Chapter 3
Variable Selection in Generalized Linear Models

Introduction

Confounders, Mediators, and Effect Modifiers

Understanding the impact of other variables in a statistical model is essential for producing valid and interpretable results. In epidemiological studies, where the aim is often to infer the relationships between exposures and health outcomes, other factors such as confounders, mediators, and effect modifiers play distinct roles and must be handled appropriately [1]. While these variables can all influence the observed relationship between exposure and outcome, they differ fundamentally in their roles. A clear understanding of these roles is essential not only for model specification and interpretation, but also for choosing appropriate strategies in study design and analysis.

A confounder is a variable that is associated both with the exposure and the outcome, and it can bias the estimation of the exposure–outcome relationship if not controlled. Formally, a confounder must satisfy three criteria:

- It is associated with the exposure
- It is an independent risk factor for the outcome
- It is not in the causal pathway between exposure and outcome

Consider a study assessing the association between the use of invasive ventilation and hospital-acquired infection (HAI). Suppose that severity of illness at admission is related both to the likelihood of receiving invasive ventilation (more severe patients are more likely to be intubated) and to the risk of developing an HAI (more severe patients have weaker immunity and longer hospital stays). In this case, severity of illness would be a confounder. Failing to adjust for it may lead to a spurious or exaggerated association between ventilation and infection risk [2]. To control for confounding, both design-level (randomization, matching, restriction)

© The Author(s) 2026

N. M. Khan et al., *Classical and Bayesian Statistical Approaches in Infectious Disease Data Analysis*, https://doi.org/10.1007/978-3-032-06747-0_3

and analysis-level (stratification, regression adjustment, Mantel–Haenszel methods) strategies can be used [3].

A mediator lies on the causal relationship between the exposure and outcome. Unlike a confounder, which distorts the observed relationship, a mediator explains how or why the exposure leads to the outcome. For example, suppose we are investigating whether recent antibiotic use increases the risk of *Clostridioides difficile* infection (CDI). If antibiotic use disrupts the gut microbiota, which then increases the susceptibility to CDI, the microbiota disruption serves as a mediator. Adjusting for the mediator estimates the direct effect of antibiotics on CDI, while omitting it estimates the total effect [4].

An effect modifier, or interaction term, occurs when the effect of the exposure on the outcome differs depending on the level of a third variable. In other words, the relationship between the exposure and outcome is not uniform across strata defined by the modifier [5]. As an example, consider the effect of catheter use on the development of urinary tract infections (UTIs). If the association is stronger among female patients than among males due to anatomical differences, then sex is an effect modifier [6]. Unlike confounding, which is a bias to be eliminated, effect modification is a real and meaningful phenomenon that should be described and interpreted. Statistical models can incorporate interaction terms to formally test for modification, or stratified analyses can be presented to show differential effects across subgroups.

Therefore, identifying confounders and effect modifiers requires both statistical tools and subject matter expertise. We suggest to apply a roadmap, as follows:

- Calculate the crude measure of association between exposure and outcome (e.g., odds ratio or risk ratio)
- Stratify the data by the potential third variable
- Calculate the stratum-specific estimates
- Use a test for homogeneity (e.g., Breslow–Day test or the test statistics in linear regression for the interaction coefficient) to assess whether the associations differ across strata
- If homogeneity is confirmed, compute an adjusted estimate to control for confounding
- If effect modification is detected, present and interpret stratified associations rather than an adjusted average

This roadmap helps ensure that confounding and effect modifiers are identified and described to enhance scientific insight and public health relevance.

Table 3.1 shows the results of three logistic models fitted on simulated data. The main predictor of interest is the invasive ventilation on the probability to infection. Here we try to understand the role of severity of disease as a confounder/mediator or as effect modifier. The crude odds Ratio (OR) from the univariable model for invasive ventilation is equal to 4.3 (95%CI: 3.26, 5.67). As the roadmap suggests, we stratify the analysis by Severity, introducing the interaction term between invasive ventilation and Severity. The OR for invasive ventilation in the group severity = no is 2.86 (95%CI: 1.87, 4.39) while the OR for invasive ventilation in

Table 3.1 Odds ratios (OR) and 95% confidence intervals to understand the role of disease severity in modifying the effect of invasive ventilation

Variable	Model 1	Model 2	Model 3
Invasive Ventilation	4.30 [3.26, 5.67]	2.86 [1.87, 4.39]	3.14 [2.33, 4.24]
Severity (Yes vs. No)	–	2.19 [1.41, 3.39]	2.41 [1.79, 3.23]
Invasive Ventilation × Severity	–	1.20 [0.66, 3.82]	–

the group severity = yes is obtained by multiplying the main effect of *invasive ventilation* ($\widehat{\beta}_1$) and the interaction term *invasive ventilation* × *Severity* ($\widehat{\beta}_3$). This gives an OR of:

$$\mathrm{OR}_{\mathrm{vent}\,|\,\mathrm{Severity}=\mathrm{yes}} = \exp\left(\widehat{\beta}_1 + \widehat{\beta}_3\right) = 2.86 \times 1.20 = 3.43$$

To compute the 95%CI for this marginal OR, we use the delta method to estimate the standard error (SE) of the linear combination $\left(\widehat{\beta}_1 + \widehat{\beta}_3\right)$, where:

$$\mathrm{SE}_{\mathrm{sum}} = \sqrt{\mathrm{Var}\left(\widehat{\beta}_1\right) + \mathrm{Var}\left(\widehat{\beta}_3\right) + 2 \cdot \mathrm{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_3\right)}$$

Then the 95% CI on the log-odds scale is:

$$\left(\widehat{\beta}_1 + \widehat{\beta}_3\right) \pm 1.96 \cdot \mathrm{SE}_{\mathrm{sum}}$$

Exponentiating the lower and upper bounds gives the CI for the marginal odds ratio.

In our example, plugging in the values from the R output, we have an $\mathrm{SE}_{\mathrm{sum}} \approx 0.375$ and consequentely a 95%CI equal to (1.65, 7.17).

To test the difference between severity groups, we can refer to the test on the interaction term. In our example, the interaction term is greater than 1 but not statistically significant. Therefore, we opt to include Severity just as confounder and select the additive model (Model 3) rather than the interaction model (Model 2).

Frequentist Variable Selection

The Likelihood Ratio Test

The decision to adopt Model 3 over Model 2 can also be supported by performing a Likelihood Ratio Test (LRT). The LRT is a widely used statistical procedure that evaluates whether the inclusion of one or more additional terms significantly improves the fit of a GLM. Specifically, it compares two nested models—one

Table 3.2 Comparison of GLMs using likelihood ratio test

Comparison	Df	Deviance	*P*-value
Model 1 vs. Model 2	2	33.554	<0.001***
Model 1 vs. Model 3	1	33.19	<0.001***
Model 3 vs. Model 2	1	0.364	0.546

more complex (the full model) and one simpler (the reduced model)—to determine whether the extra parameters in the full model significantly enhance the model's explanatory power [7].

In our example, Model 1 represents the simplest structure, a univariable logistic regression including only *invasive ventilation* as the predictor. Model 2 is the most complex specification, incorporating both *invasive ventilation*, *severity*, and their interaction term. Model 3 includes the same two main effects as Model 2 but excludes the interaction, thereby representing an additive model. Since Model 1 and Model 3 are both nested within Model 2, we can perform pairwise comparisons using the LRT to evaluate whether the added complexity in Model 2 significantly improves model fit. Specifically:

- **Model 1 vs. Model 2:** Tests whether including both severity and the interaction term improves the fit compared to a model with only *invasive ventilation*.
- **Model 3 vs. Model 2**: Tests whether the inclusion of the *invasive ventilation* × *severity* interaction term significantly improves the model beyond the additive effects.

By following the principle of model parsimony [8], we prefer simpler models unless the added parameters in a more complex model lead to a statistically significant improvement in fit. Therefore, LRTs offer a formal statistical approach to guide model selection in the context of nested logistic regression models.

The LRT statistic is then computed as follows:

$$\mathrm{D} = -2\left[\log \mathrm{L}\left(\widehat{\beta}_{\text{reduced}}\right) - \log \mathrm{L}\left(\widehat{\beta}_{\text{full}}\right)\right]$$

where $\mathrm{L}\left(\widehat{\beta}\right)$ denote the maximum likelihood estimates of the model.

Under the null hypothesis, which assumes that the additional predictor(s) do not significantly improve model fit of the simpler model, the test statistic (D) follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. A non-significant LRT result suggests that the more complex model does not offer a significantly better fit, and the simpler model should be preferred. In R, the LRT can be conducted using the *anova()* function from the *stats* package, specifying the argument *test* = *"LRT"*. This approach allows the comparison of nested models to assess whether the inclusion of additional predictors significantly improves model fit.

Table 3.2 summarizes the results of three pairwise comparisons. In the first and second comparisons, the LRT yields highly significant results: for both cases, the inclusion of severity as a main effect and the interaction between severity and

invasive ventilation significantly improve model fit compared to the univariable model (Model 1), with test statistics of ($\chi^2 = 33.55$), (df = 2), ($p < 0.001$) and ($\chi^2 = 33.19$), (df = 1), ($p < 0.001$), respectively.

However, in the third comparison—testing whether adding the interaction term (Model 2: interaction model) significantly improves fit over the additive model (Model 3: main effects only)—the LRT is not statistically significant (($\chi^2 = 0.36$), (df = 1), ($p = 0.546$)). This suggests that the interaction between invasive ventilation and disease severity does not contribute meaningful explanatory power beyond the additive effects already captured in Model 3. Given that the interaction term does not significantly enhance the model, Model 3 is the most appropriate and interpretable choice for estimating the effect of invasive ventilation on infection risk while accounting for severity.

Information Criteria (AIC/BIC)

While the LRT is a robust approach for comparing nested models, its scope is limited to assessing whether the inclusion of additional parameters yields a statistically significant improvement in model fit. In practical applications, particularly when faced with multiple competing models or non-nested alternatives, model selection must also consider the trade-off between model fit and complexity.

To address this trade-off, model selection criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are commonly used. These metrics extend the evaluation beyond significance testing by introducing penalties for the number of estimated parameters, thus reducing the risk of overfitting. The AIC is rooted in information theory and prioritizes predictive performance, while the BIC adopts a more conservative stance, favoring simpler models, particularly in larger samples.

The AIC is defined as follows [9]:

$$\mathrm{AIC} = -2 \cdot \log \mathrm{L}\left(\widehat{\beta}\right) + 2\mathrm{p}$$

where p denotes the total number of estimated parameters, including the intercept. A lower AIC indicates a better model in terms of the trade-off between goodness-of-fit and complexity. Unlike LRT, AIC can be used to compare models regardless of whether they are nested.

The BIC, also known as the Schwarz Information Criterion, is formulated as [10]:

$$\mathrm{BIC} = -2 \cdot \log \mathrm{L}\left(\widehat{\beta}\right) + p \cdot \log(n)$$

where n is the sample size. By incorporating the sample size into the penalty term, BIC imposes a stronger penalty for model complexity than AIC. As a result, BIC tends to select more parsimonious models, especially when n is large.

In R, AIC and BIC can be conveniently computed using the built-in *AIC()* and *BIC()* functions from the *stats* package. These functions allow for direct comparison of candidate models based on their balance between explanatory power and parsimony, thus serving as valuable complements to LRT in the model selection process.

Stepwise Regression

Stepwise regression [11] is a widely used automated procedure for model selection that iteratively adds or removes explanatory variables from a candidate set based on predefined statistical criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). This approach is particularly valuable in epidemiological and clinical settings where a large number of potential covariates may be available, and where manual model building would be impractical or subjective. There are three primary types of stepwise procedures: forward selection, backward elimination, and bidirectional selection. Each method uses information-theoretic criteria to evaluate model performance at each step and aims to strike an optimal balance between model complexity and explanatory or predictive power.

Forward Selection The forward selection algorithm begins with the simplest model, typically an intercept-only model, and incrementally adds predictor variables one at a time. At each step, the variable whose inclusion yields the greatest improvement in model fit (typically measured by the largest reduction in AIC or BIC) is added, provided that the improvement is statistically or practically significant. The process continues until no further variables contribute meaningfully to model performance, as determined by the stopping criterion.

Backward Elimination In contrast, backward elimination starts with the most complex model, containing all candidate predictors. Variables are then sequentially removed, beginning with the least contributive one, based on criteria such as the increase in AIC or BIC after removal. The process continues until eliminating additional predictors results in a significant deterioration of model quality. This technique is particularly sensitive to multicollinearity among variables and should be applied with caution in high-dimensional settings.

Bidirectional Selection Bidirectional (or stepwise) selection combines elements of both forward and backward methods. Variables can be added or removed at each iteration, allowing the procedure to adapt flexibly as the model evolves. This approach is advantageous when relationships among predictors are complex or when model parsimony must be balanced with performance.

In R, stepwise model selection can be implemented using the *stepAIC()* function from the *MASS* package. This function automates the iterative inclusion and

Table 3.3 Final model predictors selected via stepwise regression

Variable	Description
sex_at_birthMale	Male vs. female sex at birth
raceWhite	White vs. other racial groups
ethnicityOther	Other ethnicity (vs. reference)
respiratory_dis	Presence of respiratory disease
endocrine_dis	Presence of endocrine disease
cancer_dis	Presence of cancer
cathet_drainage_dialysis	Catheterization, drainage, or dialysis procedure
intubation	Underwent intubation
transplant	History of organ transplant

exclusion of variables based on AIC minimization. If BIC-based selection is preferred, it can be achieved by specifying the appropriate penalty term (i.e., $k = \log(n)$, where n is the sample size) within the same function.

Ultimately, while stepwise regression provides a convenient and objective method for variable selection, it should be interpreted cautiously [12]. The final model reflects a data-driven selection process and may be sensitive to sample size, collinearity, and initial model specification. Therefore, it is advisable to combine stepwise procedures with domain knowledge and cross-validation to ensure that the final model is both statistically robust and clinically meaningful.

Returning to our real-world case study based on the All of Us Research Program cohort, our objective was to identify a parsimonious, yet informative set of predictors associated with a binary infection outcome using a bidirectional stepwise regression approach. This procedure began with a full logistic regression model that included a broad range of candidate predictors, encompassing demographic variables, comorbidities, and recent medical procedures. Using AIC as the selection criterion, the algorithm iteratively assessed the impact of adding or removing variables, seeking to optimize model fit while avoiding unnecessary complexity. The stepwise procedure terminated when no further AIC improvement could be achieved.

The resulting model, shown in Table 3.3, achieved the lowest AIC value (AIC = 584.93) and includes nine predictors. These reflect a refined subset that balances statistical robustness and clinical interpretability. Table 3.4 presents the adjusted ORs and their 95%CI from the final logistic regression model.

Among the most prominent risk factors, intubation emerges as the strongest predictor of infection, with an OR of 4.98 (95% CI: 2.65–9.37), followed by transplant (OR = 4.29, 95% CI: 1.74–10.57) and endocrine disease (OR = 2.67, 95% CI: 1.66–4.30). These associations suggest that intensive interventions and pre-existing immune-modulating conditions play a critical role in infection susceptibility. Respiratory disease is also significantly associated with infection (OR = 3.29, 95% CI: 1.40–7.76), reinforcing the known vulnerability of patients with pulmonary conditions to secondary infections.

Demographic variables such as sex at birth, race, and Ethnicity did not reach statistical significance, although the direction of effects aligns with previous

Table 3.4 Odds ratios (ORs) and 95% confidence intervals from final logistic regression model (stepwise selection)

Variable	OR	95% CI lower	95% CI upper
Sex at birth: Male	1.50	0.97	2.27
Race: White	0.68	0.43	1.06
Ethnicity: Other	0.36	0.08	1.34
Respiratory disorder	3.29	1.40	7.76
Endocrine disorder	2.67	1.66	4.30
Cancer	1.64	0.90	2.64
Catheter/drainage/dialysis	2.26	0.98	5.28
Intubation	4.98	2.65	9.37
Transplant	4.29	1.74	10.57

findings. Finally, catheterization/drainage/dialysis was retained in the final model with borderline significance (p-value $= 0.063$), suggesting potential roles that merit further investigation in larger samples.

Bayesian Variable Selection

Bayesian variable selection offers a principled alternative to frequentist approaches, moving beyond point estimates and single-model decisions. Instead of relying solely on information criteria such as AIC or BIC, the Bayesian paradigm incorporates model uncertainty and evaluates predictive performance more holistically. Three widely used strategies in this context are Leave-One-Out Cross-Validation (LOO-CV), Bayesian Model Averaging (BMA), and Projection Predictive Feature Selection.

Leave-One-Out Cross-Validation (LOO-CV)

Leave-One-Out Cross-Validation (LOO-CV) is a widely used method in Bayesian model selection. Its main goal is to assess how well a model generalizes to new, unseen data. The idea is simple: remove one observation from the dataset, fit the model on the remaining data, and use it to predict the omitted value. Repeat this process for every observation in the dataset [9].

The predictive performance is then summarized using the expected log predictive density (ELPD)}, calculated as [13]:

$$\mathrm{ELPD}_{\mathrm{LOO}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \log p(y_i | y_{-i})$$

where y_i is the i-th observation and y_{-i} is the dataset excluding y_i. A higher ELPD value indicates better predictive accuracy.

To facilitate model comparison, we often use the LOO Information Criterion (LOOIC), which is similar in form to the AIC [9]:

$$\text{LOOIC} = -2 \times \text{ELPD}_{\text{LOO}}$$

Lower LOOIC values indicate better models. While conceptually straightforward, performing exact LOO-CV for Bayesian models can be computationally expensive because it requires refitting the model n times. To overcome this, the Pareto-smoothed importance sampling method—known as PSIS-LOO—offers a fast and accurate approximation [14].

The reliability of PSIS-LOO is evaluated using the Pareto k diagnostic. These diagnostic checks whether the importance weights used for approximating leave-one-out predictions are stable. See Box 3.1 for a concise walkthrough of how PSIS-LOO and the Pareto k diagnostic make leave-one-out cross-validation computationally feasible for Bayesian GLMs.

Box 3.1: Pareto-smoothed Importance Sampling (PSIS) and the Pareto k Diagnostic

What is PSIS-LOO?

In Bayesian analysis, exact LOO-CV requires refitting the model n times—once for each observation—which is computationally expensive. Pareto-smoothed importance sampling (PSIS) is an efficient approximation method that avoids full model refits by using importance sampling.

Given a posterior distribution ($p(\theta|\, y)$) based on full data ($y = (y_1, \ldots, y_n)$), we estimate the leave-one-out predictive density for observation (y_i) using:

$$p(y_i|y_{-i}) \approx \frac{\int p(y_i|\theta)p(\theta|y)\, w_i(\theta)\, d\theta}{\int w_i(\theta)\, d\theta} \quad \text{with} \quad w_i(\theta) = \frac{1}{p(y_i|\theta)}$$

Why is Pareto Smoothing?

The raw importance weights (w_i) can be highly variable and unstable, particularly when the omitted observation (y_i) has high influence. To improve stability, the upper tail of the weight distribution is smoothed by fitting a Generalized Pareto Distribution (GPD). This leads to more robust and accurate approximations.

(continued)

Box 3.1 (continued)

What is the Pareto k Diagnostic?

Pareto k is the shape parameter of the fitted GPD to the tail of the importance weights. It quantifies how influential an observation is on the posterior.

Interpretation of k values:

- ($k < 0.5$): Reliable approximation (low variance in weights).
- ($0.5 \leq k < 0.7$): Still acceptable, but some caution advised.
- ($0.7 \leq k < 1.0$): High variability; PSIS-LOO may be unstable.
- ($k \geq 1.0$): Unreliable approximation—consider exact LOO-CV or K-fold CV.

The k value is estimated from the tail weights by fitting a GPD:

$$\mathrm{GPD}(w) \sim 1 - \left(1 + \frac{kw}{\sigma}\right)^{-1/k}, \quad \text{for } k \neq 0$$

where w are the rescaled tail weights and (σ) is the scale parameter.

Table 3.5 Comparison of Bayesian model (BM) performance using LOO-CV estimates

Metric	BM 1: Age only	BM 2: Additive	BM 3: Interaction
ELPD_LOO	−322.82 (SE = 19.36)	−310.65 (SE = 19.18)	−310.93 (SE = 19.25)
p_LOO	2.03 (SE = 0.16)	3.17 (SE = 0.25)	4.32 (SE = 0.34)
LOOIC	645.65 (SE = 38.72)	621.29 (SE = 38.36)	621.86 (SE = 38.51)

To demonstrate the applicability of the LOO-CV method, Table 3.5 presents a comparison of three Bayesian models using LOO-CV performance metrics. In R, this analysis can be performed using the *loo()* function from the *loo* package. The goal is to identify the most suitable model for estimating the risk of infection within 180 days post-intervention in the All of Us cohort.

Model comparison is based on the ELPD and the LOOIC, where a higher ELPD and a lower LOOIC indicate better predictive performance. All models were estimated using default priors, and their specifications are as follows:

- **Bayesian Model 1** includes only the variable *age_60*
- **Bayesian Model 2** adds endocrine disorder as a main effect (additive model)
- **Bayesian Model 3** includes the interaction between *age_60* and endocrine disorder

The additive model (Bayesian Model 2) achieves the highest ELPD and the lowest LOOIC, suggesting that adding endocrine disorder as a predictor substantially improves model performance compared to using age alone. Although Bayesian Model 3 (interaction model) yields similar performance metrics, it has a slightly higher LOOIC and includes more parameters (as reflected in the higher p_{LOO}),

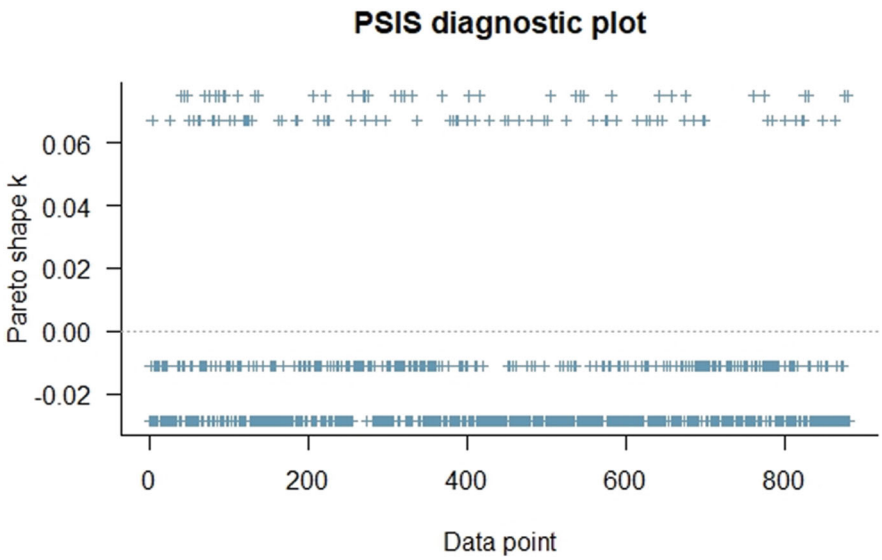

Fig. 3.1 Pareto K diagnostics – Bayesian Model 1

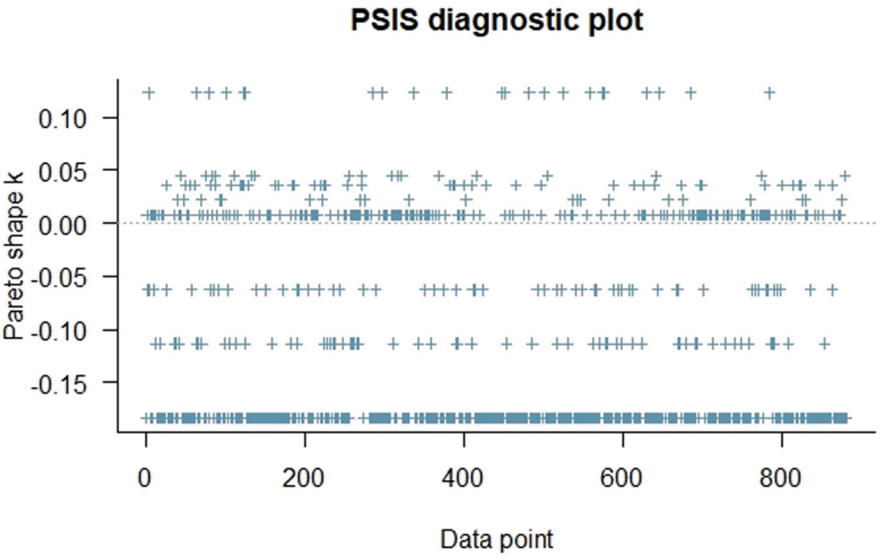

Fig. 3.2 Pareto K diagnostics – Bayesian Model 2

indicating greater complexity without significant predictive gains. Based on these results, Bayesian Model 2 provides the best balance of accuracy and parsimony.

Figures 3.1, 3.2, and 3.3 display the Pareto-smoothed importance sampling (PSIS) diagnostic plots for the three Bayesian models under comparison. Each "+"

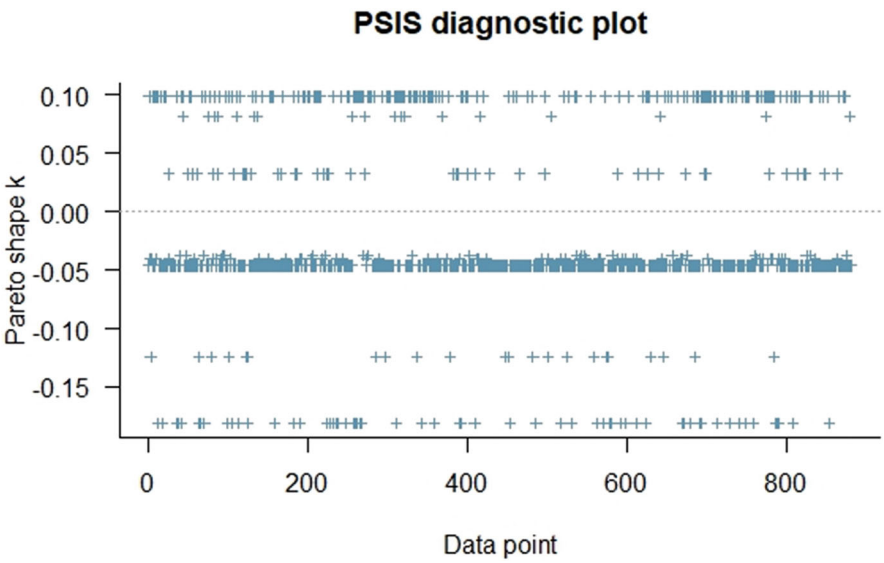

Fig. 3.3 Pareto K diagnostics – Bayesian Model 3

symbol represents an observation, with its corresponding estimated Pareto shape parameter k. In all three models, the Pareto k values remain well below the 0.7 threshold across all data points, suggesting that the PSIS-LOO estimates are stable and trustworthy. No influential observations are detected that would compromise the reliability of the LOO-CV estimates. This confirms the validity of using LOO-CV for model comparison in our case study and supports the interpretation of the results in Table 3.5.

Model Averaging Using Bayesian Adaptive Sampling

Bayesian Model Averaging (BMA) offers an alternative to assess model by model, by explicitly acknowledging model uncertainty. For instance, BMA combines results across many plausible models. Each model contributes to the final inference in proportion to its posterior model probability, resulting in more robust and stable estimates that help mitigate risks like overfitting and underfitting [15].

Formally, BMA estimates the posterior distribution of a parameter of interest, θ, by averaging over a set of candidate models $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$:

$$p(\theta|D) = \sum_{m=1}^{k} p(\theta|M_m, D) \cdot p(M_m|D)$$

where:

- $p(\theta|M_m, D)$ is the posterior distribution of θ under model M_m.

$p(M_m|D)$ is the posterior probability of model M_m given the data D

The posterior model probability is computed using Bayes' theorem:

$$p(M_m|D) = \frac{p(D|M_m) \cdot p(M_m)}{\sum_{j=1}^{k} p(D|M_j) \cdot p(M_j)}$$

where $p(D|M_m)$ is the marginal likelihood (also called Bayesian evidence), and $p(M_m)$ is the prior probability of model M_m. The marginal likelihood is calculated by integrating out the model parameters:

$$p(D|M_m) = \int p(D|\theta, M_m) \cdot p(\theta|M_m)\, d$$

In practice, computing this integral can be computationally intensive. Tools like Bayesian Adaptive Sampling (BAS) allow efficient exploration of this large model space without exhaustively evaluating every possible model [15]. In fact, BAS approximates these quantities by iteratively sampling models with high posterior probabilities, focusing computational resources on the most promising regions of the model space. It adaptively updates the exploration strategy based on previous sampling results, leveraging stochastic search techniques such as Metropolis–Hastings or MCMC. In order to provide the Bayesian counterpart of the Full model variable selection based on the All of Us cohort, we have applied BAS through the function *bas.glm()* from the *BAS* R package.

Figures 3.4 and 3.5 provide diagnostic insights from the Bayesian Adaptive Sampling (BAS) procedure used in our BMA analysis.

Figure 3.4 displays the Cumulative Model Probability as a function of the model search order. The steep initial rise indicates that the majority of the posterior model probability mass is concentrated in a small number of top-ranking models. This rapid convergence demonstrates that BAS efficiently identifies the highest-probability models early in the search process. The flat tail suggests diminishing returns from continued exploration, confirming that the posterior distribution is dominated by a few highly informative models.

Figure 3.5 shows the relationship between Model Complexity—measured by the number of included variables—and the log marginal likelihood. The plot reveals that moderate-sized models tend to achieve the highest marginal likelihoods, balancing fit and parsimony. As model dimension increases, a noticeable decline in log marginal likelihood appears, indicating that overly complex models are penalized and do not provide better support from the data. This behavior aligns with Bayesian principles that inherently guard against overfitting by integrating model complexity into the marginal likelihood calculation.

Finally, Table 3.6 reports the Posterior Inclusion Probabilities (PIP) from the BMA analysis [16]. The PIP represents the probability that each variable is included in the true model, given the data and prior information. Values close to 1 indicate strong evidence that a variable is relevant for predicting the outcome, while values

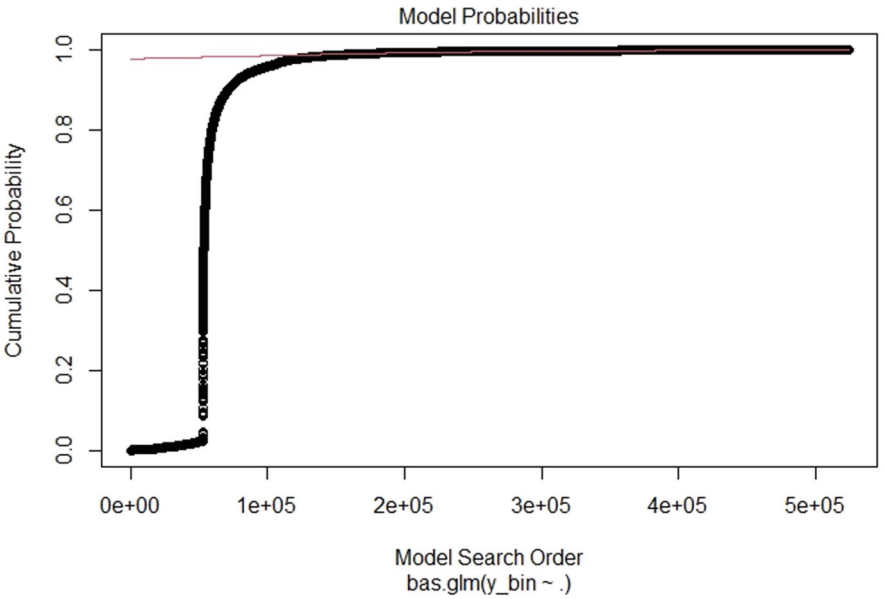

Fig. 3.4 Cumulative model probabilities as a function of the model search order in BAS

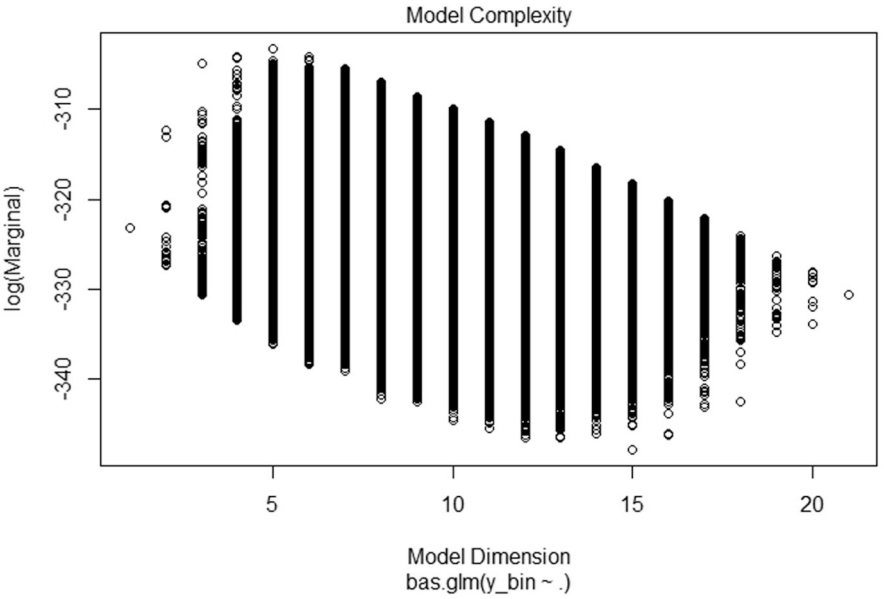

Fig. 3.5 Model complexity in BAS

Table 3.6 Posterior inclusion probabilities (PIP) from BMA

Variable	Posterior inclusion probabilities (PIP)
Intercept	1.0000
Sex at birth: Male	0.2698
Circulatory disorder	0.2585
Respiratory disorder	0.7043
Endocrine disorder	0.9963
Catheter/drainage/dialysis	0.2244
Endoscopy	0.3018
Intubation	0.9965
Transplant	0.6957

close to 0 suggest weak or negligible evidence. As shown, intubation and endocrine disorder have the highest inclusion probabilities (PIP = 0.997 and 0.996, respectively), providing strong evidence for their importance in modeling infection risk. Respiratory disorder and transplant also show relatively high PIP values (0.70), supporting their relevance. These findings align with clinical expectations, where intensive procedures and chronic conditions increase vulnerability to infections. Conversely, variables such as catheter/drainage/dialysis, endoscopy, circulatory disorder, and sex at birth exhibit relatively low PIP, indicating weaker support for their inclusion in the model based on the available data. Moreover, variables not displayed in the table reported PIP values below 20%, reinforcing the notion that they offer limited explanatory value and are unlikely to play a significant role in the prediction of infection risk.

Projection Predictive Feature Selection

The projection predictive variable selection method is a principled Bayesian approach to feature selection that aims to retain predictive performance while simplifying models. Its development is rooted in foundational work by Goutis and Robert (1998) [17], who introduced the idea of projecting the full posterior distribution onto a sub model space, and was further extended by Dupuis and Robert (2003) [18], who explored projection in the context of Bayesian model selection.

The Projective Predictive approach rather than selecting a sub model based on marginal posterior summaries, first fits a full (reference) model that aims to approximate the data-generating process as accurately as possible and then projects its information onto simpler candidate sub models. This strategy ensures that the sub models retain the predictive capabilities of the full model, while achieving greater simplicity.

Building upon these theoretical underpinnings, the method gained practical form through the work of Piironen and Vehtari, who systematically developed and formalized projection predictive techniques in the context of modern Bayesian

computation. Their seminal comparison study in Piironen and Vehtari (2017) [19] positioned the method among a variety of alternative selection approaches, demonstrating its favorable balance between parsimony and predictive accuracy.

Catalina et al. (2022) [20] extend the method beyond standard GLMs to include generalized linear mixed-effects models (GLMMs) and generalized additive mixed models (GAMMs), enabling structure and variable selection in models with random effects and nonlinear smooth terms.

More recently, Weber, Glass, and Vehtari (2023) [21] advanced the methodology with enhanced algorithms and diagnostics, solidifying its role as a competitive and interpretable alternative to classical selection methods like stepwise regression or penalization-based techniques.

The projection predictive workflow consists of three main components:

Fitting the Reference Model A full Bayesian model (the reference model) is fitted using all available covariates and model terms, including group-level effects and nonlinear smooths when relevant. The posterior distribution of the model parameters λ^* is obtained:

$$p(\lambda^*|\mathcal{D})$$

where λ^* denotes the full model parameters and $\mathcal{D}$ the observed data.

Projection Onto Sub models For a given sub model $\mathcal{M}$ (with a restricted set of terms), the goal is to find parameters $\lambda \in \mathcal{M}$ that minimize the Kullback–Leibler (KL) divergence [22] between the predictive distribution of the reference model and that of the sub model. The KL divergence is defined as:

$$\mathrm{KL}(\mathrm{p}(\tilde{y} \mid \mathcal{D}) \parallel \mathrm{q}(\tilde{y})) = \mathrm{E}_{\lambda^* \mid \mathcal{D}} \left[\mathrm{E}_{\tilde{y} \mid \lambda^*} \log \frac{p(\tilde{y} \mid \lambda^*)}{q(\tilde{y}|\lambda)} \right]$$

where $\mathrm{q}(\tilde{y}|\lambda)$ is the predictive distribution under the projected sub model.

The projection step finds λ for each posterior draw λ^* such that this divergence is minimized. This step effectively transfers the predictive information from the full model into the sub model.

Model Selection via Forward Search The projection predictive method relies on a search part and an evaluation part. The search part defines the predictor ranking, i.e., the best sub model considering each sub model size, where the size is defined by the number of predictor. The evaluation phase then examines how the predictive performance changes as progressively more complex sub models are built by adding predictors in order of their ranking.

Figures and Tables below present the results from the application of projection predictive variable selection method, through the *projpred* R package, to our All Of Us selected cohort, including all the available predictor to fit the full model.

Specifically, in *projpred* R package, there are two main functions for executing the Model Selection through combined search and evaluation: *varsel()* and *cv_varsel*

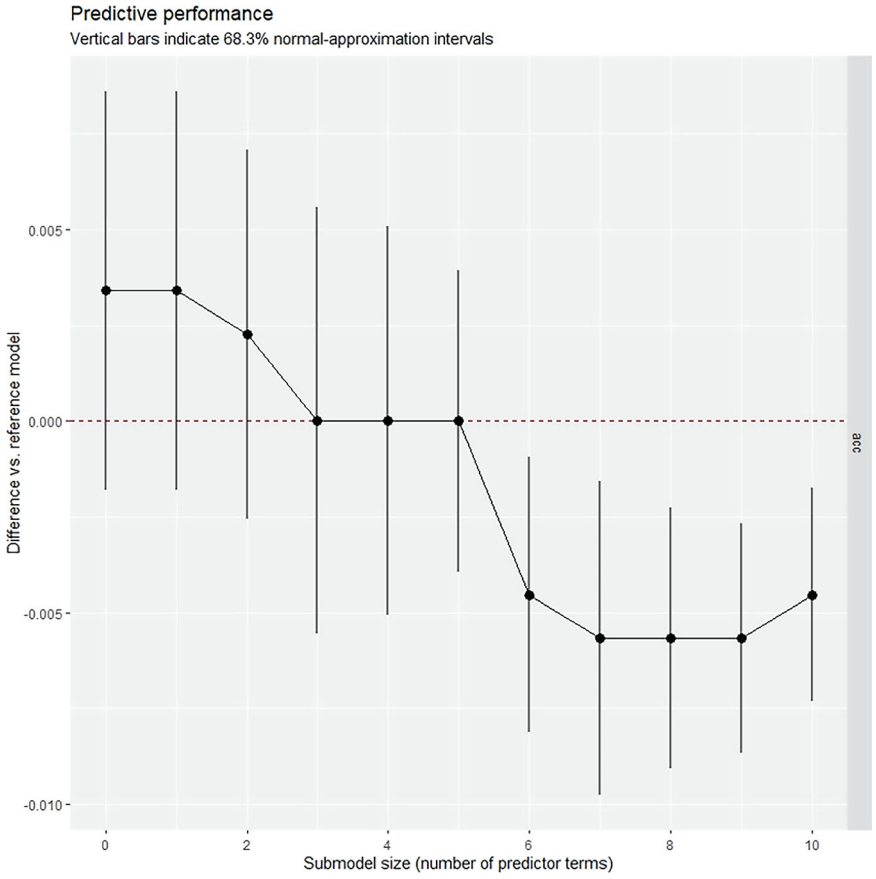

Fig. 3.6 Predictive performance in terms of accuracy: difference between each sub model and the full reference model

(). Specifically, we have used the function *cv_varsel()* to perform the cross-validation (CV). In particular, to perform the CV, we have specified (cv_method = "kfold") with 10 folds.

Figure 3.6 shows the difference in predictive accuracy (ACC) between each sub model and the full reference model, plotted against the number of predictor terms included (i.e., sub model size). The full model is used as the baseline (dashed line at 0). The performance gap decreases as more predictors are added: sub models with 6 or more variables reach a performance not worse than the full model.

In Fig. 3.7, the difference in AUC (area under the ROC curve) is reported for each sub model relative to the reference model. Sub models with fewer than 4 predictors suffer from substantial degradation in AUC, but performance improves sharply with the addition of variables. Starting from 6 predictors, the AUC difference stabilizes and becomes negligible, indicating that most of the discriminative power of the full model can be preserved in a smaller model. The narrower bootstrap intervals around the 6–10 variable range confirm this stability.

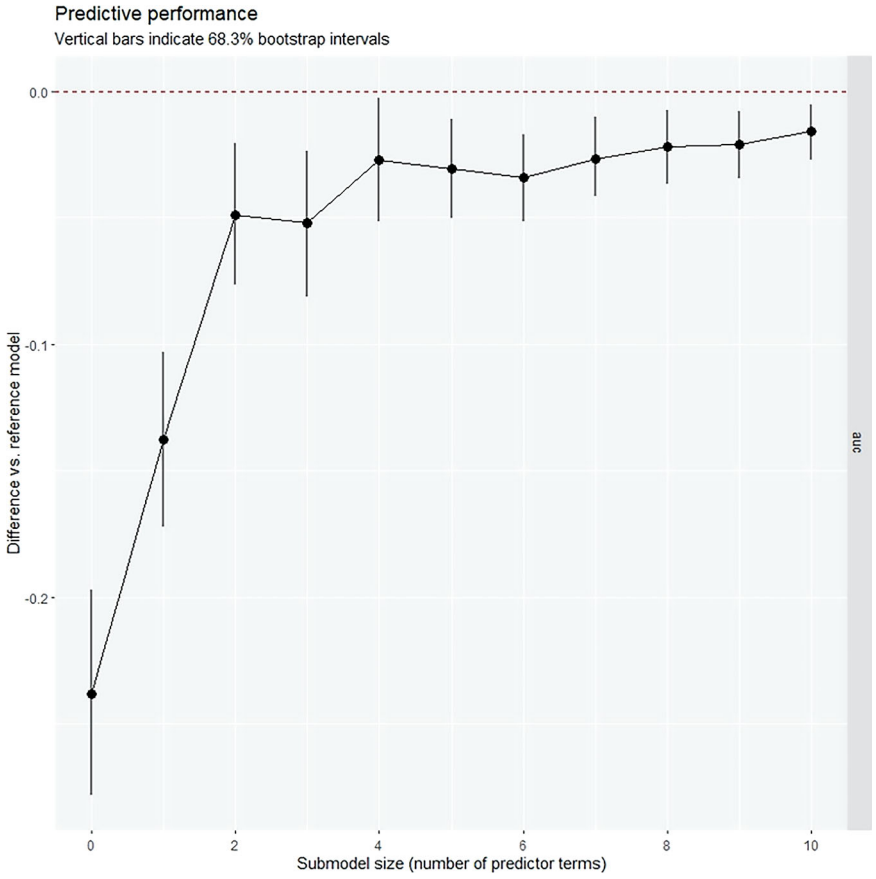

Fig. 3.7 Predictive performance in terms of AUC: difference between each sub model and the full reference model

Figure 3.8 presents a heatmap of variable inclusion frequencies across cross-validation folds as a function of sub model size. Predictors are ordered by their frequency of selection. Intubation and endocrine disorders are the most frequently selected variables, appearing in 100% of folds already in small sub models, suggesting they are highly informative. Transplant and respiratory disorders follow closely, becoming stable components from sub model sizes 4–5 onward. Variables like replacement and ethnicity are included less frequently, and only in larger sub models, suggesting marginal contributions to predictive performance.

Therefore, based on the variable ranking and inclusion stability, the optimal sub model with 6 predictors comprises intubation, endocrine disorders, transplant, respiratory disorders, sex, and endoscopy. Taken together, these figures support the identification of a reduced, high-performing sub model that maintains predictive accuracy close to that of the full model while enhancing interpretability and minimizing overfitting risks.

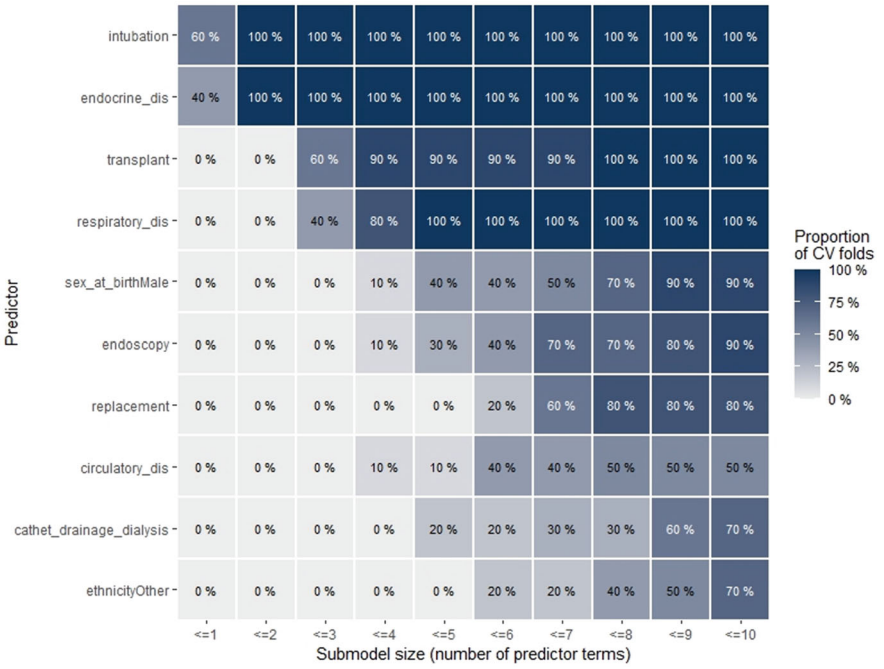

Fig. 3.8 Variable inclusion frequency across sub models of increasing size. Each cell shows the proportion of cross-validation folds in which the predictor was selected

Discussion

This chapter offers a multifaceted examination of variable selection methods for GLMs, with a particular focus on their application in high-stakes healthcare prediction. The comparison of frequentist and Bayesian approaches—grounded in both theoretical considerations and empirical evaluations—underscores the trade-offs between interpretability, flexibility, and predictive performance in real-world modeling tasks.

From a frequentist perspective, the likelihood ratio test and information criteria (AIC/BIC) offer rigorous decision criteria for both nested and non-nested model comparisons, reinforcing the principle of model parsimony when adding predictors or interaction terms did not yield significant improvements. In particular, stepwise regression provided a transparent and interpretable model, identifying a concise subset of nine predictors associated with infection risk. This method, guided by AIC minimization, yielded clinically plausible variables such as intubation, transplant, endocrine and respiratory disorders. However, it also revealed potential limitations: inclusion of variables with borderline significance (e.g., catheterization procedures) and sensitivity to model-building heuristics.

Bayesian methods, by contrast, provided a complementary approach. LOO-CV allowed model comparisons grounded in predictive accuracy, accounting for

uncertainty and reducing overfitting risks. The additive model (with age and endocrine disorder) was consistently preferred over the interaction model, based on both ELPD and LOOIC values. BMA further addressed model uncertainty by aggregating over many plausible sub models, identifying intubation and endocrine disorder as highly robust predictors based on their near-unity PIPs. This probabilistic measure of relevance, unavailable in classical regression, offers a clear and interpretable framework for feature selection, particularly when no single model dominates.

The most interpretable and powerful insights emerged from projection predictive variable selection, which takes a pragmatic and predictive view of variable selection. This approach projected the full model's inferential strength onto smaller sub models, preserving predictive accuracy while enhancing parsimony. From both accuracy (ACC) and AUC perspectives, sub models with six or more predictors matched the full model's performance. The ranking and stability of variable inclusion across cross-validation folds revealed a core subset of predictors (intubation, endocrine disorders, transplant, respiratory disorders, sex, and endoscopy) that jointly maintained performance and interpretability.

A central theme across all methods is the consistent identification of a small number of highly informative predictors. Intubation emerged as the most reliable and influential variable in nearly every analysis, aligning with clinical intuition around invasive procedures as strong infection risk factors. Endocrine disorders and transplant history also maintained prominence, reflecting their association with immune modulation and complex care pathways. Interestingly, demographic variables such as sex and ethnicity showed weaker, less consistent effects, suggesting that clinical history may be more predictive than socio-demographic characteristics in this context.

Methodologically, this chapter illustrates how different selection paradigms approach the bias–variance trade-off. It also reinforces the importance of conceptual clarity in interpreting variable roles. The initial example on confounding and interaction provided a concrete application of theory to real data, highlighting how careful model specification can yield more accurate and actionable insights. Whether a variable acts as a confounder, mediator, or effect modifier has profound implications for both statistical modeling and clinical interpretation.

In summary, the convergence of results across selection strategies offers a robust roadmap for variable selection in epidemiological modeling. When computational resources and expertise allow, Bayesian model selection techniques, particularly projection predictive methods, represent a best-in-class strategy for building interpretable and high-performing predictive models in healthcare and beyond.

References

1. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol. 1989;18:269–74.
2. Monegro AF, Muppidi V, Regunath H. Hospital-acquired infections. 2017.

3. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench. 2012;5:79.
4. Sehgal K, Khanna S. Gut microbiome and Clostridioides difficile infection: a closer look at the microscopic interface. Ther Adv Gastroenterol. 2021;14:1756284821994736.
5. Marlin N, Allotey J. The difference between effect modification and covariate confounding. BJOG. 2021;128:1471–0528.
6. Deltourbe L, Mariano LL, Hreha TN, Hunstad DA, Ingersoll MA. The impact of biological sex on diseases of the urinary tract. Mucosal Immunol. 2022;15:857–66.
7. Buse A. The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note. Am Stat. 1982;36:153–7.
8. Goloboff PA. Parsimony, likelihood, and simplicity. Cladistics. 2003;19:91–103.
9. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017;27:1413–32.
10. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
11. Johnsson T. A procedure for stepwise regression analysis. Stat Pap. 1992;33:21–9.
12. Lewis-Beck MS. Stepwise regression: a caution. Political Methodol. 1978:213–40.
13. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Stat Comput. 2014;24:997–1016.
14. Vehtari A, Simpson D, Gelman A, Yao Y, Gabry J. Pareto smoothed importance sampling. J Mach Learn Res. 2024;25:1–58.
15. Clyde MA, Ghosh J, Littman ML. Bayesian adaptive sampling for variable selection and model averaging. J Comput Graph Stat. 2011;20:80–101.
16. Horvath R. Research & development and growth: a Bayesian model averaging analysis. Econ Model. 2011;28:2669–73.
17. Goutis C, Robert CP. Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections. Biometrika. 1998;85:29–37.
18. Dupuis JA, Robert CP. Variable selection in qualitative models via an entropic explanatory power. J Stat Plan Infer. 2003;111:77–94.
19. Piironen J, Vehtari A. Comparison of Bayesian predictive methods for model selection. Stat Comput. 2017;27:711–35.
20. Catalina A, Bürkner P-C, Vehtari A. Projection predictive inference for generalized linear and additive multilevel models. In: International conference on artificial intelligence and statistics. PMLR; 2022. p. 4446–61.
21. Weber F, Glass Ä, Vehtari A. Projection predictive variable selection for discrete response families with finite support. Comput Stat. 2025;40:701–21.
22. Kullback S. Kullback-Leibler divergence. Tech. Rep; 1951.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.