

# Bayesian Decision Rules Under Asymmetric Loss: A Parkinson’s Disease Case Study and Simulation Analysis

Om Lala  
Whitney High School, Cerritos, CA

## Abstract

A probabilistic classifier does not determine an action until prediction is combined with a loss function. This paper studies that distinction through a synthetic Parkinson’s disease case study and a controlled simulation analysis. Using 2105 synthetic records, we compare maximum-likelihood, ridge, lasso, and Bayesian logistic regression under repeated stratified cross-validation, evaluating discrimination, calibration, proper scoring rules, and decision loss rather than accuracy alone. For false-positive cost  $c_{FP}$  and false-negative cost  $c_{FN} = \lambda c_{FP}$ , the Bayes action refers an observation when its calibrated risk exceeds  $\tau^*(\lambda) = 1/(1 + \lambda)$ . At  $\lambda = 3$ , this gives  $\tau^* = 0.25$ ; the out-of-fold Bayesian model yielded sensitivity 0.959 and specificity 0.442, compared with 0.856 and 0.704 at the conventional 0.50 cutoff. Removing UPDRS and the cardinal motor symptoms, which are components of the diagnostic definition rather than independent measurements, drops ROC AUC from 0.881 to 0.668 while leaving calibration intact. We further propagate posterior uncertainty into referral decisions and study threshold regret over prevalence–cost combinations. A simulation with known data-generating probabilities tests when Bayesian shrinkage changes calibration and decision regret relative to maximum-likelihood and ridge estimators. The emergent result is that threshold choice is a decision problem conditioned on probability quality, target prevalence, and relative error costs. This is not strictly a property of the classifier.

**Keywords:** Bayesian decision theory; asymmetric loss; Bayesian logistic regression; calibration; screening; decision regret; synthetic data; uncertainty quantification

## 1 Introduction

Binary prediction is often presented as if estimating a probability and making a decision were the same problem. They are not. A statistical model estimates a risk such as

$$\pi(x) = P(Y = 1 \mid X = x),$$

whereas a decision rule maps that risk to an action. The distinction becomes important whenever the consequences of false positives and false negatives differ, such as in diagnoses of Parkinson’s Disease. A 0.50 cutoff, while being intuitive, is not a methodologically backed default; it is optimal only under a particular combination of calibrated probabilities and symmetric classification costs.

Parkinson’s disease (PD) provides a useful motivating example because screening and referral decisions

naturally involve asymmetric consequences. Missing a potentially relevant case can delay specialist evaluation, while a false-positive screen more often creates additional evaluation and resource use than immediate treatment. At the same time, clinical PD prediction is a difficult domain in which symptoms used as predictors may overlap with the process used to define the diagnosis itself [1, 3, 2]. A statistically careful analysis must therefore distinguish prediction, calibration, and utility from claims about biological risk factors or clinical deployment.

The dataset being synthetic shapes much of the direction of the paper. We do not claim this framework to be sufficient under clinical trials. Instead, we look in an alternative direction.

*How do calibration, posterior uncertainty, prevalence, and asymmetric error costs interact when a probabilistic classifier is converted into a screening decision?*

The paper aims to accomplish four different, yet intersecting, tasks. It evaluates classical and Bayesian logistic models under repeated out-of-sample validation, reporting calibration and proper scoring rules next to discrimination. It derives the Bayes action under an explicit false-negative to false-positive cost ratio and evaluates that action on held-out predictions. It separates the posterior-mean Bayes action from uncertainty about that action across posterior draws, which are different questions and are often conflated. And it uses a controlled simulation with known ground-truth probabilities to ask when, if ever, Bayesian shrinkage changes decision regret relative to maximum-likelihood and ridge estimators.

## 2 Background and Statistical Motivation

### 2.1 Literature Review

PD is a progressive neurodegenerative disorder with heterogeneous motor and nonmotor manifestations [1, 5]. Bradykinesia, rest tremor, and rigidity are central to the clinical syndrome, while cognition, sleep, autonomic symptoms, gait, and functional impairment can also contribute to evaluation [4]. The Movement Disorder Society diagnostic criteria use combinations of motor findings, supportive criteria, exclusion criteria, and red flags rather than a single decisive biomarker [3].

The MDS-UPDRS likewise measures multiple dimensions of PD-related impairment [6]. Consequently, a model that predicts a recorded PD label using UPDRS or cardinal motor symptoms can exhibit strong apparent predictive performance partly because these features resemble information involved in defining the label. In the present synthetic dataset, the problem is rather elementary: the fitted associations describe the synthetic data-generating mechanism, not biological effect sizes. Section 6.1 quantifies how much of the model’s apparent performance rests on these four variables.

### 2.2 Prediction quality, calibration, and decisions

Logistic regression estimates conditional probabilities while retaining a transparent likelihood and coefficient interpretation [7]. Alternative links such as probit and complementary log-log regression

provide useful model-form comparisons, but a model’s predictive utility cannot be summarized by in-sample likelihood criteria alone. Out-of-sample evaluation must separate at least three questions.

**Discrimination** asks whether higher-risk observations tend to receive higher scores. ROC AUC and precision–recall AUC are ranking measures. **Calibration** asks whether numerical predictions agree with empirical frequencies. A perfectly discriminating model can still be badly calibrated. **Probabilistic accuracy** can be assessed using proper scoring rules such as the Brier score and logarithmic score [11, 12]. A utility or loss model defines when to take action. [14].

Bayesian logistic regression uses the same Bernoulli likelihood as ordinary logistic regression but represents uncertainty through a posterior distribution over coefficients and predictions [8, 10]. This doesn’t inherently improve prediction or decisions. Its potential advantage is that prior information or shrinkage can stabilize estimation and that posterior draws can be propagated into downstream decision quantities. The paper aims to see whether Bayesian methods change calibrated predictive potential.

## 3 Data and Preprocessing

### 3.1 Synthetic case-study dataset

The empirical dataset contains 2105 synthetic records with 1304 positive and 801 negative diagnosis labels, for an observed positive-label prevalence of 61.9%. Variables include demographics, medical history, physiological measurements, cognitive and functional assessments, cardinal motor symptoms, and a binary diagnosis indicator [15].

The primary model uses 14 predictors: age, gender, BMI, family history of PD, hypertension, diabetes, UPDRS, MoCA, functional assessment, tremor, rigidity, bradykinesia, systolic blood pressure, and total cholesterol. Patient identifiers and non-informative administrative fields are excluded. A reduced specification dropping UPDRS, tremor, rigidity, and bradykinesia is fitted alongside it and reported in Section 6.1.

### 3.2 Leakage-controlled preprocessing

Continuous predictors are standardized within each training fold,

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_{j,\text{train}}}{s_{j,\text{train}}},$$

with the training-fold means and standard deviations then applied unchanged to the corresponding assessment fold. Binary indicators remain coded as 0/1. Missing-data handling, factor coding, fold assignments, seeds, and software versions are preserved within the github repository.

This fold-specific standardization serves two purposes. First, it prevents information from the assessment set from entering preprocessing. Second, it places continuous predictors on comparable scales so that a common weakly informative prior on slopes has a coherent interpretation.

## 4 Statistical Framework

### 4.1 Generalized linear models

For observation  $i$ , let  $Y_i \in \{0, 1\}$  and let  $x_i$  denote its predictor vector. A logistic model assumes

$$Y_i \mid \beta \sim \text{Bernoulli}(\pi_i), \quad (1)$$

$$\text{logit}(\pi_i) = x_i^T \beta, \quad (2)$$

so that

$$\pi_i = \frac{1}{1 + \exp(-x_i^T \beta)}.$$

For descriptive model-form comparison, probit and complementary log-log links are also fitted. Their AIC and BIC values are reported as in-sample criteria rather than evidence of out-of-sample superiority. The classical link-model comparison yields AIC/BIC values of 1760.0/1844.7 for logit, 1765.8/1850.6 for probit, and 1804.4/1889.2 for complementary log-log.

### 4.2 Frequentist regularization and Bayesian logistic regression

The out-of-sample comparison includes maximum-likelihood logistic regression, ridge logistic regression, lasso logistic regression, and Bayesian logistic regression. Ridge and lasso tuning parameters are selected by inner cross-validation using only the outer training fold.

For the Bayesian model, standardized slopes receive weakly informative priors

$$\beta_j \sim N(0, 1.5^2),$$

with a separate intercept prior

$$\beta_0 \sim N(0, 2.5^2).$$

The posterior is

$$p(\beta \mid Y, X) \propto \left[ \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \right] \left[ \prod_{j=1}^p \phi(\beta_j; 0, 1.5^2) \right] \phi(\beta_0; 0, 2.5^2).$$

Sampling is performed with the No-U-Turn Sampler [9, 10]. For a new predictor vector  $x$ , posterior predictive risk is estimated as

$$\bar{\pi}(x) = E_{\beta \mid D}[\pi(x, \beta)] \approx \frac{1}{S} \sum_{s=1}^S \frac{1}{1 + \exp(-x^T \beta^{(s)})}.$$

The full-data Bayesian fit had maximum  $\hat{R}$  of 1.001, minimum bulk effective sample size 5847, minimum tail effective sample size 5578, and 0 divergent transitions. Its PSIS-LOO expected log predictive density was -880.1 (SE 25.6) [13].

### 4.3 Bayes action under asymmetric loss

Let  $\delta \in \{0, 1\}$  denote the action, where  $\delta = 1$  means referral and  $\delta = 0$  means no referral. Consider the loss

$$L(y, \delta) = \begin{cases} 0, & y = \delta, \\ c_{FN}, & y = 1, \delta = 0, \\ c_{FP}, & y = 0, \delta = 1. \end{cases}$$

Given calibrated posterior risk  $\pi$ , the conditional risk of no referral is  $\pi c_{FN}$ , while the conditional risk of referral is  $(1 - \pi)c_{FP}$ . Referral is preferred when

$$(1 - \pi)c_{FP} < \pi c_{FN},$$

which gives

$$\tau^* = \frac{c_{FP}}{c_{FN} + c_{FP}}. \quad (3)$$

If  $c_{FN} = \lambda c_{FP}$ ,

$$\tau^*(\lambda) = \frac{1}{1 + \lambda}. \quad (4)$$

Thus  $\lambda = 3$  implies  $\tau^* = 0.25$ .

Equation 4 is conditional on the loss ratio and on probabilities being calibrated for the target population. This does not describe the optimal cutoff learned via the dataset. A different loss ratio, target population, calibration function, or capacity constraint can change the appropriate action rule.

### 4.4 Population expected loss and prevalence

For a threshold  $t$ , sensitivity  $\text{Sens}(t)$ , specificity  $\text{Spec}(t)$ , positive prevalence  $q$ , and costs  $c_{FN}$  and  $c_{FP}$ , population-level normalized expected loss is

$$R(t; q, c_{FN}, c_{FP}) = q c_{FN}[1 - \text{Sens}(t)] + (1 - q)c_{FP}[1 - \text{Spec}(t)]. \quad (5)$$

This representation makes explicit that even when an individual risk threshold follows Equation 4, the aggregate burden of false positives and false negatives depends on the prevalence of the target population. Sensitivity and specificity are properties of the fitted model and the threshold alone, so  $q$  can be varied in Equation 5 without refitting anything.

For finite out-of-sample data, define empirical decision regret relative to the best threshold on a specified grid as

$$\text{Regret}(t) = \widehat{R}(t) - \min_{u \in \mathcal{T}} \widehat{R}(u).$$

The regret surface provides a direct check of how costly a theoretically derived threshold becomes when probabilities are imperfectly calibrated or the evaluation sample is finite.

## 4.5 Posterior decision uncertainty

The Bayes action uses posterior *expected* risk,

$$\delta^*(x) = \mathbf{1}\{\bar{\pi}(x) > \tau^*\}.$$

This differs from asking how stable that action is across plausible parameter values. For posterior draw  $s$ , define

$$\delta^{(s)}(x) = \mathbf{1}\{\pi^{(s)}(x) > \tau^*\}.$$

Then

$$P(\delta = 1 \mid D, x) \approx \frac{1}{S} \sum_{s=1}^S \delta^{(s)}(x)$$

measures posterior uncertainty about which side of the decision boundary an observation lies on. This quantity does not replace the Bayes action; it describes uncertainty surrounding it. It also does not include dataset shift or sampling uncertainty from observing a new population.

## 5 Validation and Simulation Design

### 5.1 Repeated nested validation

The empirical comparison uses stratified five-fold cross-validation repeated five times. All standardization occurs inside each outer training fold. Ridge and lasso penalties are tuned through an inner five-fold cross-validation using the outer training data only. The Bayesian model is refit within each outer fold, and posterior-mean predictions are generated for observations not used in fitting.

For each outer assessment fold, we compute ROC AUC, precision–recall AUC, Brier score,

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n (\hat{\pi}_i - y_i)^2,$$

log loss,

$$-\frac{1}{n} \sum_{i=1}^n [y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)],$$

and calibration intercept and slope. Threshold-specific sensitivity, specificity, positive predictive value, negative predictive value, and expected loss are evaluated from pooled out-of-fold predictions.

### 5.2 Simulation study

The simulation isolates statistical behavior from the inherent design of the synthetic Parkinson’s dataset. Predictors are generated from a correlated Gaussian design with

$$\Sigma_{jk} = \rho^{|j-k|},$$

and a sparse logistic signal. Sample size, outcome prevalence, predictor correlation, and model specification are varied. In correctly specified scenarios, data are generated from the same linear logistic family used for fitting. In misspecified scenarios, the true log-odds additionally contain an interaction and nonlinear quadratic component omitted from the fitted models.

The simulation compares maximum-likelihood logistic regression, ridge logistic regression, and a Gaussian-prior Bayesian logistic estimator computed by a Laplace approximation around the posterior mode. The approximation is used only for the simulation study and was validated against NUTS on 10 representative simulated datasets, with checks saved in the reproducibility directory.

Each fitted model is evaluated on a large independent test population where the true risk  $\pi_{\text{true}}(x)$  is known. For cost ratio  $\lambda$ , the oracle action is

$$\delta_{\text{oracle}}(x) = \mathbf{1} \left\{ \pi_{\text{true}}(x) > \frac{1}{1 + \lambda} \right\}.$$

Decision regret is then

$$\text{Regret}(\hat{\delta}) = R(\hat{\delta}) - R(\delta_{\text{oracle}}).$$

This simulation directly tests the claim that shrinkage or posterior averaging can improve downstream decisions. It also permits a null result: if Bayesian estimation does not reduce regret in a scenario, the analysis reports that result rather than assuming an advantage.

## 6 Results

### 6.1 Out-of-sample probabilistic performance

Table 1 summarizes pooled repeated-validation performance. Because the models share the same binary outcome and outer folds, differences in Brier score, log loss, calibration, and ranking metrics can be interpreted as differences in out-of-sample predictive behavior on this synthetic case study.

**Table 1:** Repeated out-of-sample performance. Values are generated by the R analysis pipeline from 25 outer assessment folds.

| Model                              | ROC AUC | PR AUC | Brier | Log loss | Cal. int. | Cal. slope |
|------------------------------------|---------|--------|-------|----------|-----------|------------|
| MLE logistic                       | 0.881   | 0.922  | 0.136 | 0.419    | 0.005     | 0.983      |
| Ridge logistic                     | 0.881   | 0.922  | 0.137 | 0.426    | -0.092    | 1.310      |
| Lasso logistic                     | 0.881   | 0.923  | 0.136 | 0.418    | -0.015    | 1.058      |
| Bayesian logistic                  | 0.881   | 0.922  | 0.136 | 0.419    | 0.002     | 0.999      |
| Bayesian, criteria-free predictors | 0.668   | 0.755  | 0.217 | 0.623    | 0.018     | 0.959      |

The four full-predictor models rank observations identically to three decimal places and are nearly indistinguishable on the proper scoring rules. Their calibration slopes are not identical: ridge sits at 1.310 against 0.999 for the Bayesian fit. Since Equation 3 operates on the numerical probability scale, that gap matters more for what follows than the shared AUC does. A model can preserve its ranking

while moving probabilities enough to change who crosses a cutoff.

The last row of Table 1 is a different kind of check. Dropping UPDRS and the three cardinal motor symptoms takes ROC AUC from 0.881 to 0.668 and Brier score from 0.136 to 0.217. Most of the full model’s apparent predictive strength therefore comes from four variables that stand in for the diagnostic definition itself, which is what Section 2.1 anticipated. Calibration survives the removal almost unchanged, at slope 0.959. So the criteria-free model ranks much worse and still produces probabilities that mean what they say, and the decision analysis below could be run on either one. Only one of them is measuring something beyond how the label was constructed.

## 6.2 Posterior structure and predictive diagnostics

The full-data Bayesian fit describes posterior structure within the synthetic generator. It does not estimate clinical effect sizes. Tremor, bradykinesia, and rigidity carry the largest positive posterior means, with UPDRS close behind; functional assessment and MoCA carry the largest negative ones. All six have 95% credible intervals well clear of zero. The demographic and cardiometabolic predictors — BMI, hypertension, systolic blood pressure, total cholesterol, family history — sit near zero with intervals that cross it. The three cardinal symptoms are close enough to one another that the posterior does not support ranking them, and no such ranking is claimed.

Because continuous variables are standardized, their coefficients are read per standard deviation while the binary indicators remain 0-to-1 contrasts, so magnitudes are not directly comparable across the two groups. The point of carrying the posterior forward is not the coefficient table itself. It is that the same uncertainty can be pushed into predicted risks and then into referral decisions, which is what Section 6.4 does.

## 6.3 Asymmetric loss changes the operating point

For the Bayesian out-of-fold predictions, Table 2 compares the conventional 0.50 threshold with  $\tau^* = 0.25$ , corresponding to  $\lambda = 3$ . The change in cutoff alters the action rule while leaving the fitted probability model unchanged.

**Table 2:** Out-of-fold threshold performance for the Bayesian model.

| Threshold | Sensitivity | Specificity | PPV   | NPV   | Expected loss ( $\lambda = 3$ ) |
|-----------|-------------|-------------|-------|-------|---------------------------------|
| 0.50      | 0.856       | 0.704       | 0.825 | 0.750 | 0.381                           |
| 0.25      | 0.959       | 0.442       | 0.737 | 0.869 | 0.288                           |

At the observed synthetic prevalence, the empirically loss-minimizing threshold for  $\lambda = 3$  was 0.21, while the regret of the theoretical  $\tau^* = 0.25$  relative to the best threshold on the finite grid was 0.005. Agreement would be expected when the risk estimates are well calibrated and sampling noise is modest. Disagreement is indicative of finite-sample error, calibration error, or both.

The prevalence sweep in Appendix C qualifies this. At 61.9% the theoretical rule is close to optimal,

but the loss-minimizing threshold moves as  $q$  falls, and the regret surface shows the theoretical rule paying most where positives are rare. That regime, not the one this dataset sits in, is where real screening happens. The agreement reported above should be read as a property of this prevalence rather than as a general endorsement of the rule.

## 6.4 Posterior uncertainty in the referral decision

At  $\lambda = 3$ , 1.5% of held-out observations had posterior referral probability between 0.40 and 0.60, while 6.4% had posterior-mean risk within 0.05 of the 0.25 threshold. These quantities identify cases for which parameter uncertainty can materially change the side of the decision boundary. The mass is heavily concentrated at 0 and 1, so most actions are stable and a small minority are not.

This distinction is conceptually important. A posterior interval for sensitivity conditional on one fitted dataset is not the same as sampling uncertainty under a new dataset, and thresholding each posterior draw is not itself the Bayes action. The Bayes action is based on posterior expected risk; draw-level actions are used here to quantify its stability.

## 6.5 Prior sensitivity

Bayesian conclusions can be sensitive to prior scale when the likelihood is weak, predictors are correlated, or data are sparse. None of those conditions holds strongly here, and the refits confirm it. Under a tighter  $N(0, 0.75^2)$  slope prior, a heavier-tailed Student- $t(3, 0, 2.5)$ , and the baseline  $N(0, 1.5^2)$ , the posterior means barely move. The tighter prior pulls UPDRS, MoCA, and functional assessment slightly toward zero, as it should; the heavier-tailed prior nudges the motor coefficients marginally outward. The ordering of predictors is identical across all three specifications and the credible intervals overlap throughout.

That is the result one wants, and it is not automatic. A decision analysis should not hinge on a choice among defensible weakly informative priors. Where it does, the right reading is not that one prior is correct but that the data do not determine the action.

## 6.6 Simulation: estimator quality and decision regret

The simulation evaluates model behavior against known truth. The main quantity is decision regret relative to the oracle action, rather than whether any method “wins” on a particular synthetic dataset. The R analysis reports results across sample size, prevalence, predictor correlation, and correct versus misspecified functional form.

In the small-sample correctly specified settings, the Bayesian estimator’s median decision regret relative to maximum likelihood differs by 10.2%. At larger sample sizes the same comparison gives -1.8%, and under misspecification 0.1%. Overall: Bayesian shrinkage and MLE had similar median decision regret in the correctly specified simulation settings

The honest conclusion is therefore conditional and fairly narrow. Shrinkage buys a little when  $n$  is small and the logistic form is right. It buys essentially nothing once the sample grows, and nothing at

all when the functional form is wrong, since posterior precision cannot recover a missing interaction or a missing quadratic term. This is worth stating plainly rather than burying: the simulation was constructed so that a null result would be visible if one existed, and one did.

## 7 Discussion

### 7.1 Why Bayesian Analysis?

The Bayesian and frequentist logistic models share the same basic likelihood. With a large and informative dataset, weakly informative priors should produce posterior means close to maximum-likelihood estimates, and Table 1 shows exactly that. The main Bayesian contribution is therefore not a guaranteed improvement in point prediction. It is the ability to combine explicit regularization with a posterior distribution that can be propagated into risk estimates and decision quantities.

That advantage should not be overstated. Bootstrap procedures can quantify frequentist uncertainty in calibration and decision metrics; ridge and lasso provide effective regularization; and a Bayesian model can be badly calibrated, misspecified, or poorly chosen. The relevant comparison is empirical: which procedure produces more reliable probabilities and lower decision regret under the conditions studied?

### 7.2 Calibration

The threshold derivation assumes that the input probability has decision-theoretic meaning. If observations assigned risk 0.25 actually experience the outcome at a materially different frequency in the target population, applying a 0.25 cost-derived cutoff does not implement the intended loss ratio. Calibration is therefore not an optional diagnostic after threshold selection; it is part of the foundation of the decision rule.

This is why the analysis emphasizes out-of-sample Brier score, log loss, calibration slope and intercept, and calibration curves. ROC AUC alone can remain unchanged under monotone transformations that substantially distort numerical probabilities and hence change threshold-based actions. The criteria-free fit makes the converse point: discrimination collapsed and calibration held, so the two are separable in both directions.

### 7.3 Interpretation of the synthetic Parkinson’s case study

The empirical results do not establish a clinical screening tool. The dataset is synthetic, and its strongest predictors may simply reveal how the synthetic label was generated. The drop from 0.881 to 0.668 puts a number on that suspicion. Motor symptoms and UPDRS-like variables are closely connected to the clinical phenotype used to recognize PD in real practice, and once they are removed the remaining demographic, cardiometabolic, and cognitive variables carry considerably less signal about the label. Their strong synthetic associations are unsurprising and should not be presented as medical discoveries.

The case study is useful for a different reason: it provides a concrete probability-estimation problem on which the consequences of asymmetric loss can be demonstrated. The simulation then supplies known ground truth, allowing calibration error and decision regret to be measured directly.

## 7.4 Limitations

Several limitations remain even after repeated validation. First, no external clinical cohort is used, so there is no evidence of transportability to real-world care settings. Second, the loss ratio  $\lambda$  is illustrative rather than elicited from affected individuals, clinicians, or a health system. Third, the simple loss matrix excludes downstream treatment effects, specialist capacity, evaluation burden, and time-to-diagnosis. Fourth, a linear logistic model may be misspecified even when sampling diagnostics are excellent. Fifth, the simulation necessarily represents a limited family of data-generating mechanisms.

The prevalence–cost analysis should likewise be read as a sensitivity analysis, not an estimate of real-world policy. Changing  $q$  in Equation 5 changes the population mix used to weight sensitivity and specificity; deploying probabilities in a population with different baseline prevalence may additionally require recalibration or model transport.

## 8 Conclusion

A probability estimate becomes a decision only after consequences are specified. Under asymmetric false-positive and false-negative losses, the appropriate risk threshold follows from expected loss, not from the convention of 0.50. That result is simple; using it responsibly is not. The numerical probability must be calibrated, the cost ratio must be defensible, the target prevalence and resource context must be understood, and uncertainty about the fitted model must be separated from uncertainty about future populations.

The synthetic Parkinson’s case study illustrates these distinctions without supporting clinical claims. Repeated validation evaluates whether the models produce reliable out-of-sample probabilities; the criteria-free comparison shows how much of the apparent performance rests on variables that encode the label; posterior draws quantify uncertainty around the decision boundary; prevalence–cost surfaces expose the assumptions hidden by a fixed cutoff; and simulation measures decision regret against an oracle rule with known probabilities. The resulting framework is best viewed as a general statistical argument for moving from classification accuracy toward calibrated probabilistic prediction and explicit decision analysis.

## Acknowledgments

The author thanks Dr. Olga Korosteleva for research guidance and continued discussion. He also thanks Dr. Rebecca Le, Dr. Sung Kim, Dr. Kathryn McCormick, Dr. Rolando de Santiago, Dr. Simon Rubenstein-Salzedo, Ms. Stephanie Machado, Mrs. Cindy Carlson, and Dr. Tuesday Stoffers for their instruction and support.

## References

- [1] L. V. Kalia and A. E. Lang, “Parkinson’s disease,” *The Lancet*, vol. 386, no. 9996, pp. 896–912, 2015.
- [2] M. J. Armstrong and M. S. Okun, “Diagnosis and treatment of Parkinson disease: A review,” *JAMA*, vol. 323, no. 6, pp. 548–560, 2020.
- [3] R. B. Postuma et al., “MDS clinical diagnostic criteria for Parkinson’s disease,” *Movement Disorders*, vol. 30, no. 12, pp. 1591–1601, 2015.
- [4] J. Jankovic, “Parkinson’s disease: Clinical features and diagnosis,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.
- [5] W. Poewe et al., “Parkinson disease,” *Nature Reviews Disease Primers*, vol. 3, Art. no. 17013, 2017.
- [6] C. G. Goetz et al., “Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale,” *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.
- [7] A. Agresti, *Categorical Data Analysis*, 3rd ed., Wiley, 2013.
- [8] A. Gelman et al., *Bayesian Data Analysis*, 3rd ed., Chapman and Hall/CRC, 2013.
- [9] M. D. Hoffman and A. Gelman, “The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo,” *Journal of Machine Learning Research*, vol. 15, pp. 1593–1623, 2014.
- [10] B. Carpenter et al., “Stan: A probabilistic programming language,” *Journal of Statistical Software*, vol. 76, no. 1, pp. 1–32, 2017.
- [11] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
- [12] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.
- [13] A. Vehtari, A. Gelman, and J. Gabry, “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC,” *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.
- [14] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer, 1985.
- [15] R. El Kharoua, *Parkinson’s Disease Dataset Analysis*, Kaggle dataset, 2024. Synthetic educational dataset, CC BY 4.0.

## A Decision Rule Derivation

For posterior risk  $\pi(x)$ , the conditional risks are

$$\rho(1 \mid x) = c_{FP}[1 - \pi(x)], \quad (6)$$

$$\rho(0 \mid x) = c_{FN}\pi(x). \quad (7)$$

Choosing referral when  $\rho(1 \mid x) < \rho(0 \mid x)$  gives

$$\pi(x) > \frac{c_{FP}}{c_{FN} + c_{FP}},$$

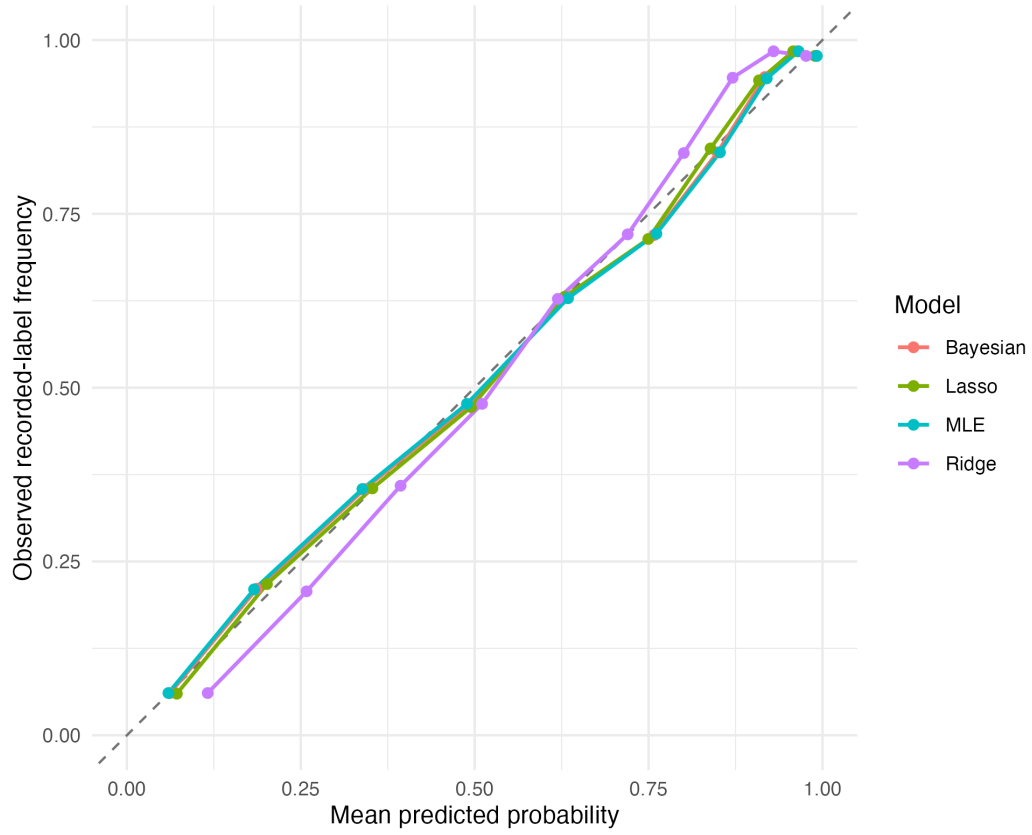
and with  $c_{FN} = \lambda c_{FP}$ ,

$$\tau^*(\lambda) = \frac{1}{1 + \lambda}.$$

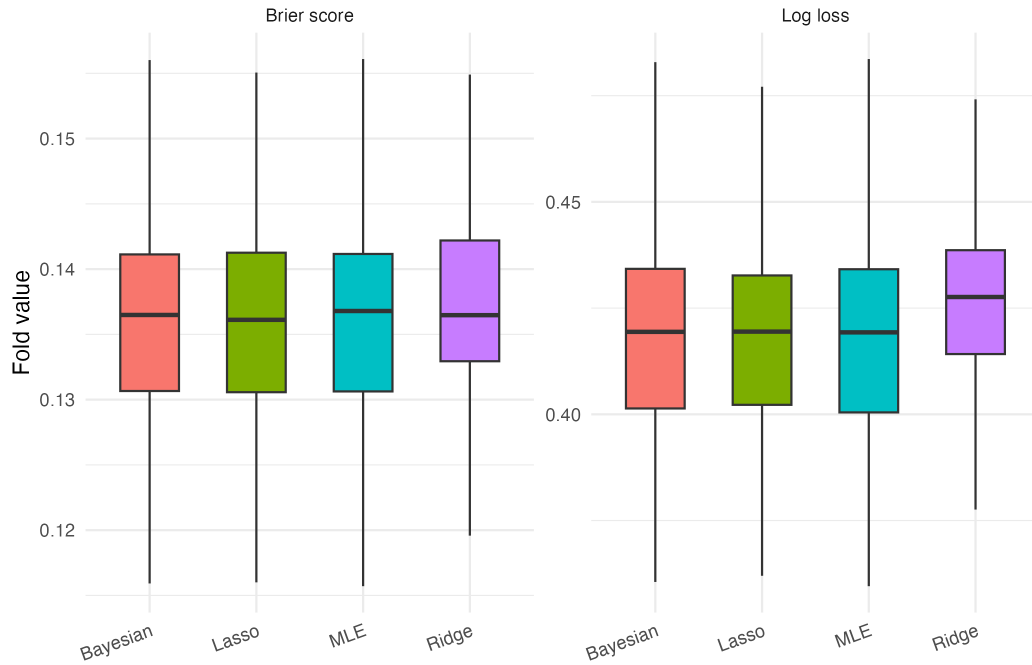
## B Reproducibility

All experiments ran from the single R entry point `analysis/run_analysis.R`, with master seed 20260823. The script writes machine-readable result tables, a preprocessing manifest, session information, 300-dpi PNG figures, and `analysis/results/results_macros.tex`, which supplies the numerical values in this manuscript. Please reference [github.com/zovial/parkinsonsdiseaseprediction](https://github.com/zovial/parkinsonsdiseaseprediction) for the most updated version of the pipeline and data.

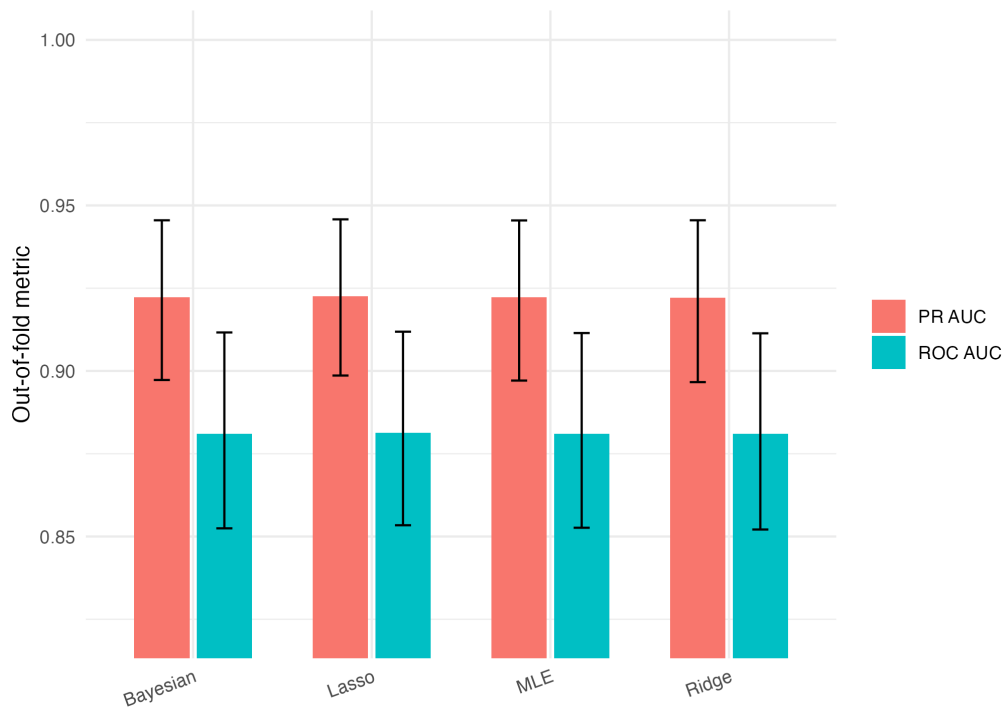
## C Figures



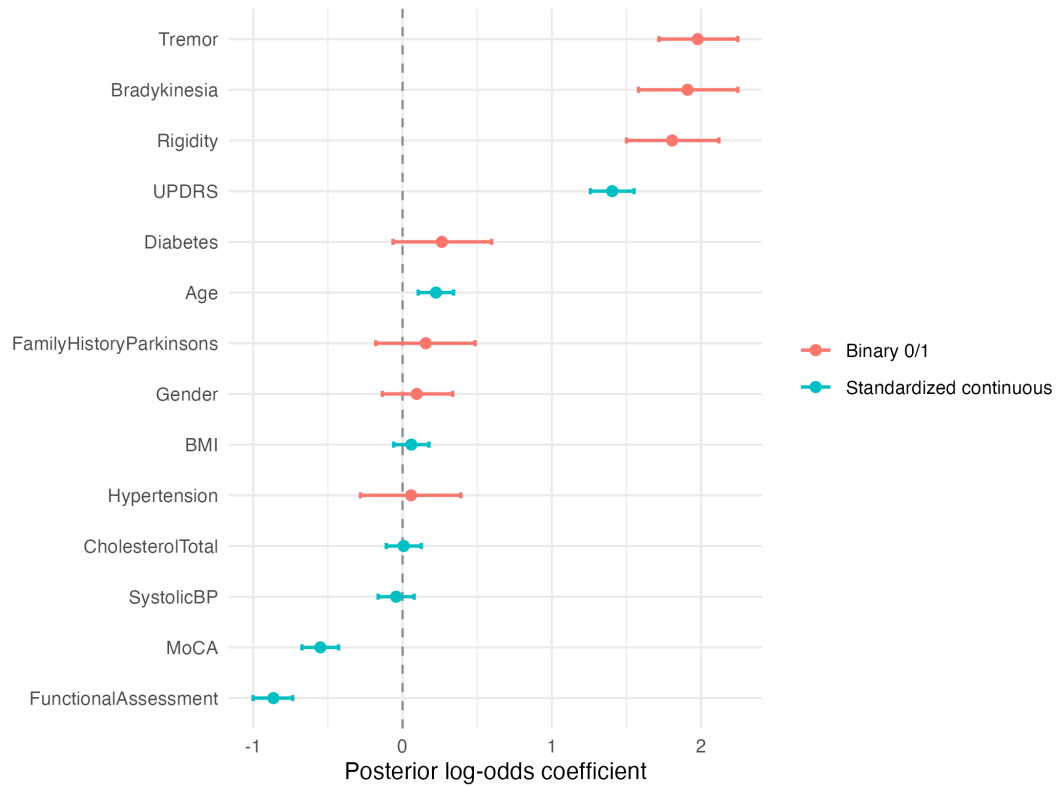
**Figure 1:** Pooled out-of-fold calibration curves for maximum-likelihood, ridge, lasso, and Bayesian logistic regression. The 45-degree line represents perfect calibration.



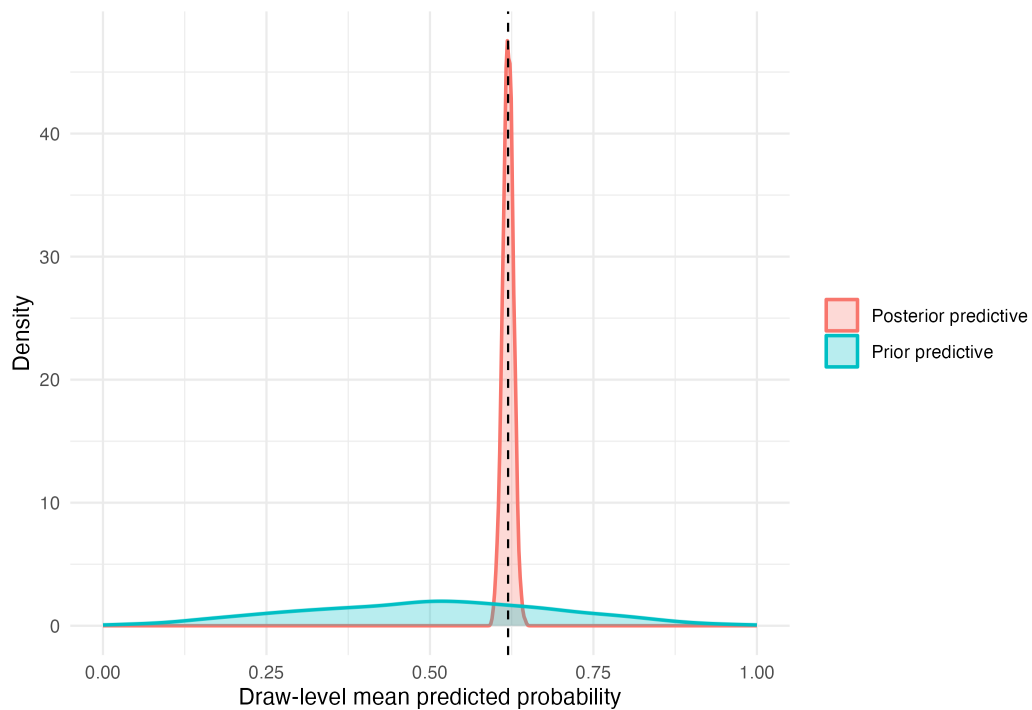
**Figure 2:** Distribution of Brier score and logarithmic loss across repeated outer assessment folds. Lower values indicate better probabilistic prediction.



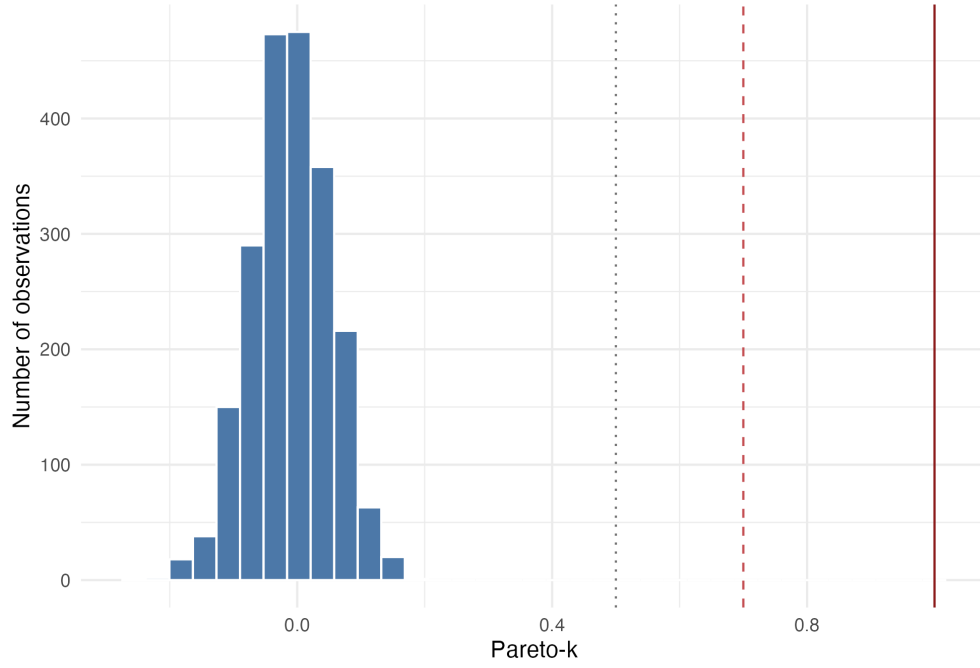
**Figure 3:** Out-of-fold discrimination comparison. ROC AUC and precision–recall AUC measure ranking rather than calibration and are therefore interpreted alongside proper scoring rules.



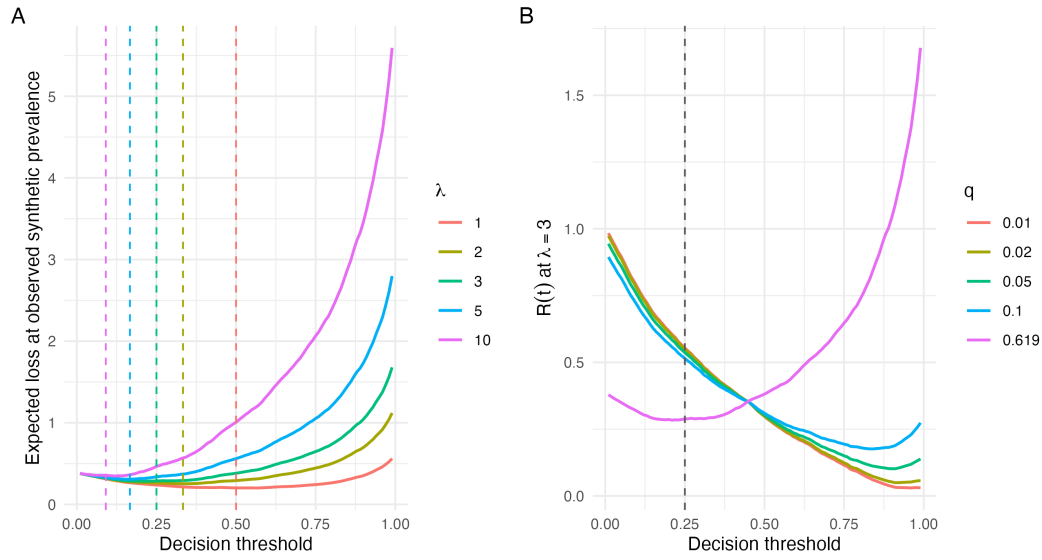
**Figure 4:** Posterior means and 95% credible intervals for the Bayesian logistic coefficients. These are associations with the synthetic diagnosis label and are not interpreted as clinical risk-factor effects.



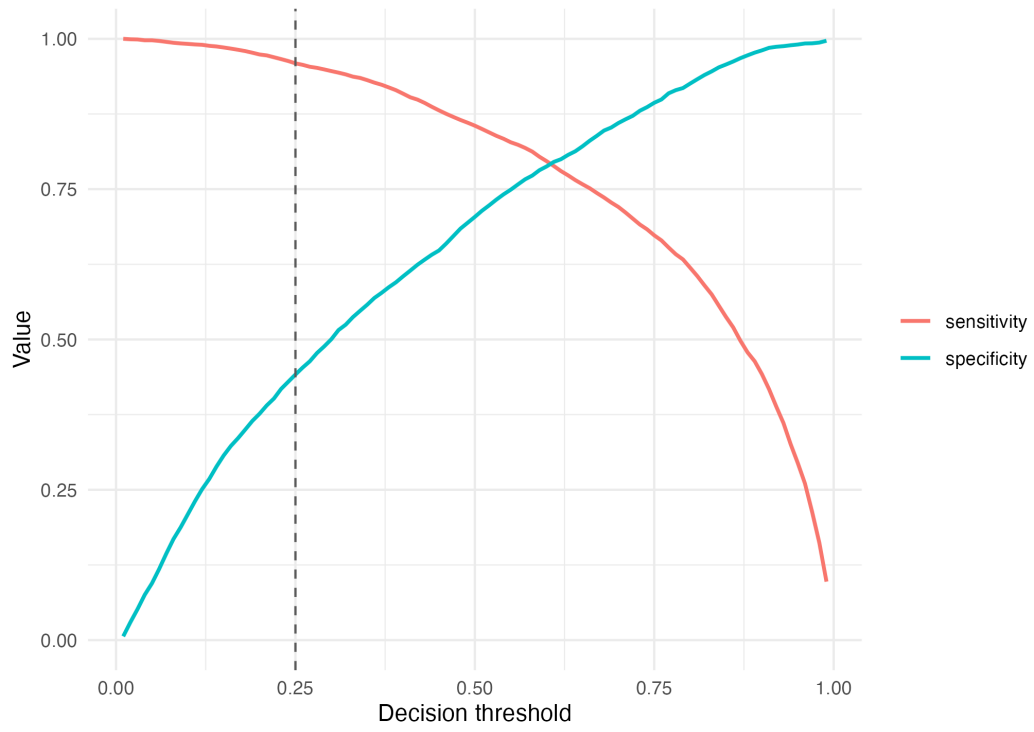
**Figure 5:** Prior and posterior predictive checks for the Bayesian logistic model. Prior predictive simulation evaluates whether the specified priors imply plausible probability distributions before observing the data.



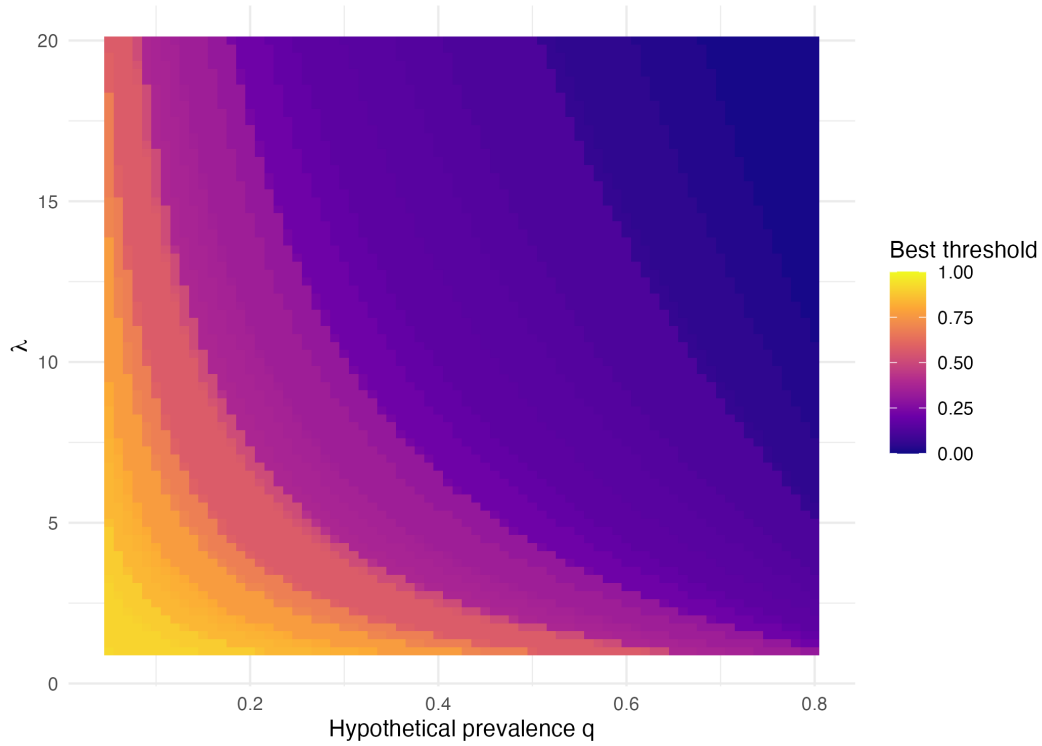
**Figure 6:** PSIS-LOO Pareto- $k$  diagnostics for the full-data Bayesian logistic model. Lower values indicate more stable leave-one-out importance sampling.



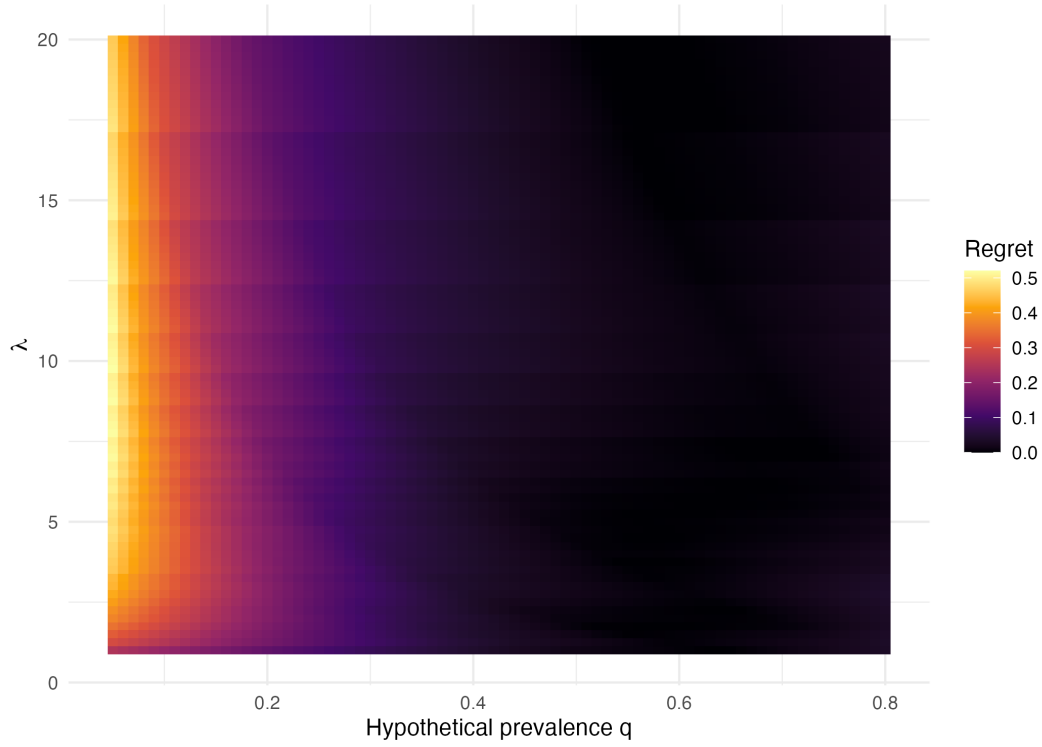
**Figure 7:** Out-of-fold expected loss across decision thresholds. Panel A shows several false-negative cost ratios at the observed synthetic prevalence. Panel B shows the prevalence sweep at  $\lambda = 3$ . Vertical markers show the theoretical threshold  $\tau^* = 1/(1 + \lambda)$ .



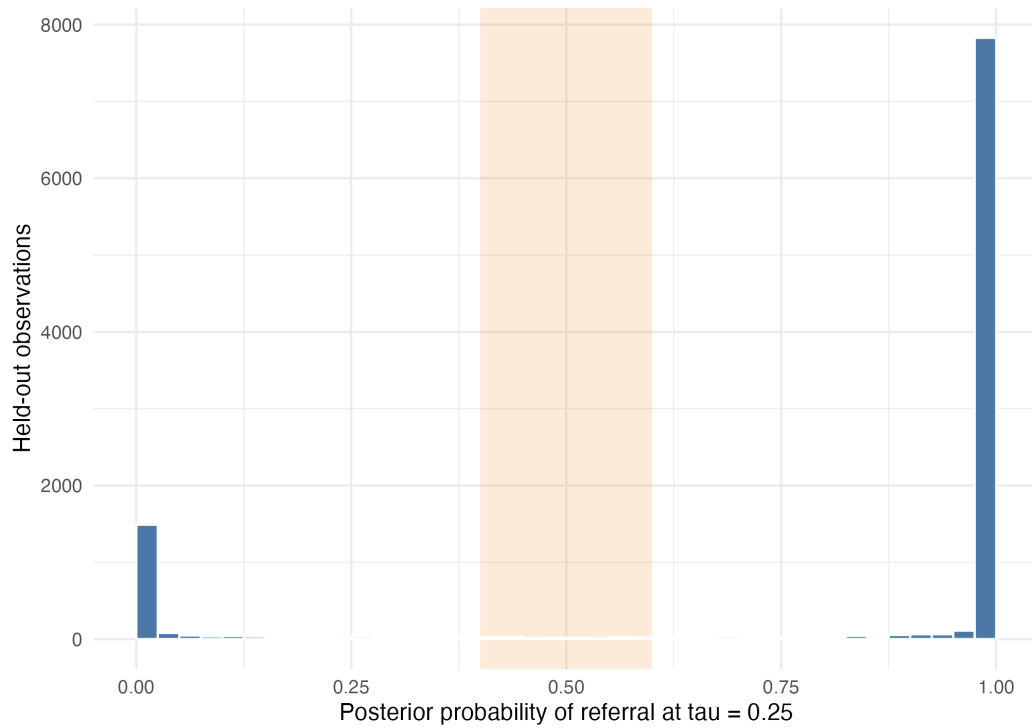
**Figure 8:** Sensitivity and specificity as continuous functions of the decision threshold for pooled Bayesian out-of-fold predictions.



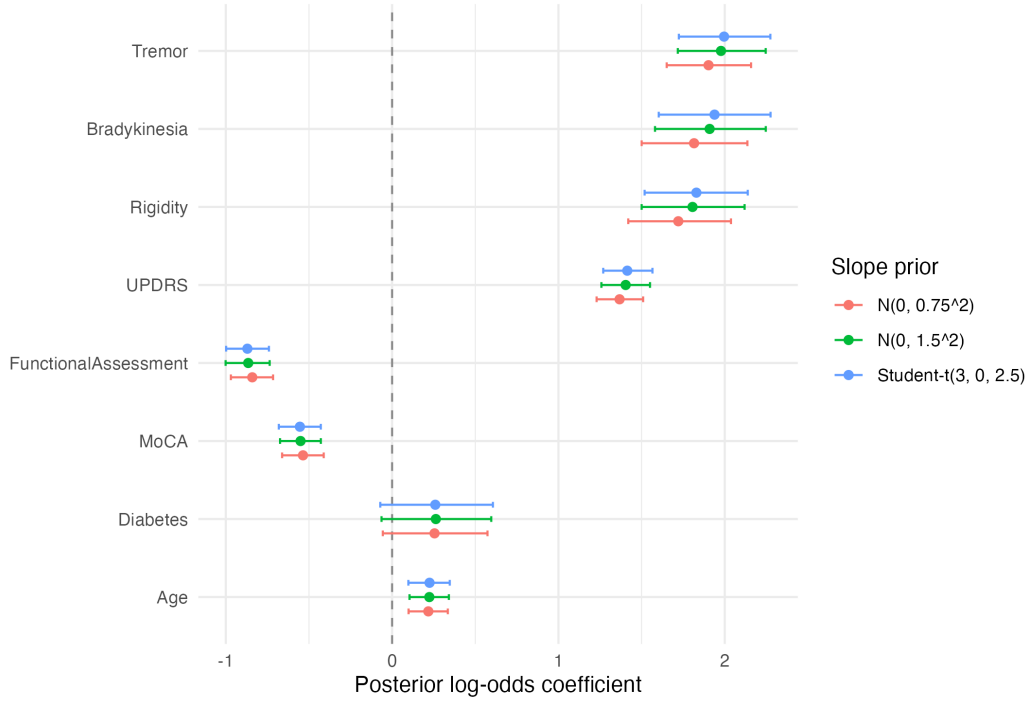
**Figure 9:** Empirically loss-minimizing threshold across hypothetical prevalence  $q$  and false-negative to false-positive cost ratio  $\lambda$ .



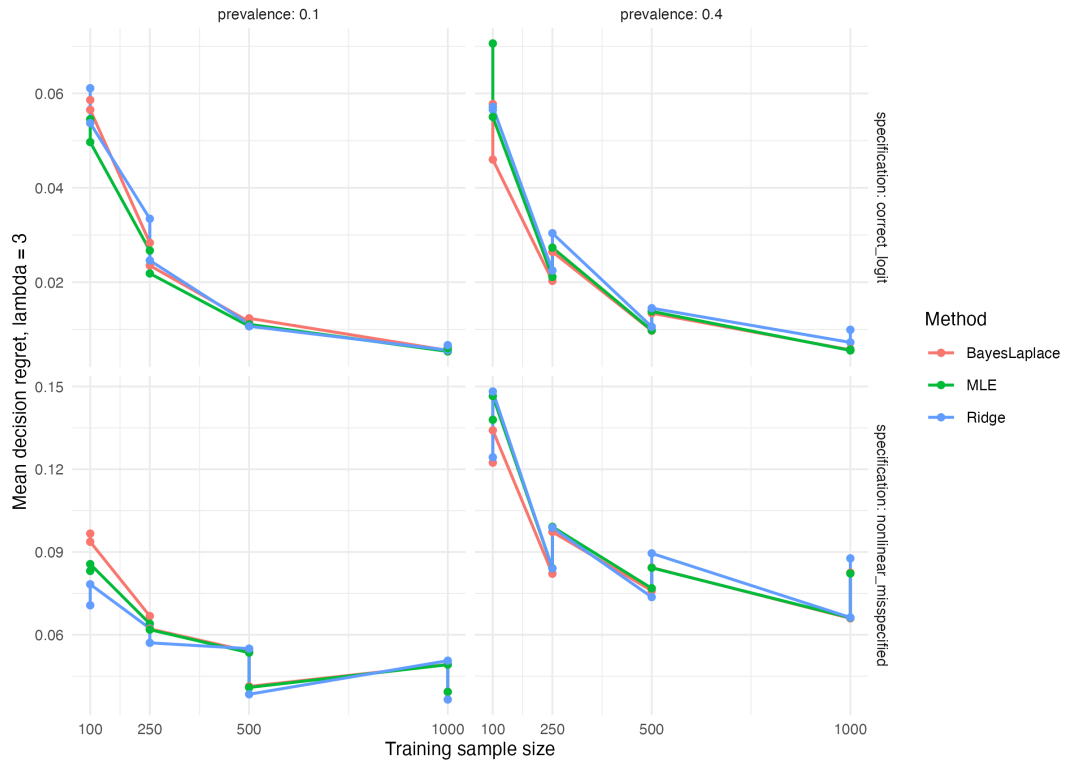
**Figure 10:** Regret of the theoretical threshold  $1/(1 + \lambda)$  relative to the empirically best threshold on the finite out-of-fold sample across prevalence–cost combinations.



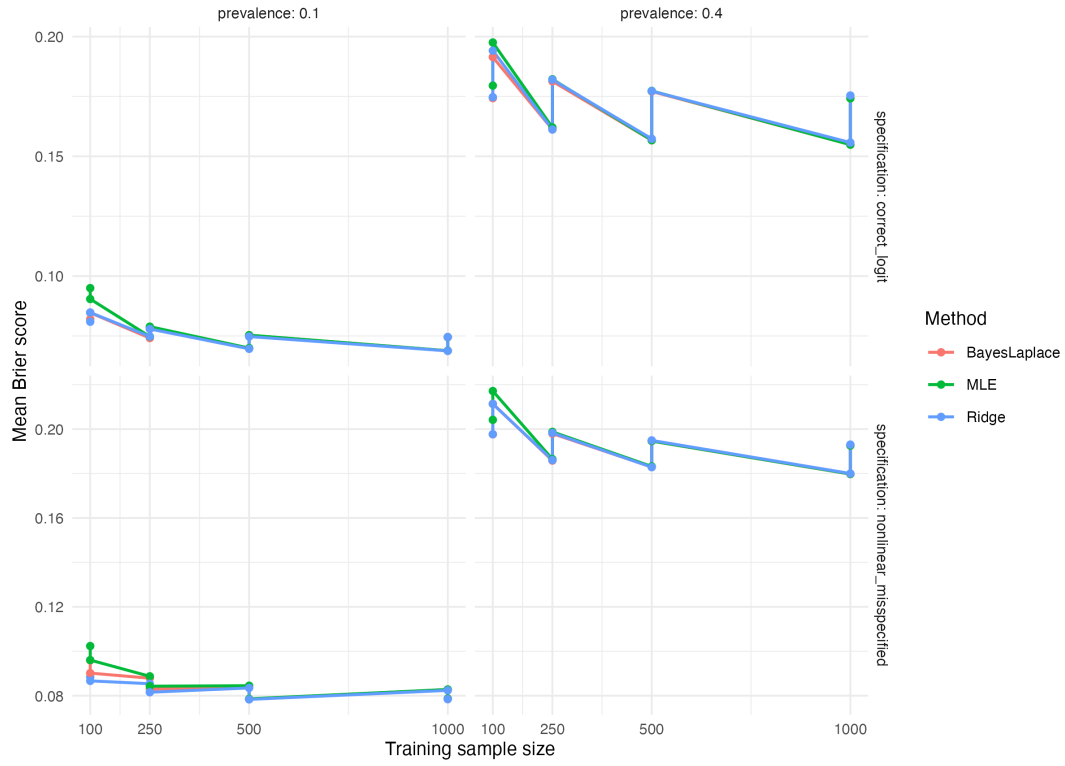
**Figure 11:** Posterior probability of referral at  $\lambda = 3$  for held-out observations. Values near 0.5 indicate substantial posterior uncertainty about which side of the threshold the observation lies on.



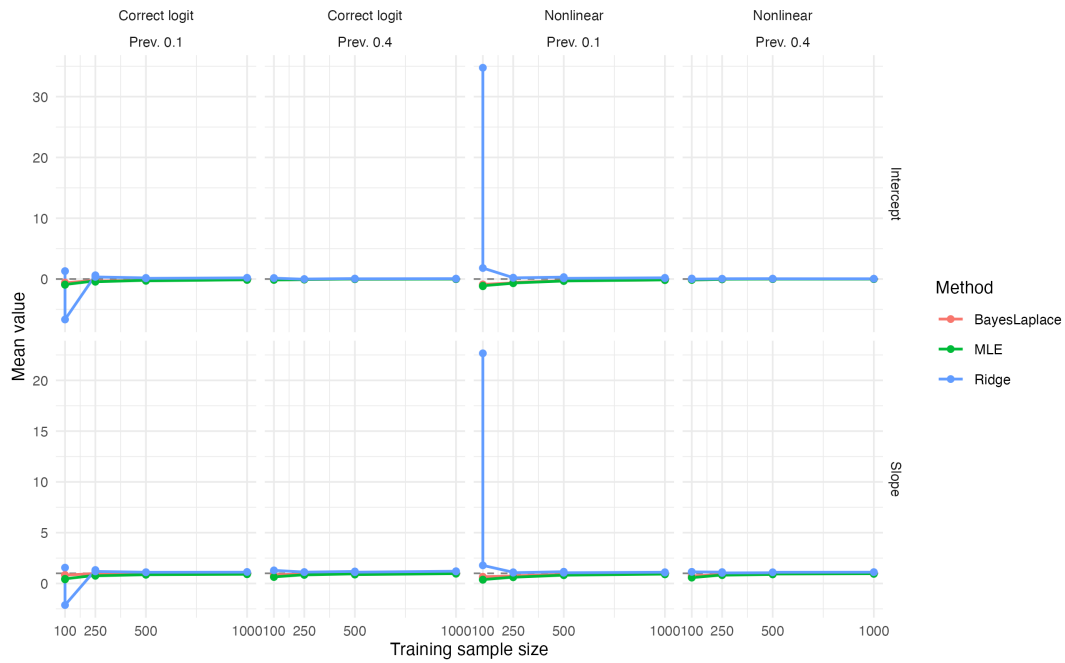
**Figure 12:** Sensitivity of posterior coefficients or predicted risks to alternative weakly informative priors.



**Figure 13:** Decision regret relative to the oracle rule across simulated sample sizes and model conditions. Lower regret indicates decisions closer to those made using the true data-generating probabilities.



**Figure 14:** Out-of-sample Brier score across simulated sample sizes and model conditions.



**Figure 15:** Calibration behavior across simulated sample sizes and model conditions.