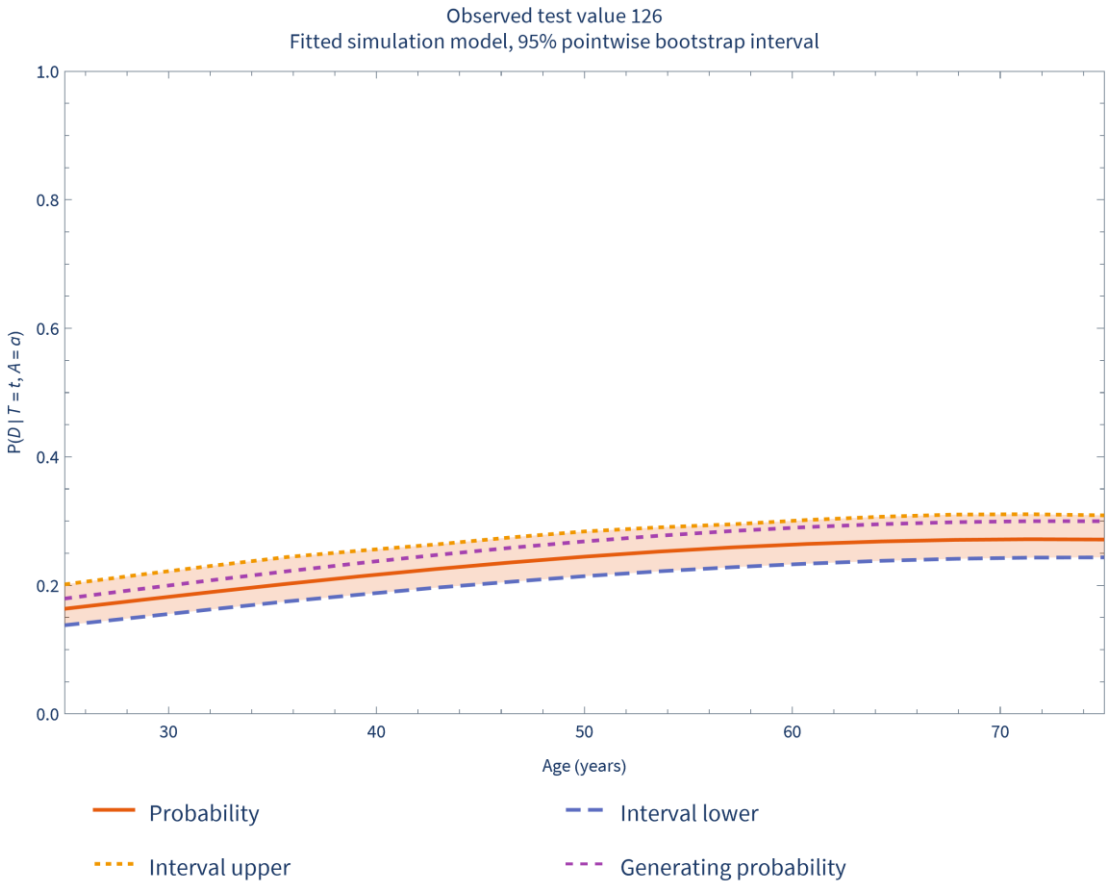


Figure 9. Simulated-data probability over age at exactly test measurement 126, using the Normal anchor settings and age slopes specified for Figures 5 and 6. Orange denotes the probability moment-fitted to simulated samples, with pointwise 95% percentile-bootstrap intervals; the known generating probability is shown separately. Each age uses 500 observations per class and 200 bootstrap replicates with the recorded simulation and bootstrap seeds. The age coefficients and prevalence inputs are fixed scenario assumptions.

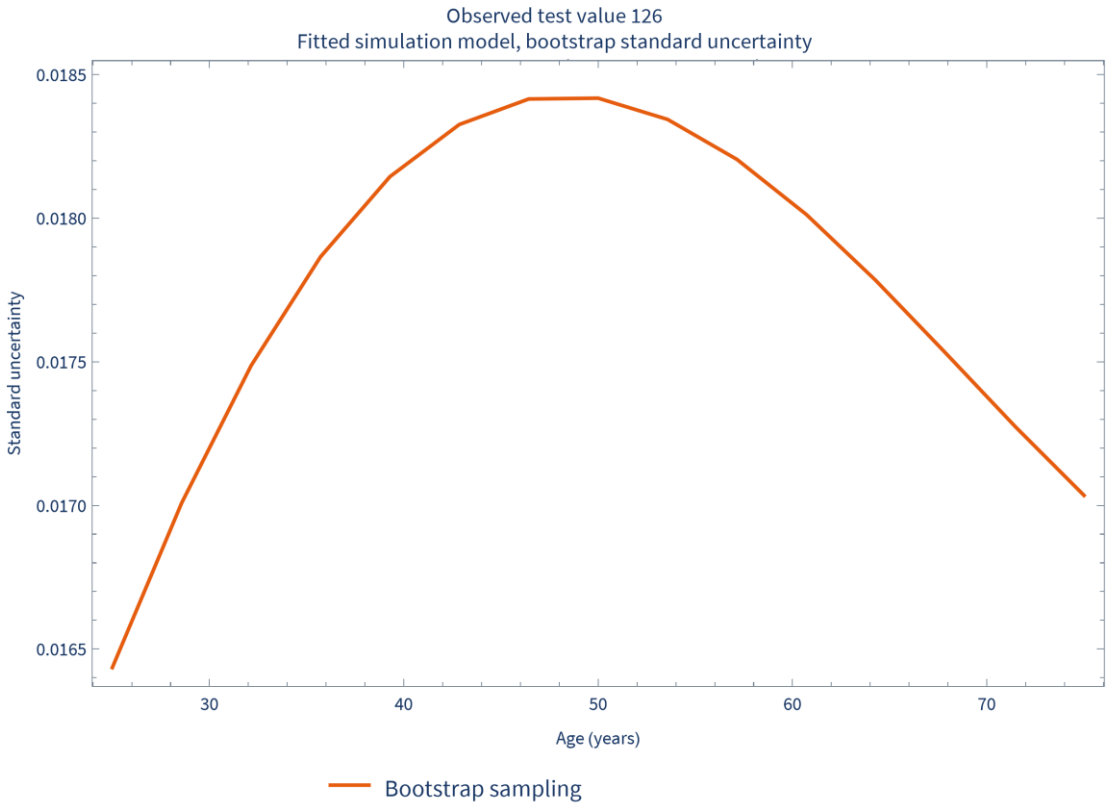


Figure 10. Bootstrap sampling standard uncertainty over age for the same fixed measurement and simulation scenario as Figure 9. This uncertainty describes the sample-fitted probability under the specified class families, prevalence and age slopes. It does not include a separate analytical measurement component or uncertainty in the fixed scenario inputs.

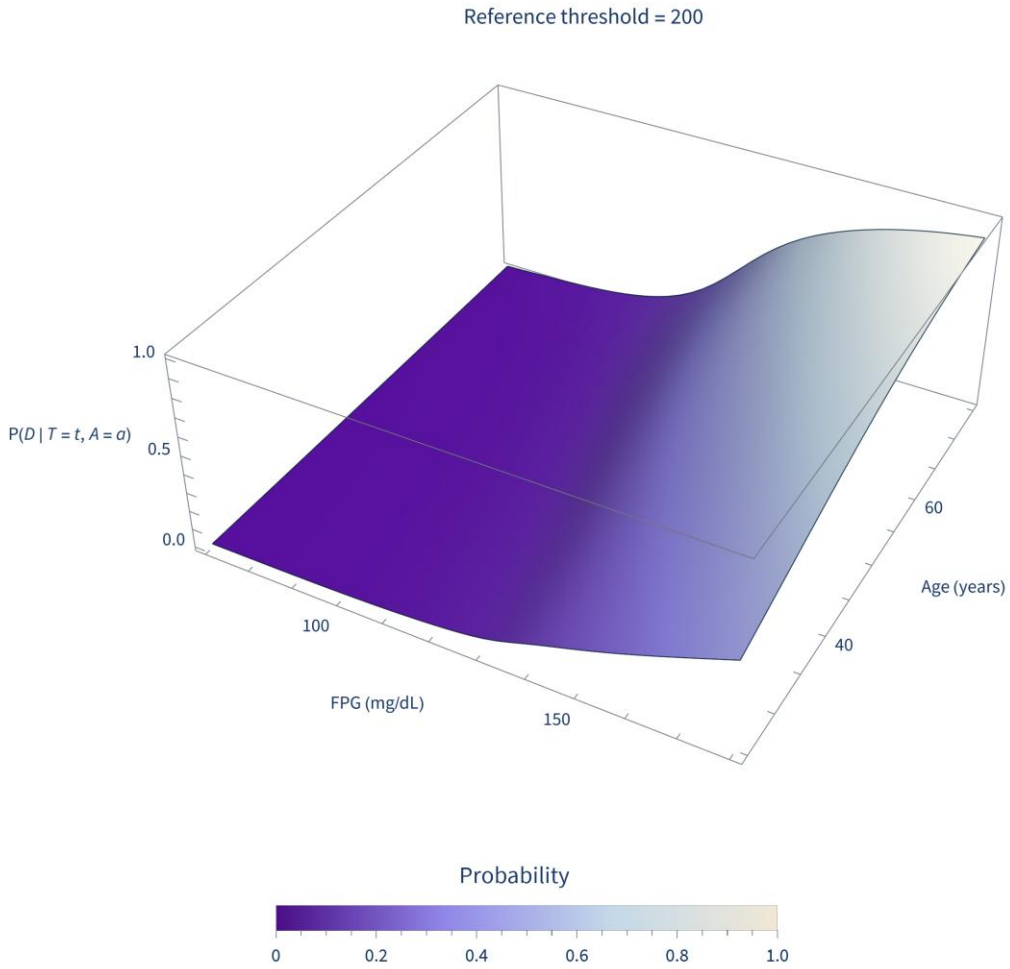


Figure 11. Real-data probability as a function of observed FPG and recorded age, conditional on the saved three-dimensional posterior and latent OGTT threshold 200 mg/dL. The vertical axis and color scale both show probability. The surface displays the posterior mean at each measurement and age. The displayed ages remain within the fitted range.

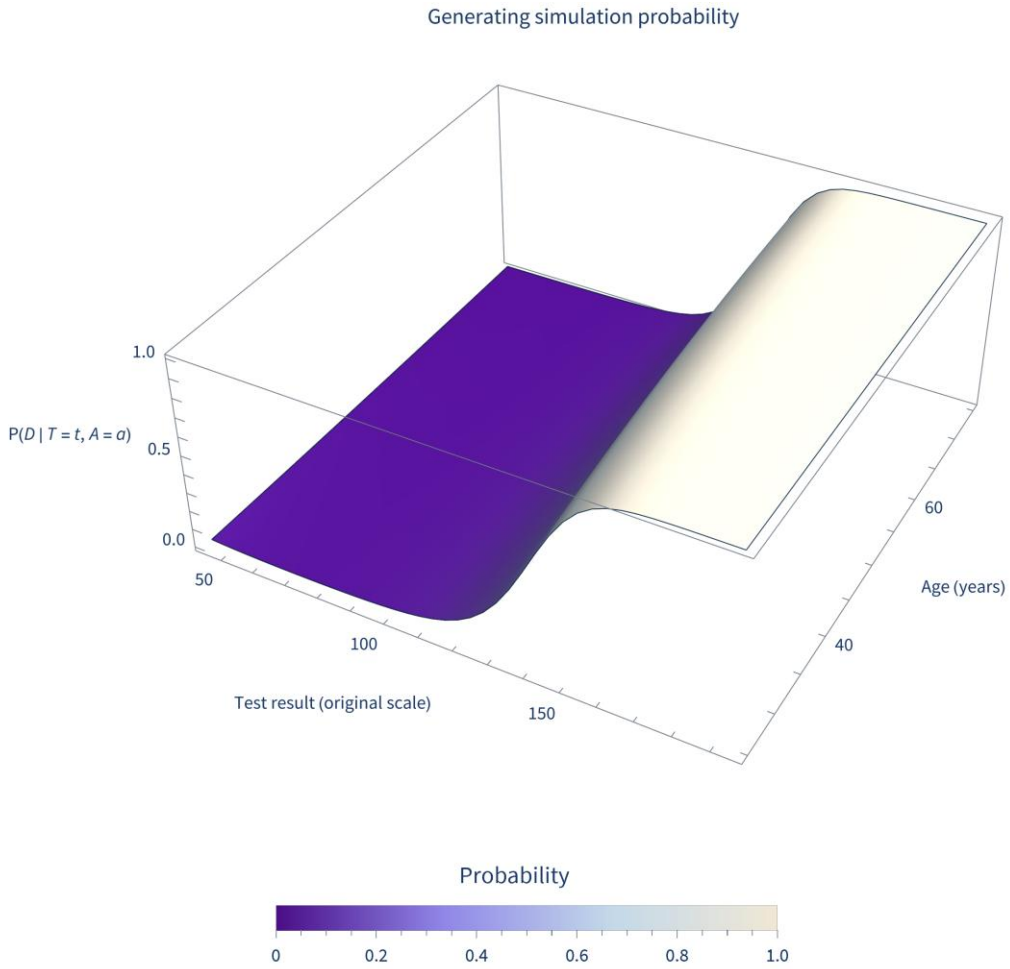


Figure 12. Known generating probability as a function of test measurement and age for the Normal simulation scenario specified for Figures 5 and 6. The surface follows the supplied shifts in class means and prevalence and does not estimate their uncertainty. It is a scenario display, not a fitted age effect from the NHANES data.

4. Discussion

BDPU-Threshold implements a defined inferential framework for interpreting a quantitative diagnostic measurement when population-distribution parameters and analytical measurements are uncertain. The real-data model combines an explicit prior, a likelihood for paired test and reference measurements, and posterior conditioning on a new observed test result. Reference positivity is defined by a threshold on the latent continuous reference measurement, rather than by assigning mixture components to disease classes. This construction permits analytical imprecision in the training reference measurements to affect parameter inference. The contribution demonstrated here is a Wolfram implementation that integrates the stated inferential models, explicit uncertainty definitions, checked numerical evaluation and reproducible applied-result records. The underlying Bayesian conditioning and law of total variance are established principles; their implementation here does not establish comparative methodological superiority or clinical benefit.

The methodological extension is an explicit joint model for the latent test and reference measurements, with posterior uncertainty over its fitted parameters. Numerical integration and the law of total variance define the real-data uncertainty components; parameter and total-Q intervals summarize different conditional distributions. This changes the inferential construction rather than relabelling a first-order confidence interval as a credible interval. It does not establish more accurate uncertainty estimates or better interval coverage. The separate simulation workflow estimates sampling variation by bootstrap refitting while leaving fixed scenario assumptions outside that uncertainty calculation.

The comparison also defines scope boundaries. The kernel-density and two-diagnostic-test analyses described for Bayesian Diagnosis (3) are not general alternatives implemented in *BDPU-Threshold*. The positive and negative predictive values in Bayesian Diagnostic Insights (2) condition on threshold classification, whereas the present calculator conditions on an observed quantitative result. Thus *BDPU-Threshold* extends selected modelling and delivery functions without replacing every predecessor output. No matched-data comparison of these programs, cross-program timing study, usability study or clinical-outcome evaluation was performed; the reported holdout evaluation is an internal comparison of the current models, not a comparison with predecessor software.

4.1.Endpoint and interpretation

At FPG 126 mg/dL, the age-free model gives a probability of 0.2228 that the analytical-error-free, same-visit OGTT concentration is at least 200 mg/dL. The companion calculator in Technical Report XXXII (BDPU-LatentClass) gives 0.4411 for membership in the latent class with the higher mean log FPG at the same observed FPG. The first probability depends on the chosen reference threshold, the second on the fitted class distributions. Both estimates are conditional on the selected cohort and measurement model. An independently confirmed diagnosis would require additional clinical evidence.

4.2.Uncertainty and measurement assumptions

The separation of sampling parameter uncertainty and new-result measurement uncertainty makes the reported uncertainty interpretable, provided that its conditioning is explicit. The parameter component quantifies variation in the integrated reference-positive probability over the posterior parameter distribution after incorporating the evidence from the new observation. It includes the effects of training-measurement imprecision on parameter inference and, in the age model, uncertainty in the fitted age distribution and its associations. The measurement component quantifies the remaining variation in the latent-test-dependent probability within a parameter draw, arising from uncertainty about the analytical-error-free value underlying the new FPG result. The law of total variance combines the two variances; the total standard uncertainty is obtained by taking the square root of their sum. The standard uncertainties themselves are not added arithmetically.

This total uncertainty applies to a conditional probability under the fitted model. It is distinct from the variance of a future binary reference outcome, which would additionally concern outcome variability. The parameter credible interval summarizes variation in the integrated probability, whereas the total-Q credible interval also reflects uncertainty in the latent test value. Their different widths therefore represent different inferential quantities, rather than competing diagnoses. Both intervals are conditional on the assumed distributions and fixed analytical CVs. Neither their numerical construction nor their agreement across implementations establishes frequentist coverage. A relatively narrow interval may coexist with systematic model error, and a smaller standard uncertainty does not demonstrate greater predictive accuracy.

The analytical measurement assumptions remain a material limitation. The case study uses a high-concentration FPG quality-control estimate as a proxy for OGTT analytical imprecision and assumes independent, zero-mean errors on the log scale. Those assumptions determine how observed covariance is separated into latent covariance and analytical-error covariance. A single paired observation per participant cannot identify assay bias, analytical-error correlation or within-person biological variability separately from the modeled population heterogeneity. The reported decomposition also does not integrate uncertainty in the fixed CV estimates, distributional misspecification or transport to another population. These omissions limit the meaning of the term total uncertainty to the specified probability and model. Assay-specific information, repeated measurements and justified covariance assumptions would be required to support a broader uncertainty analysis.

4.3.Age conditioning and model comparisons

Optional age conditioning changes both the statistical model and the information used for prediction. The three-dimensional mixture models the joint distribution of log test concentration, log reference concentration and standardized recorded age. Conditioning on age updates component weights and conditional measurement distributions; it also updates the relative weights of posterior parameter draws through the age likelihood. The subsequent test-result likelihood completes the conditioning. Updating

only the component weights would produce a different inferential calculation. Recorded age is treated as observed without measurement or rounding error, although uncertainty in the fitted distribution involving age remains included through the posterior parameters. The interface restricts predictions to the fitted age range, but the underlying Gaussian age density is not a truncated age model.

The real-data comparison contrasts separately fitted two- and three-dimensional models using the same training measurement pairs, analytical CVs, measurement units, reference threshold and prediction grid. At FPG 126 mg/dL, the age-disabled two-dimensional estimate is 0.2228 with total standard uncertainty 0.0455; the separate three-dimensional model conditioned on age 50 years gives 0.1366 and 0.0395, respectively. These values demonstrate that conditioning and the fitted joint distribution can materially change the calculated probability. Their difference is not an estimated causal age effect: the models have different dimensions and corresponding priors, and the age-disabled comparator is not obtained by marginalizing the fitted age model. The intervals shown for the two curves do not provide an interval for their difference. This fixed-result comparison describes model-dependent predictions. The separate holdout comparison in Table 6 evaluates their measured-reference predictions and did not demonstrate improved calibration from adding age.

4.4.Simulation scope

The simulated-data workflow serves a different purpose from the real-data posterior analysis. Its known generating distributions permit the generating probability to be distinguished from the probability fitted to finite simulated samples. Population prevalence is specified separately from the numbers allocated to the disease and nondisease classes. The parametric bootstrap refits class means and SDs within each replicate and estimates sampling variation in the fitted probability, conditional on the chosen family, prevalence, degrees of freedom and age-effect inputs. It does not add the real-data new-result analytical measurement component or assign posterior distributions to these fixed simulation inputs. The resulting pointwise percentile intervals must consequently be distinguished from the real-data credible intervals. Requiring all requested replicates to be valid at a test value, and retaining unsupported values as gaps, prevents a selectively retained subset from silently determining a reported interval.

The simulated age comparison contrasts the anchor scenario with a scenario defined by supplied changes in class means and prevalence. Its baseline is not a probability averaged over a population age distribution, and its age coefficients are scenario assumptions rather than effects estimated from the NHANES data. Class SDs remain fixed between generating scenarios but are re-estimated in every bootstrap replicate, so their sampling uncertainty is included. The demonstrated identities at zero age effects or at the anchor help verify the implementation. They do not establish coverage of the bootstrap intervals, adequacy of a family for another dataset or a causal interpretation of the scenario changes. Likewise, the shading attached to each simulated curve is an interval for that scenario's fitted probability, not a simultaneous band or an uncertainty interval for the difference between scenarios.

4.5.Presentation and numerical reproducibility

The Tables and Plots hierarchy keeps the primary probability and uncertainty outputs visible while fitting details, age analyses and exports remain in expandable sections. Display choices preserve the applied numerical result. Dark navy labels identify the plot axes and legend entries. Line colors and styles distinguish the probability estimates, interval bounds and uncertainty components. The design is intended to make the calculation target and applied context easier to inspect; no comparative usability study has been performed.

BDPU-Threshold provides two defined workflows: posterior inference from paired real measurements and parametric-bootstrap exploration of simulated classes. The interface presents common inputs and results while preserving their different conditioning and interval meanings. Comparisons on identical posterior arrays, fitted simulation inputs, quadrature and prediction inputs support numerical consistency within the checked tolerances. A shared integer seed alone does not guarantee identical streams across environments. The difference between earlier and shared-input total-Q intervals illustrates prediction-input provenance without changing the probability, standard uncertainties or parameter interval. Appendix B5 reports execution times for posterior prediction.

The displayed total-Q interval endpoint digits identify the retained Monte Carlo realization; they do not establish endpoint precision at every reported decimal place.

4.6.Predictive evidence and further evaluation

The held-out comparison supports a bounded interpretation of the NHANES results. Both models had Brier scores below the constant training-fraction comparator, but adding age did not demonstrate improved calibration. The age model predicted a lower average risk despite the same observed reference-positive frequency; its calibration intercept was more positive and its slope farther above one. The paired Brier and absolute mean calibration-error intervals included zero, so the data do not establish a difference in overall calibration quality. Higher AUROC and lower logarithmic loss were also uncertain and should not be described as established benefits. The 341 participants included only 14 positives, limiting calibration precision. Residual tail misfit and a same-cycle, unweighted holdout restrict generalization. This evaluation concerns measured reference classification, not the latent endpoint, and does not establish clinical validity or comprehensive convergence.

Population scope and reference ascertainment must also remain explicit. The primary likelihood is unweighted and describes the selected cohort of adults meeting the specified eligibility and measurement requirements. The descriptive OGTT-weighted summaries do not convert that posterior into a national survey-design posterior or remove the consequences of additional selection. Predictions in another setting may depend on differences in measurement procedures, reference ascertainment, age distribution and other aspects of case mix not represented by the current model. The absence of a sex covariate and the assumed Gaussian representation of age are additional boundaries of the present extension. A survey-aware or population-transport extension would require its own model specification and evaluation; it cannot be inferred from the current descriptive weighting or interface controls.

The additional age displays make the dependence of a model-based prediction on recorded age visible, but a smooth surface does not establish calibration or biological continuity. A fixed-measurement profile answers a conditional question at that measurement; it does not describe longitudinal change in an individual. Pointwise age intervals do not jointly cover an entire age curve at their nominal level. Simulation profiles remain conditional on fixed age slopes and prevalence inputs. Retaining exact measurements and applied settings, and preserving unsupported regions, is therefore necessary for faithful interpretation of the new displays.

Further evaluation should distinguish numerical verification, statistical performance and clinical usefulness. Numerical work should assess sensitivity to priors, covariance restrictions, analytical CV assumptions, mixture specification, quadrature and Monte Carlo inputs. Repeated synthetic-data experiments with known generating mechanisms could assess estimation error and interval coverage under defined conditions, including support boundaries and deliberate departures from the fitted families. Both models require evaluation on larger independent datasets, with adequate reference-positive observations and explicit assessment of calibration; the present small internal holdout does not establish external validity. Evaluation in additional populations should retain a clearly defined reference endpoint and justified measurement assumptions. Claims about clinical benefit would require a separate assessment of the decisions and consequences associated with using the probabilities. The current evidence supports further methodological research and reproducible software use within the documented scope.

5. Conclusions

BDPU-Threshold provides a Wolfram notebook implementation for estimating probabilities conditional on quantitative measurements and examining their uncertainty. Real-data analysis propagates posterior parameter uncertainty and uncertainty in the latent test value underlying the observation. The variance decomposition, parameter credible interval and total-Q credible interval describe distinct conditional quantities. Simulation separately reports generating and sample-fitted probabilities, with parametric-bootstrap sampling uncertainty under fixed scenario assumptions. The software presents these results through a compact interface with optional age conditioning, detailed diagnostics and reproducible exports.

The optional age extension incorporates recorded age into a joint model and applies both component and parameter-draw age weighting before test-result conditioning. Comparison views display probabilities and uncertainties from matched real-data fits and specified simulation scenarios. Probability surfaces and exact fixed-measurement age profiles make the conditioning explicit over a selected range. Real-data profiles provide parameter and total-Q credible intervals and all three standard uncertainty components; simulated profiles provide pointwise bootstrap intervals and sampling standard uncertainty. Embedded data, callable Wolfram functions and retained calculation records support inspection and reevaluation of the specified computations.

The NHANES illustration demonstrates the computational framework within a selected-cohort analysis with fixed measurement assumptions, residual tail misfit and limited reference-positive holdout evidence. The paired holdout evaluation of both models concerns measured reference classification and did not demonstrate better calibration after adding age. External predictive validity, validity for the latent endpoint and interval coverage remain to be established. Differences between age models or assumed simulation scenarios do not identify causal effects, and reduced uncertainty does not establish improved calibration or clinical benefit. The supported conclusion is that the Wolfram implementation provides a reproducible and numerically checked means of examining defined diagnostic probabilities and their stated uncertainty components. Broader statistical and clinical claims require separate validation against appropriate data, reference endpoints and intended decisions.

6. Supplemental material

The following supplemental files are available for download as a ZIP archive at: https://www.hcsl.com/Supplements/SBDPU-T.zip (accessed on September 22, 2026):

a) Supplemental File I:
BDPU-Threshold.nb: The Wolfram notebook containing the interactive calculator and saved reference fits.
b) Supplemental File II:
BDPU-ThresholdHelper.nb: The helper notebook documents the model, probability estimands, uncertainty decomposition, input validation and native calculation functions.

7. List of abbreviations

AUROC, area under the receiver operating characteristic curve; BDPU, Bayesian Diagnostic Probability and Uncertainty; CI, confidence interval; CV, coefficient of variation; ESS, effective sample size; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin; iid, independent and identically distributed; JSON, JavaScript Object Notation; MCMC, Markov chain Monte Carlo; NCHS, National Center for Health Statistics; NHANES, National Health and Nutrition Examination Survey; NIW, normal–inverse-Wishart; OGTT, oral glucose tolerance test; PSU, primary sampling unit; QC, quality control; SD, standard deviation.

8. Declarations

8.1.Ethics approval and consent to participate

This work uses publicly released, deidentified NHANES data and involved no new recruitment or intervention. Original NHANES data collection was approved by the NCHS Research Ethics Review Board under protocol #2005-06 for 2005–2010 and protocol #2011-17 for 2011–2016, including the primary 2013–2014 cycle (12). The authors confirm that the data collection followed the Declaration of Helsinki and that all participants provided written informed consent. No additional institutional ethics determination was obtained for this secondary computational analysis. The original survey approvals are not presented as a separate approval or exemption for this work.

8.2.Consent for publication

Not applicable. No identifiable individual clinical information or patient images are reported.

8.3.Availability of data and materials

NHANES source components are publicly available from NCHS (4–6,8). The accompanying Wolfram notebooks embed the selected training measurements, reference posterior draws and numerical examples needed by their saved calculations, including the age-conditioned model. The reported 18 October 2026 holdout comparison and its paired-bootstrap intervals are retained historical summaries. Native reevaluation of saved predictions is distinct from regeneration of the archived fitting and bootstrap analysis.

8.4.Competing interests

The authors declare no conflicts of interest.

8.5.Funding

This research received no external funding.

8.6.Authors' contributions

For the contribution statement, TC denotes Theodora Chatzimichail, RCA denotes Rallou A. Chatzimichail and ATH denotes Aristides T. Hatjimihail. Conceptualization, TC and RCA; methodology, TC, RCA and ATH; software, TC, RCA and ATH; validation, TC; formal analysis, RCA and ATH; investigation, TC; resources, ATH; data curation, TC; writing—original draft, RCA; writing—review and editing, ATH; visualization, TC and RCA; supervision, ATH; project administration, RCA. All authors have read and agreed to the published version of the manuscript.

8.7.Acknowledgements

The authors acknowledge assistance from OpenAI Codex with data retrieval and verification scripts, calculator implementation, preparation of supplementary materials, and manuscript drafting and editing. The authors retain responsibility for the scientific claims and interpretation.

8.8.Use of generative artificial intelligence

Codex assisted with retrieval scripts, numerical implementation and verification, consistency checks and computational cross-checks, software interfaces, and document preparation. The statistical-analysis procedural skill informed analysis planning and reporting (13). Independently implemented numerical calculations checked the predictions and calibration estimates; these are computational cross-checks, not independent investigator validation. Earlier coordinated agents also contributed implementation and verification. These activities do not replace the authors' verification of scientific claims. The tools are not authors.

9. References

1. Chatzimichail T, Hatjimihail AT. A Software Tool for Estimating Uncertainty of Bayesian Posterior Probability for Disease. Diagnostics (Basel) [Internet]. 2024 Feb 12;14(4). Available from: http://dx.doi.org/10.3390/diagnostics14040402

2. Chatzimichail T, Hatjimihail AT. A software tool for applying Bayes' theorem in medical diagnostics. BMC Med Inform Decis Mak. 2024 Dec 21;24(1):399.

3. Chatzimichail T, Hatjimihail AT. A Bayesian Inference Based Computational Tool for Parametric and Nonparametric Medical Diagnosis. Diagnostics. 2023 Oct 5;13(19):3135.

4. National Center for Health Statistics. National Health and Nutrition Examination Survey 2013-2014. 2016 [cited 2026 Oct 25]. Oral glucose tolerance test (OGTT_H). Available from: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2013/DataFiles/OGTT_H.htm

5. National Center for Health Statistics. National health and nutrition examination survey 2013-2014. 2015 [cited 2026 Oct 25]. Demographic variables and sample weights (DEMO_H). Available from: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2013/DataFiles/DEMO_H.htm

6. National Center for Health Statistics. National Health and Nutrition Examination Survey 2013-2014. 2015 [cited 2026 Oct 25]. Diabetes (DIQ_H). Available from: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2013/DataFiles/DIQ_H.htm

7. American Diabetes Association Professional Practice Committee for Diabetes*. 2. Diagnosis and classification of diabetes: Standards of care in diabetes-2026. Diabetes Care. 2026 Jan 1;49(Supplement_1):S27–49.

8. Little R. Laboratory procedure manual: Fasting glucose in plasma [Internet]. National Center for Health Statistics; 2013. Available from: https://wwwn.cdc.gov/nchs/data/nhanes/public/2013/labmethods/GLU_H_MET.pdf

9. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with discussion). Bayesian Anal. 2021 June 1;16(2):667–718.

10. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec 16;17(1):230.

11. National Center for Health Statistics. National Health and Nutrition Examination Survey. [cited 2026 Oct 27]. NHANES tutorials: Weighting module. Available from: https://wwwn.cdc.gov/nchs/nhanes/tutorials/weighting.aspx

12. National Center for Health Statistics. National Health and Nutrition Examination Survey. [cited 2026 Oct 27]. NHANES NCHS ethics review board approval. Available from: https://www.cdc.gov/nchs/nhanes/about/erb.html

13. Kassis T, Agarwal V, He Y, Patel D, Brueckner AM. Scientific Agent Skills: A library of procedural knowledge for research agents [Internet]. arXiv [cs.CL]. 2026 [cited 2026 Sept 5]. Available from: http://arxiv.org/abs/2609.00065

10. Appendix A Mathematical details

10.1. A1 Observation model and posterior sampling

This appendix writes the total component covariance as Ω_k, corresponding to T_k in the main text. With observed original-scale test concentration t and reference threshold c, let $y = \log t$ and $g_c = \log c$ denote their log values. Recorded age a is standardized as $z = (a - a_c)/s_a$, where a_c is the fixed centering age and s_a its scale. Concentrations are divided by 1 mg/dL before taking logs in the case study. $X = (F, G)$ denotes the latent log test/reference pair. The latent vector Z_i is (F_i, G_i) in two dimensions and (F_i, G_i, z_i) in three dimensions. The observed vector has dimension $d = 2$ without age and $d = 3$ with standardized age. Components represent heterogeneity, not disease labels. The component count is K, with index $k = 1, \dots, K$ and weight vector $\pi = (\pi_1, \dots, \pi_K)$. The symbols N_d, φ_d and IW_d extend the normal distribution, its density and the inverse-Wishart distribution to dimension d; R_d is the analytical-error covariance in that dimension.

$$W_i = Z_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}_d(0, R_d), \quad \Omega_k = \Sigma_k + R_d, \quad \Sigma_k \succ 0 \quad \text{(A1)}$$

$$R_3 = \mathrm{diag}(\log(1 + {c_F}^2), \log(1 + {c_G}^2), 0) \quad \text{(A2)}$$

The first two diagonal entries give R_2. Zero age error makes the total and latent age variance identical. The Gaussian age density remains untruncated; the prediction-range guard is an interface restriction, not a truncated-likelihood normalization. The selected-cohort likelihood is unweighted and conditionally iid.

$$p(W_i \mid \theta) = \sum_{k=1}^{K} \pi_k \varphi_d(W_i;\, \mu_k, \Omega_k) \quad \text{(A3)}$$

$$\pi \sim \mathrm{Dirichlet}(1, \dots, 1), \quad \Omega_k \sim \mathrm{IW}_d(\nu_0, \Psi_0) \mid (\Omega_k - R_d \succ 0) \quad \text{(A4)}$$

$$\mu_k \mid \Omega_k \sim \mathrm{N}_d\left(m_0, \frac{\Omega_k}{\kappa_0}\right) \quad \text{(A5)}$$

For two dimensions, $m_0 = (\log 100, \log 110)$, $\kappa_0 = 0.1$, $\nu_0 = 4$ and $\Psi_0 = \mathrm{diag}(0.04,0.25)$. For three dimensions, append zero to the two-dimensional prior mean, retain $\kappa_0 = 0.1$, and use $\nu_0 = 5$ and $\Psi_0 = \mathrm{diag}(0.04,0.25,1)$. The fixed centering and scale are 50 and 20 years. Covariance degrees of freedom must exceed $d - 1$. The priors are exchangeable across components; they are not disease-class prevalence priors. For one component, suppress its index and let n be its assigned-observation count, $\bar{w}$ its sample mean and S its scatter matrix. The updated hyperparameters κ_n, ν_n, m_n and Ψ_n have the same roles as in the main text.

$$\kappa_n = \kappa_0 + n, \quad \nu_n = \nu_0 + n, \quad m_n = \frac{\kappa_0 m_0 + n\bar{w}}{\kappa_0 + n} \quad \text{(A6)}$$

$$\Psi_n = \Psi_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{w} - m_0)(\bar{w} - m_0)^{\mathrm{T}} \quad \text{(A7)}$$

Gibbs allocation probabilities are proportional to $\pi_k \varphi_d$; weight updates use the Dirichlet conditional with counts added to its unit concentrations. Covariance candidates are redrawn until $\Omega_k - R_d$ is positive definite, followed by the conditional Normal mean update. Finite chains approximate the restricted posterior; common integer seeds do not imply identical independently generated chains across environments.

10.2. A2 Conditioning on recorded age

For each posterior draw, partition its mean and total covariance into the log measurement pair X and standardized age coordinate z. The age likelihood L_A is a mixture density in standardized age; its constant original-age Jacobian cancels in normalized component and parameter weights. The block labels X and A select the measurement-pair and standardized-age coordinates; Ω always denotes observed covariance and Σ latent covariance. The scalar normal density φ is parameterized by its mean and variance. The weights $\omega_k(a, \theta)$ are component probabilities conditional on age; a vertical bar in block subscripts denotes age conditioning for fixed θ.

$$L_A(\theta, a) = \sum_{k=1}^{K} \pi_k \varphi(z;\, \mu_{A,k}, \Omega_{AA,k}) \quad \text{(A8)}$$

$$\omega_k(a, \theta) = \frac{\pi_k \varphi(z;\, \mu_{A,k}, \Omega_{AA,k})}{L_A(\theta, a)} \quad \text{(A9)}$$

$$\mu_{X|a,k} = \mu_{X,k} + \Omega_{XA,k}{\Omega_{AA,k}}^{-1}(z - \mu_{A,k}) \quad \text{(A10)}$$

$$\Omega_{X|a,k} = \Omega_{XX,k} - \Omega_{XA,k}{\Omega_{AA,k}}^{-1}\Omega_{AX,k}, \quad \Sigma_{X|a,k} = \Omega_{X|a,k} - R_2 \quad \text{(A11)}$$

The likelihood $L_T(\theta, y|a)$ is the observed-log-test density at y conditional on age and model parameters.

$$L_T(\theta, y \mid a) = \sum_{k=1}^{K} \omega_k(a, \theta) \varphi(y;\, \mu_{F|a,k}, \Omega_{FF|a,k}) \quad \text{(A12)}$$

$$p(\theta \mid \mathcal{D}, y, a) \propto L_A(\theta, a) L_T(\theta, y \mid a)\, p(\theta \mid \mathcal{D}) \quad \text{(A13)}$$

With equal-weight saved posterior draws, normalized parameter weights are proportional to the product of these two likelihood factors. Both the age factor and the new-test factor are required. For original-scale test density, division by t supplies the log-transform Jacobian; that constant also cancels from normalized parameter weights. In the age-disabled model the age likelihood is absent. Taking the first two coordinates of a fitted $3D$ model without conditioning would be a different comparator from the separately fitted $2D$ model used in the interface.

10.3. A3 Latent measurements and uncertainty

For component k after age conditioning, the latent test given observed y has mean $m_{F,k}$ and variance $v_{F,k}$ below. Its component weights are proportional to $\omega_k(a, \theta)$ times the component observed-test density. The zero-test-CV case gives a point mass at y and zero new-result measurement variance. The arguments y, a and θ are suppressed in $m_{F,k}$ and $v_{F,k}$; R_{FF} is the FPG variance entry of R_2.

$$m_{F,k} = \mu_{F|a,k} + \frac{\Sigma_{FF|a,k}}{\Omega_{FF|a,k}}(y - \mu_{F|a,k}) \quad \text{(A14)}$$

$$v_{F,k} = \frac{\Sigma_{FF|a,k} R_{FF}}{\Omega_{FF|a,k}} \quad \text{(A15)}$$

For latent test f, the component latent-reference conditional mean and variance use Σ rather than Ω. At that f, mixture weights must be recomputed from the latent-test density; retaining observed-test weights in this inner probability would be incorrect. Write these moments as $m_{G,k}(f)$ and $v_{G,k}$, with conditioning on a and θ understood. The age-conditioned reference-positive probability is $q_\theta(f,a) = \Pr(G \geq g_c | F = f, a, \theta)$, extending $q(\theta, f)$, and Φ is the standard-normal distribution function. Its component weights $\lambda_k(f,a,\theta)$ are proportional to $\omega_k(a,\theta)\varphi(f; \mu_{F|a,k}, \Sigma_{FF|a,k})$, normalized over components.

$$m_{G,k}(f) = \mu_{G|a,k} + \frac{\Sigma_{GF|a,k}}{\Sigma_{FF|a,k}}(f - \mu_{F|a,k}) \quad \text{(A16)}$$

$$v_{G,k} = \Sigma_{GG|a,k} - \frac{\Sigma_{GF|a,k}{}^2}{\Sigma_{FF|a,k}} \quad \text{(A17)}$$

$$q_\theta(f,a) = \sum_{k=1}^{K} \lambda_k(f,a,\theta)\left[1 - \Phi\left(\frac{g_c - m_{G,k}(f)}{\sqrt{v_{G,k}}}\right)\right] \quad \text{(A18)}$$

The integrated component reference tails for observed y provide the analytic first moment of q. The corrected implementation integrates its centered conditional variance over latent F; the initial Gauss–Hermite rule has 32 nodes, with the convergence and fallback checks specified below. Let M be the number of retained parameter draws $\theta^{(h)}$, indexed by $h = 1, \ldots, M$. The normalized weights ρ_h are proportional to $L_A(\theta^{(h)}, a) L_T(\theta^{(h)}, y|a)$. The quantities $\bar{q}_h$ and V_h are the mean and variance of $Q = q_\theta(F,a)$ over F conditional on y, a and $\theta = \theta^{(h)}$. The following sums estimate posterior moments conditional on the training data and the new observations; $p_D(t,a)$ is the age-conditioned posterior reference-positive probability at $y = \log t$, abbreviated p_D in the variance formula.

For each retained draw, the numerical target is $V_h = \mathrm{E}[(Q - \bar{q}_h)^2 | y, a, \theta = \theta^{(h)}]$, using the analytic conditional mean $\bar{q}_h$. An initial n-node Gauss–Hermite rule, with default $n = 32$, is checked against min (256, max (64,2n)) nodes. The absolute and relative tolerances are 2×10^{-11} and 2×10^{-8}, and both numerical first moments are checked against the analytic mean. Sharp component-tail or mixture-weight transitions require adaptive integration even if the two fixed rules agree.

Adaptive integration uses the standardized conditional-normal coordinate over $[-12,12]$, whose omitted total normal mass is below 2×10^{-33}, with breakpoints resolving component-tail and mixture-weight transitions. Wolfram NIntegrate uses GlobalAdaptive with symbolic processing disabled, AccuracyGoal 12, PrecisionGoal 10 and MaxRecursion 25. A stricter pass uses goals 13 and 11 with MaxRecursion 28. Failed first-moment, agreement or variance-bound checks return an explicit failure.

Each conditional variance must satisfy $0 \leq V_h \leq \bar{q}_h(1 - \bar{q}_h)$, and the final variance must satisfy $0 \leq \mathrm{Var}(Q) \leq p_D(1 - p_D) \leq 1/4$, with numerical bound tolerance 5×10^{-11}. These bounds are verification conditions, not a clipping operation or sufficient evidence of integration accuracy by themselves.

$$p_D(t,a) = \sum_{h=1}^{M} \rho_h \bar{q}_h \quad \text{(A19)}$$

$$u_{\mathrm{s}}{}^2 = \sum_{h=1}^{M} \rho_h (\bar{q}_h - p_D)^2 \quad \text{(A20)}$$

$$u_{\mathrm{m}}{}^2 = \sum_{h=1}^{M} \rho_h V_h, \quad u_{\mathrm{total}}{}^2 = u_{\mathrm{s}}{}^2 + u_{\mathrm{m}}{}^2 \quad \text{(A21)}$$

The total is posterior uncertainty of the conditional probability Q, not the variance of a future Bernoulli outcome. The parameter term includes uncertainty from fitting the entire measurement model, including the age distribution. Bias, biological variation, uncertainty of fixed CVs and age measurement/rounding error are not added by this decomposition. Disease and nondisease SDs are equal; complementary interval bounds are one minus the reversed disease bounds. Relative uncertainty divides the common SD by the selected probability and is undefined at zero.

10.4. A4 Quantiles and numerical inputs

The parameter interval uses weighted quantiles of $\bar{q}_h$. The total-Q interval uses one latent-test draw per posterior draw, selected from its conditional mixture using one Uniform and one standard-Normal input. The stored prediction tape is therefore an M by 2 array. Weighted quantiles discard zero-weight values, renormalize the remaining weights, sort their values, form cumulative weight positions at the centers of weights, and interpolate at the requested tail probabilities. This midpoint convention differs from the

linear unweighted bootstrap quantiles used for simulation. In (A22), i and j index the sorted values, ρ_i and ρ_j are their reordered weights, and ξ_j is the cumulative-weight midpoint for $j = 1, \ldots, M$. The importance-weight effective sample size is denoted ESS.

$$\xi_j = \sum_{i=1}^{j} \rho_i - \frac{\rho_j}{2}, \quad \text{ESS} = \frac{1}{\sum_{h=1}^{M} \rho_h{}^2} \quad \text{(A22)}$$

The ESS in (A22) is importance-weight ESS and does not adjust for chain autocorrelation. The quantile formula refers to weights after sorting by their associated values; endpoint requests are bounded by the first and last sampled values. Probability and SD calculations use likelihood-weighted moments and quadrature, so changing only the total-Q prediction tape does not alter them. Identical posterior arrays and identical prediction inputs are required for an exact quantile comparison. Earlier and shared archived seed-555 streams contain different prediction inputs; the controlled check described in the Results accounts for their different total-Q intervals.

Only positive-weight values define the interpolation support and its endpoints. The quantile construction in (A22) is restricted to these retained indices; zero-weight draws remain part of the saved posterior but contribute no probability mass.

10.5. A5 Simulation distributions and bootstrap

Each generating class is specified by original-scale mean m and SD s. Normal uses mean m and SD s. Lognormal uses log variance $\log(1 + s^2/m^2)$ and log mean $\log m$ minus half that variance. Gamma uses shape m^2/s^2 and scale s^2/m. Student t uses location m, fixed degrees of freedom $\nu > 2$ and scale $s\sqrt{(\nu - 2)/\nu}$. Uniform has endpoints $m \pm \sqrt{3}s$. Positive m is required for Lognormal and Gamma; every family requires positive SD. Moment refitting uses the sample mean and sample SD for class sample size n, with denominator $n - 1$; Student t degrees of freedom are not estimated. At a fixed scenario, r is the population disease prevalence, and $f_D(t)$ and $f_H(t)$ are the test-measurement densities for disease and nondisease at observed t; subscripts D and H label those simulation classes. Here $p_D(t)$ denotes the probability of the simulated disease class conditional on the test value.

$$p_D(t) = \frac{r f_D(t)}{r f_D(t) + (1 - r) f_H(t)} \quad \text{(A23)}$$

The known probability uses generating densities; the reported fitted point estimate uses densities moment-fitted to the original simulated samples. It is not the bootstrap mean. Zero total evidence density makes a probability undefined, including the space between disjoint Uniform supports. Allocation counts affect precision and fitted densities, not r. For class $j \in \{D, H\}$, $m_j(a)$ is its mean at age a, $m_{j,0}$ its mean at simulation anchor age a_0, and b_j its mean change per decade. The prevalence at age a is $r(a)$, its anchor value is r_0, and β is the log-odds change per decade. The simulation anchor a_0 and the real-data centering age a_c have distinct roles.

$$m_j(a) = m_{j,0} + b_j \frac{a - a_0}{10}, \quad \operatorname{logit} r(a) = \operatorname{logit} r_0 + \beta \frac{a - a_0}{10} \quad \text{(A24)}$$

The age slopes, anchor age, family and class SDs are fixed scenario inputs. A nonzero log-odds slope requires prevalence strictly between zero and one in comparison mode, including at the anchor; the preserved single-scenario engine retains its identity shortcut at the anchor or with all zero slopes. Mean shifts must also satisfy the chosen family's support and moment restrictions. Generate at the target age once, fit those samples, and do not shift the fitted models again. For B valid bootstrap replicates at a fixed test value and scenario, $b = 1, \ldots, B$ indexes the replicate fitted probabilities p_b, and $\bar{p}$ is their arithmetic mean. Their sample standard deviation is u_{boot}.

$$u_{\text{boot}} = \sqrt{\frac{\sum_{b=1}^{B} (p_b - \bar{p})^2}{B - 1}} \quad \text{(A25)}$$

The bootstrap mean $\bar{p}$ is distinct from the original sample's fitted point estimate. For B completely valid bootstrap probabilities, the sample SD above and the linear-interpolated unweighted percentile interval are reported. Every bootstrap re-estimates both class means and SDs from parametric samples of the original class sizes. Reporting requires B at least 100 and all B probabilities valid at the test value. The grid combines both scenarios' boundary, interior and adjacent support probes; uncertainty or interval

failures remain gaps. Fixed prevalence, family, degrees of freedom and age-slope uncertainty are excluded. Matching seeds are not evidence of frequentist coverage or an interval for the difference between scenarios.

10.6. A6 Probability surfaces and fixed measurement age profiles

The real-data surface evaluates the probability in (A19) at each supplied test measurement and age, using the age and test likelihood factors in (A8)–(A13). The integrated reference tail supplies its probability mean directly; a probability-only surface does not require a latent-test prediction tape or an uncertainty calculation. For a fixed-measurement profile, the same measurement is retained at every age and the full calculation in (A19)–(A22) is repeated. The variance components and weighted quantile intervals therefore have the same definitions as the single-result calculation. The historical output label Variance closure reports the sum of the component variances, rather than a residual; verification separately compares that sum with the squared total SD.

A simulation surface applies the age transformation in (A24) once at each age, then evaluates (A23) using either generating or original-sample moment-fitted class parameters. A simulated uncertainty profile additionally repeats the bootstrap calculation in (A25) at each age, using the recorded class counts and seeds. These calculations do not integrate uncertainty in the age coefficients, prevalence, family choice or degrees of freedom. Complete-support and minimum-replicate requirements apply separately at every age. Each interval is pointwise; neither a surface nor a collection of age intervals constitutes a simultaneous uncertainty region.

11. Appendix B Programming and reproducibility details

11.1. B1 Modules and data contracts

The maintained source of record is the Wolfram code embedded in *BDPU-Threshold.nb* and *BDPU-ThresholdHelper.nb*, release 1.0.0. The frozen reference posteriors and archived case-study summaries are retained inputs. The helper documents the functions and provides runnable examples of the reference estimate, age conditioning, a custom refit and simulation. Either notebook can open the calculator. Table B1 lists the native interfaces. The distinction between evaluating a saved posterior and fitting a new posterior is explicit.

Table B1. Wolfram functions and numerical contracts

Responsibility	Wolfram function or expression	Input and output contract
Real fitting	`BayesianDiseaseFit`	Positive paired measurements to posterior fit Association
Probability	`BayesianDiseaseProbability`	Posterior, observed result and threshold to probability and uncertainty
Interface	`BDPUOpenCalculator[]`	Interactive calculations and applied result views
Age inference	`BayesianAgeDiseaseFit and BayesianAgeConditionFit`	Three-column data; conditioned samples and age log weights
Simulation	`BayesianSimulationUncertainty`	Bootstrap replicates and an observed value yield a fitted probability, SD and pointwise limits
Reproducibility	`Put, Get and FileHash`	Retained Wolfram expressions and content identity

'Import CSV and fit' accepts a header-free numeric file with positive continuous test and reference measurements in the first two columns. To fit an age model, supply age in years in the third column. The

fit requires at least ten finite rows, age variation and ages between 0 and 120 years. Predictions are restricted to the fitted training-age range. The embedded age-disabled training matrix has shape (1361,2). An age-enabled matrix for those same participants would have shape (1361,3), but the notebooks embed the age-conditioned posterior rather than the raw training-age column. Custom paired comparisons require the user to supply matching three-column data, fitted CVs, labels and units.

Posterior arrays have shapes (M, K), (M, K, d) and (M, K, d, d) for weights, means and total covariances. The Wolfram fit is an Association containing a list of posterior sample Associations and the analytical CVs, together with model metadata. Weights are normalized; total and implied latent covariances must satisfy the stated positive-definiteness restrictions. The age fit retains its center, scale, training range and priors. Age conditioning also produces a log age-likelihood weight for each parameter draw; these weights must be retained in any saved conditioned model.

11.2. B2 Inference and comparison algorithms

The Wolfram kernel packs compatible machine-real arrays, checks them in bulk and evaluates posterior draws in batches using compiled Wolfram virtual-machine code. Prediction inputs follow draw order. The calculation reuses component slopes and scales across quadrature nodes and applies the integration checks in Appendix A.

LOAD posterior arrays, analytical CVs, metadata, quadrature and tape

IF age is enabled:

  CHECK fitted age range

  FOR each parameter draw:

    EVALUATE age log likelihoods for all components

    NORMALIZE component age weights

    COMPUTE conditional mean and covariance by Schur complement

  STORE parameter draw log age likelihoods

FOR each observed test value:

  COMBINE parameter age and test log likelihoods; NORMALIZE weights

  EVALUATE analytic reference tails and checked centered latent-test variances

  COMPUTE weighted parameter and sampled-Q quantile intervals

  RETURN probability, parameter SD, measurement SD, total SD and intervals

The calculation conditions the fit once per requested age. A session cache reuses up to 16 conditioned fits and two curves when the full model and relevant inputs match exactly, including curve options and any supplied prediction tape. The baseline comparison uses the separate saved $2D$ fit, not the marginalized $3D$ fit. Log-sum-exp normalization avoids direct multiplication of small likelihoods. Cholesky/eigenvalue checks enforce the covariance restrictions. The real default grid has 41 points over an increasing positive range; density displays include the original-scale log-transform Jacobian.

CHECK matching real rows, fitted CVs, units and labels

EVALUATE 2D and age-conditioned 3D curves on the same grid

OR, for simulated data:

  VALIDATE anchor settings and supplied age effects

  GENERATE and bootstrap anchor scenario using the selected seeds

  GENERATE and bootstrap target-age scenario using the same seeds/counts

REUSE the identical scenario when settings do not change

FORM union of both support-aware grids

EVALUATE each scenario independently, retaining undefined gaps

COMMIT comparison only after successful calculation

RETAIN its applied age, threshold, level, settings and display provenance

The main calculator opens with plot bounds of 70 and 180. The standalone simulation interface starts at 45 and 195. The simulation grid divides the selected range into 60 equal intervals, giving 61 base-grid points. The plot bounds are not inferred from Normal or Student t tails. Within these bounds, critical points comprise the bounds themselves and every finite endpoint of the generating and original-sample fitted class supports and the complete bootstrap evidence support. The latter is the intersection across all bootstrap replicates of the union of class supports with nonzero prevalence weight; a failed replicate makes it empty.

Let the sorted distinct critical points be $c_0, \ldots, c_J$, with lower and upper bounds L and H, and let $N = 60$. Add every adjacent midpoint and, for each interior critical point c_j, probes at $c_j - \delta_j$ and $c_j + \delta_j$, where

$$\delta_j = \min\left((H - L)/1000N\,, (c_j - c_{j-1})/8\,, (c_{j+1} - c_j)/8\right).$$

The final grid is the sorted unique union of the base grid, critical points, midpoints and probes. A comparison uses the union of both scenario grids. Probability and SD gaps are not linearly bridged, and failed replicates are counted rather than silently dropped.

11.3. B3 Interface state and exports

The ordinary calculation, age comparison, probability surface and fixed-measurement age profile retain separate results and applied settings. Controls distinguish pending inputs from the successful calculation being displayed. Outcome selection and relative-uncertainty display operate on the retained result. An age profile records its exact measurement, age grid, reference threshold, interval level and model or simulation context. These records prevent a display change from silently redefining the completed calculation.

Custom real-data comparisons use matched measurement rows and compatible fitted analytical CVs and units. The paired fitting action prepares both the age-disabled and age-conditioned models before committing a new pair. A short interactive fit is a user-generated posterior sample; completion of the action does not give it the validation status of the archived case study.

Native export actions write numerical CSV and JSON records and PNG figures for the available calculation views. Age-profile and surface records retain their applied grids and context. Wolfram expressions can also be saved with Put and read with Get when exact model structure and prediction inputs must be retained. Export format availability is stated by the corresponding notebook action; a static image does not itself preserve the numerical model.

For reproducible reevaluation, retain the fit Association, analytical CVs, reference threshold, observed test value, interval level, quadrature settings and exact prediction inputs. An age-conditioned record must additionally preserve the fitted age metadata and parameter-draw age likelihoods. A simulation record must retain the generating settings, class counts, random seeds, fitted models and bootstrap settings. FileHash can identify saved file content; a matching hash establishes identity of those bytes, not model adequacy, interval coverage or clinical validity.

The uncertainty tables use the retained calculation and its applied measurement, age, reference threshold and interval level. Standard and relative standard uncertainty are displayed separately, with a point estimate followed by the uncertainty components. The interval-limits view selects Parameter credible or Total-Q credible for real data and presents the point estimate and separate bounds. Simulation intervals retain their pointwise parametric-bootstrap interpretation. Relative uncertainty is undefined at zero probability; probability estimates and bounds remain absolute, and undefined values remain missing. The

main probability and uncertainty tables follow the selected outcome. Exports retain full precision and applied settings. For real data, total standard uncertainty refers to the conditional probability Q.

11.4. B4 Presentation hierarchy and typography

The calculator displays the selected probability above 'Tables' and 'Plots'. It reports prior and posterior probabilities as decimals on the 0–1 scale. The main interval input and result caption express the interval level as a percentage. Result headlines, tables and reference checks use regular-weight text. Plot labels use regular-weight dark navy text for emphasis. Uncertainty labels identify absolute SD, relative SD or interval endpoints and their interpretation. This organization is intended to reduce simultaneous choices; no measured usability advantage is claimed.

Plot titles and axis labels identify the quantity, and legends identify the curves. A probability profile gives its interval type and level. An uncertainty profile identifies the standard uncertainty shown. Sampling-parameter, new-result measurement and total uncertainty use orange, blue and gold. Age-comparison curves distinguish the baseline by an orange dashed line and the age-conditioned result by a blue solid line. Legends remain outside the data region. Tables use readable labels and aligned numeric columns. Probability surfaces use a scale bounded by zero and one.

11.5. B5 Reproducibility inputs and validation

Each archived posterior contains 6,000 draws from four chains of 6,500 iterations, with 2,000 burn-in iterations and thinning by three. Seeds were 4101–4104 for the 2D model and 6101–6104 for the age model. The default simulation uses seed 20261010, 500 observations per class and 200 bootstrap replicates with seed 20261011. For strict numerical comparison, posterior arrays, simulated samples, fitted/bootstrap model parameters, quadrature nodes/weights and total-Q prediction inputs are frozen and shared as appropriate.

Save the full-precision posterior, model and prediction inputs with `$Version`, `$SystemID` and the notebook hash. `BDPUClearCalculationCache[]` clears stored curves and conditioned fits. `BDPUCalculationCacheInformation[]` reports the compiled backend and cache activity.

An earlier benchmark measured age-disabled predictions after initialization using 6,000 draws in Wolfram 15.0.1 for 64-bit Windows on a system reporting eight processors (OGTT threshold 200 mg/dL, 95% interval). At FPG 126 mg/dL, time for the first prediction fell from 2.45 to 0.38 s. Computing the 41-point curve over 70–180 mg/dL took 98.25 s before optimization and 14.41 s after optimization. Reusing the completed curve took 0.43 s. In a separate comparison, the uncached 41-point curve took 13.88 s in the preceding implementation and 10.43 s in the current one. The 41 results agreed exactly.

In the earlier verification, across 124 result rows in 29 test cases, probabilities, standard uncertainties and interval limits matched the reference exactly. For 82 age-curve rows from eight draws and two prediction tapes, maximum absolute differences were 1.11×10^{-16} for probability and 1.39×10^{-17} for standard uncertainty. All interval limits agreed. The diagnostic and holdout summaries refer to the archived analysis.

Additional identities include zero new-test analytical variance, variance closure, disease/nondisease complement relations, exact simulation identity at zero effects or the anchor, unchanged class samples under a prevalence-only effect, independent-age reduction for a specially constructed posterior, and unsupported Uniform gaps. Independent-age reduction requires a common age density across components and parameter draws as well as zero age cross-covariances; zero cross-covariance alone is insufficient.

11.6. B6 Notebook execution and reproducibility

Extract SBDPU-T.zip to obtain BDPU-Threshold.nb and BDPU-ThresholdHelper.nb. The main notebook contains the interactive calculator. The helper provides the methods and examples. Select 'Start analysis' in either notebook. For the reference example, choose 'Real data' and 'Disease', enter FPG 126 mg/dL and 95 for a 95% interval, then select 'Calculate probabilities'. With age conditioning off and the OGTT threshold at 200 mg/dL, the calculator shows a reference-positive probability of 0.2228.

To run code in a fresh kernel, select 'Load functions for examples' in the helper. Select a worked example and press Shift+Enter. BayesianDiseaseProbability evaluates a fitted model at a specified test value and reference threshold. BayesianAgeDiseaseProbability uses the joint fit and an age within the training range. The callable functions take the credible level as a fraction.

The following example evaluates the embedded age-free reference fit and saves the fit and result together.

```
fit = nfBayesDefaultFit;

result = BayesianDiseaseProbability[fit, 126., "ReferenceThreshold" -> 200.,
"CredibleLevel" -> 0.95];

Put[<|"Fit" -> fit, "Result" -> result, "WolframVersion" -> $Version|>,
"BDPUCalculation.wl"];

saved = Get["BDPUCalculation.wl"];

FileHash["BDPUCalculation.wl", "SHA256"]
```

12. Software availability and requirements

Software name: BDPU-Threshold

Version: 1.0.0

DOI: 10.5281/zenodo.24443936

Project home page: https://www.hcsl.com/Tools/BayesianDiagnosticUncertaintyThreshold/index.html

Computational source notebook: BDPU-Threshold.nb. Available at: https://www.hcsl.com/Tools/BayesianDiagnosticUncertaintyThreshold/BDPU-Threshold.nb

Programming language: Wolfram Language

Operating systems: Microsoft Windows 10+, Linux 4.18+, Apple macOS 13+

Other software requirements: Wolfram Player (freely available) or Wolfram Mathematica.

Hardware requirements: An x86-64-v2 or later CPU, or an Apple Silicon processor, with at least 4 GB of RAM is recommended.

Software citation: Chatzimichail T, Chatzimichail RA, Hatjimihail AT. *BDPU-Threshold: Estimating Conditional Disease Probability, Decomposing Uncertainty, and Simulating Test Results*. Ver. 1.0.0. Hellenic Complex Systems Laboratory; 2026. https://www.doi.org/10.5281/zenodo.24443936

License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The notebook carries its numerical code and embedded reference data; the software does not require a separate application server.

The embedded case-study resources include saved NHANES fitted posteriors, selected training data, numerical summaries and the archived held-out comparison. These resources retain their historical provenance. Reopening the notebooks or evaluating the calculator does not regenerate the four-chain reference fits or the holdout bootstrap analysis.

The calculator displays the inputs used for the last successful calculation alongside the results. Separate controls open model fitting, age analyses and exports.

The Wolfram source and scientific documentation are intended for inspection and noncommercial research use under those terms.

13. Citation

Chatzimichail T, Chatzimichail RA, Hatjimihail AT. *Probability and uncertainty estimation for a reference concentration threshold*. Technical Report XXXI. Hellenic Complex Systems Laboratory; 2026. https://www.doi.org/10.5281/zenodo.24445993

14. License

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

First Published: November 22, 2026