Bayesian Modelling of Metabolic Syndrome Risk in AVIS Blood Donors

An analysis of routine laboratory data

Noemi Bongiorni, Martina Caliandro, Davide Marchesi, Greta Minazzi, Elisa Nordera, Matteo Zanetti

Tutor: Simone Colombara

Bayesian Statistics 2025/2026

February 19, 2026

POLITECNICO
MILANO 1863

Context and dataset

- 100,203 laboratory measurements.
- 4,300 donors - longitudinal data.
- 5 target variables: `Blood pressure`, `BMI`, `Glucose`, `HDL Cholesterol`, `Triglycerides`.

Research question:

Is it feasible to implement a **Metabolic Syndrome screening**, based on AVIS laboratory measurements?
Is this able to identify the predictive biomarkers for high-risk patients and support clinicians in treatment strategies?

Modeling framework

Hierarchical multivariate Bayesian model:

$$\boldsymbol{y}_{ij} \mid \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma} \overset{\text{ind}}{\sim} \mathcal{N}_K(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, I, \; j = 1, \ldots, n_i$$

$$\boldsymbol{\mu}_{ij} = \boldsymbol{x}_{ij}\,\boldsymbol{\beta} + \boldsymbol{b}_i$$

K : number of target variables (5)

I : number of donors (35)

n_i: number of visits per donor

β: fixed effects

b_i: donor specific random intercept

POLITECNICO MILANO 1863

Residual covariance: LKJ

$$\boldsymbol{\Sigma} = \mathbf{L}_\Sigma \mathbf{L}_\Sigma^\top, \quad \mathbf{L}_\Sigma = \text{diag}(\boldsymbol{\tau})\mathbf{L}_\Omega, \quad \mathbf{L}_\Omega \sim \text{LKJ-Cholesky}(4)$$

$$\tau_k \overset{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5)$$

Random effects: Gaussian

$$\mathbf{b}_i = \mathbf{L}_{\Sigma_b} \mathbf{z}_{b_i}, \qquad \mathbf{z}_{b_i} \overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$

$$\mathbf{L}_{\Sigma_b} = \text{diag}(\boldsymbol{\tau}_b)\mathbf{L}_{\Omega_b}$$

$$\mathbf{L}_{\Omega_b} \sim \text{LKJ-Cholesky}(4)$$

$$\tau_{b_k} \overset{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5)$$

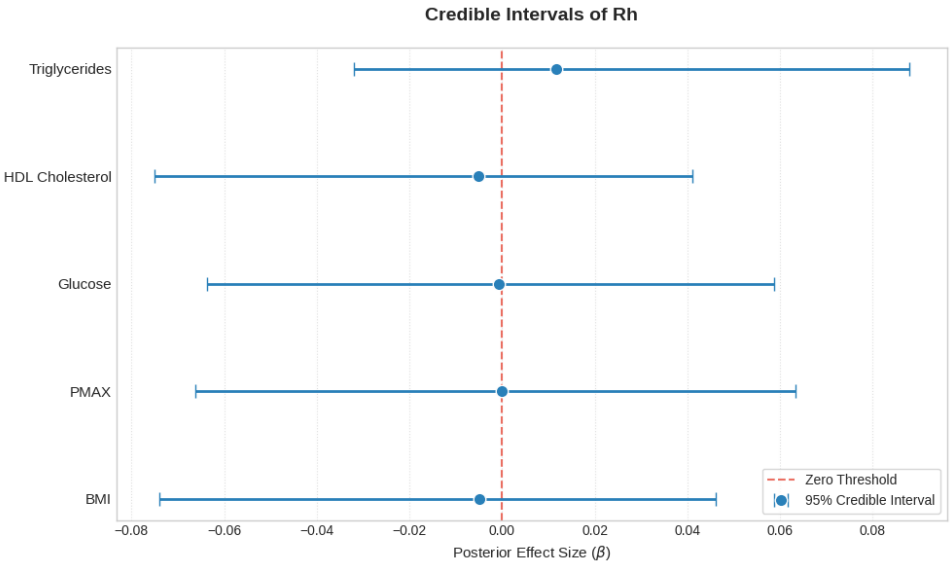
Fixed effects: Bayesian Lasso

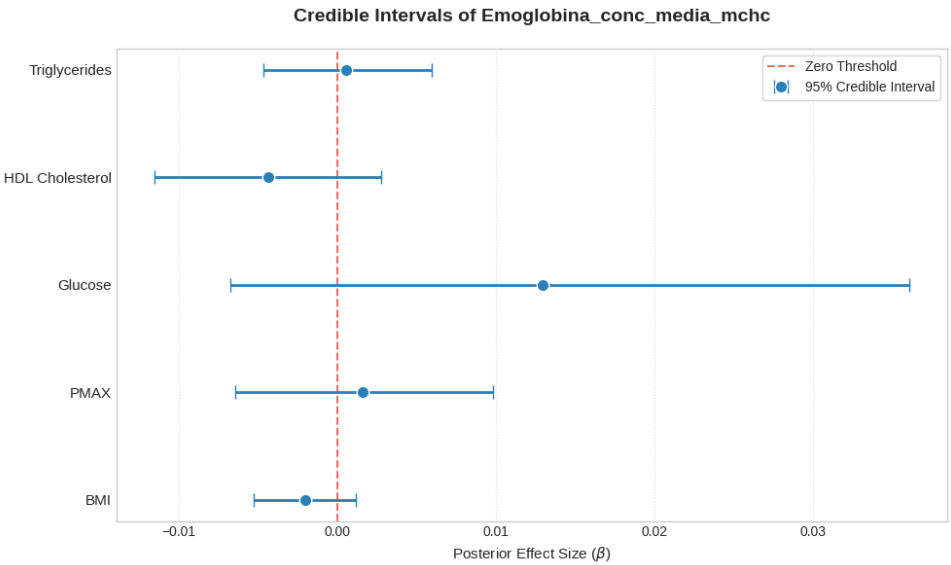
$$\beta_{pk} \mid \tau_{pk} \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau_{pk})$$

$$\tau_{pk} \overset{\text{iid}}{\sim} \text{Exponential}\left(\frac{\lambda^2}{2}\right)$$

$$\lambda \sim \text{Gamma}(1, 1)$$

Variable selection

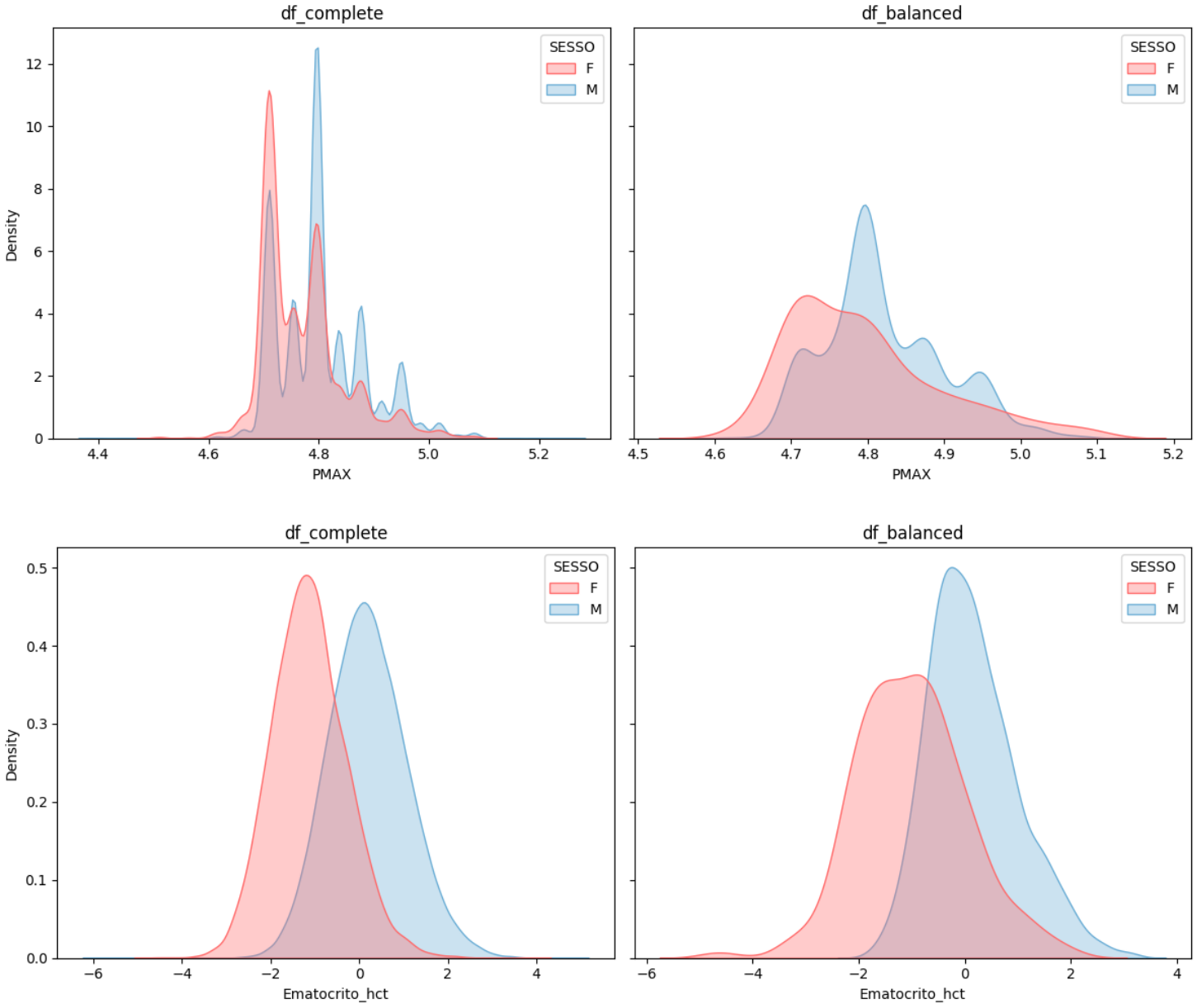




Non-informative:

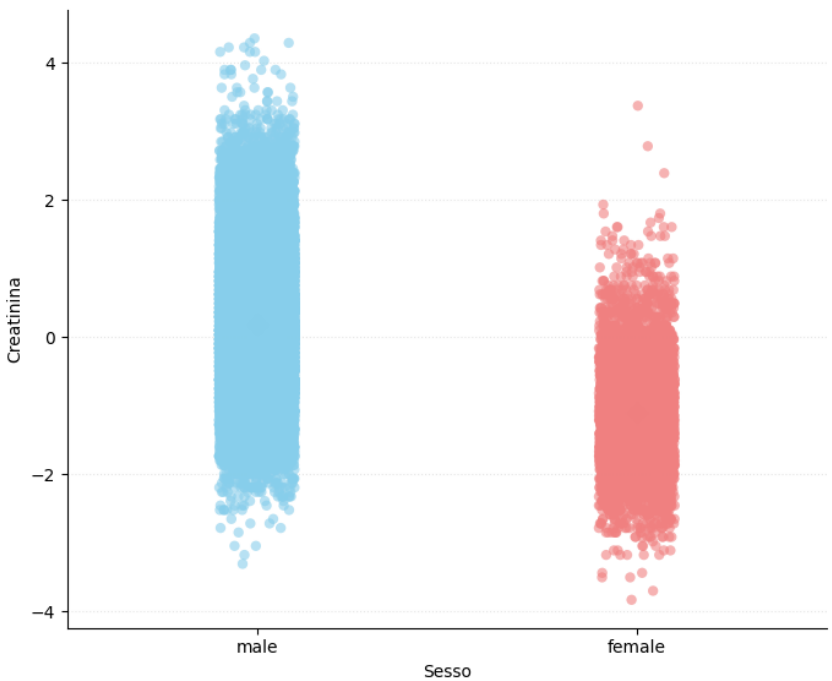
Basophili_perc
Creatinina
Emoglobina
Ferritina
Ferro_totale
Data
Data_nascita
Rh
ABO
SESSO.

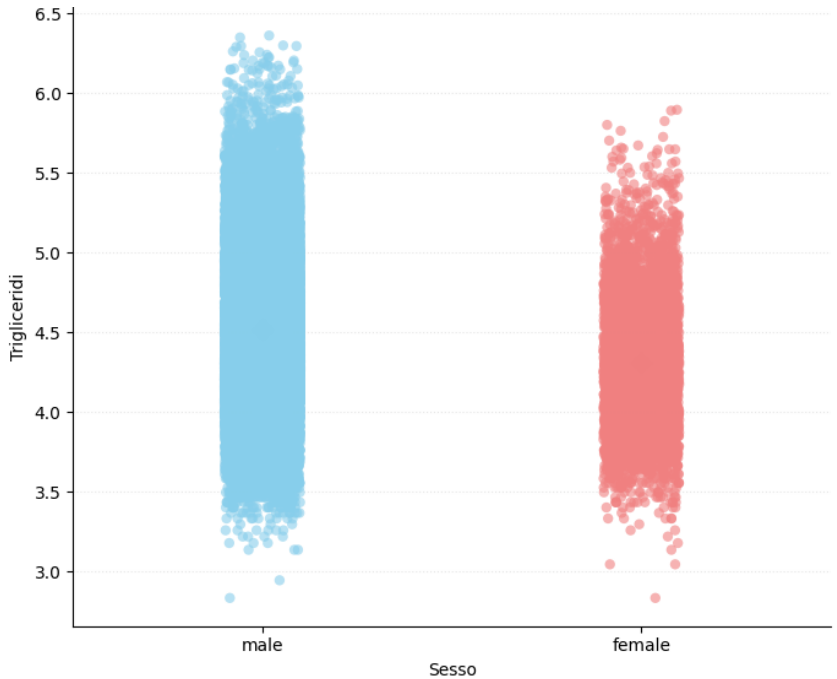
Covariate Sesso



Covariate Sesso

`Creatinina` and `Trigliceridi` values of male and females





DPMM for Random Effects

Idea:

$$\mathbf{b}_i \mid \boldsymbol{\theta}_i \overset{ind}{\sim} k(\cdot; \boldsymbol{\theta}_i)$$

$$\boldsymbol{\theta}_i \mid P \overset{iid}{\sim} P$$

$$P \sim DP(\alpha, P_0)$$

$k(\cdot; \boldsymbol{\theta}_i)$ is a multivariate Gaussian density

$i = 1, \ldots, I$ (35): donor index

Actually:

$$P = \sum_{m=1}^{M} \pi_m \delta_{\boldsymbol{\theta}_m^*}, \qquad \pi_1, \ldots, \pi_M : \text{stick-breaking weights}$$

$\boldsymbol{\theta}_m^* = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad m = 1, \ldots, M$ (20): cluster index

$\boldsymbol{\mu}_m \sim \mathcal{N}_K(\mathbf{0}, 2^2\mathbf{I}), \quad \boldsymbol{\Sigma}_m$ with LKJ prior

$K = 5$: number of targets

Fixed effects

POLITECNICO MILANO 1863

Cluster-level mean random effects

$\boldsymbol{\mu}_m =$

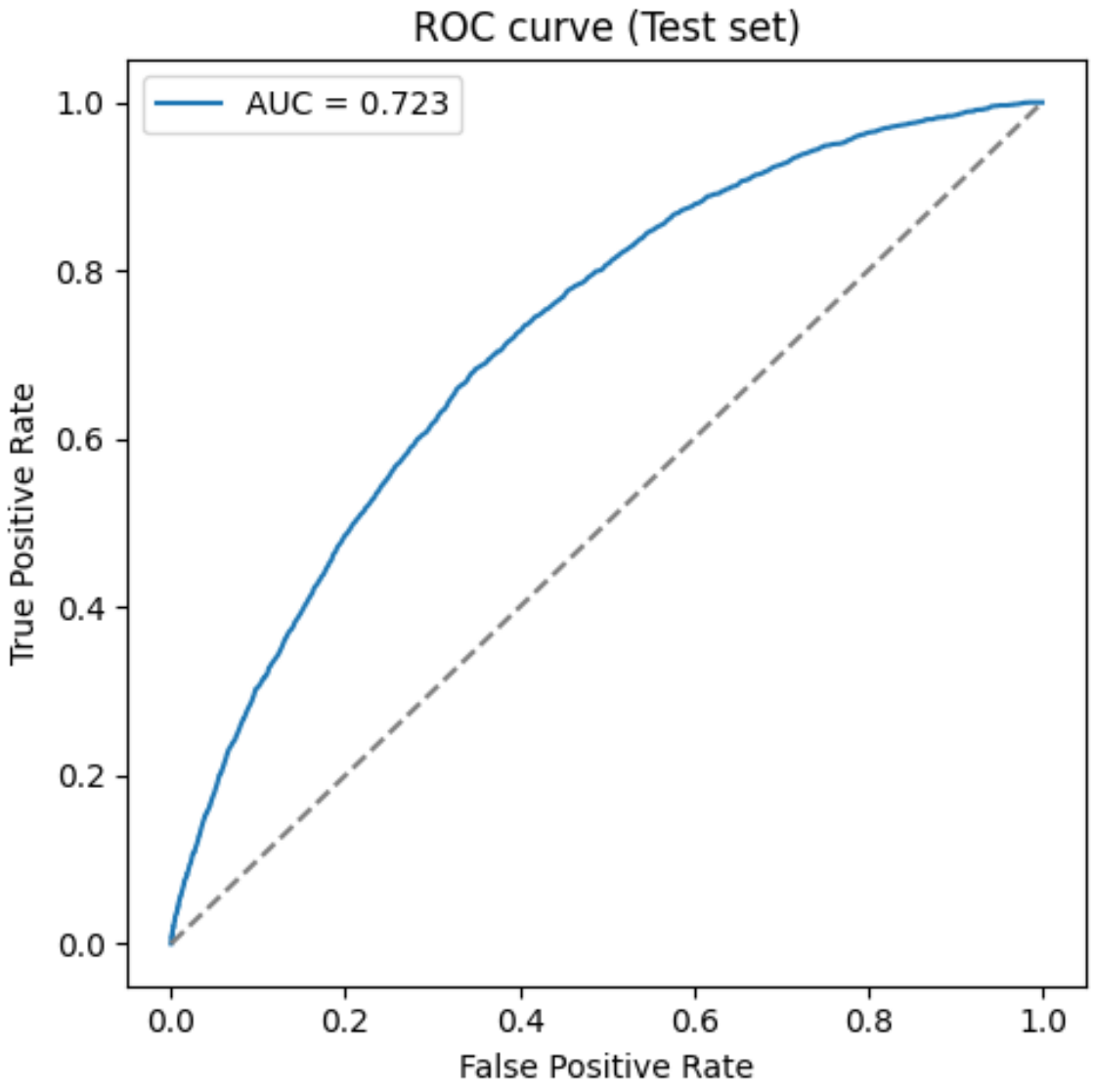
	PMAX	Glucosio	Trigliceridi	HDL	BMI
Cluster 1	4.7929	4.5373	4.6824	3.9904	3.2908

Covariance matrix of random effects

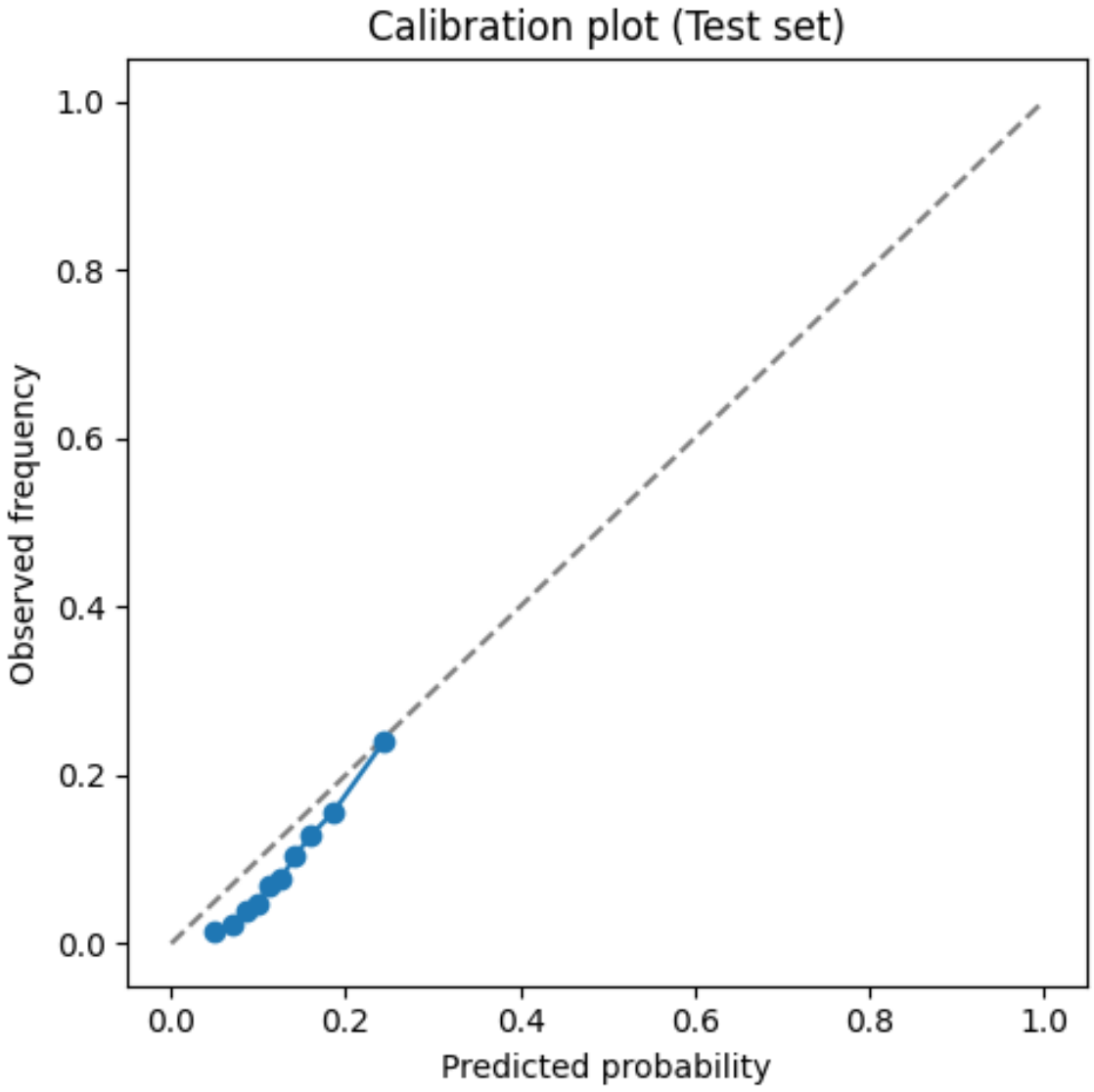
$\boldsymbol{\Sigma}_m =$

	P	G	T	H	B
P	0.0020	0.0008	0.0014	−0.0010	0.0016
G	0.0008	0.0044	0.0015	−0.0030	0.0021
T	0.0014	0.0015	0.0856	−0.0279	0.0131
H	−0.0010	−0.0030	−0.0279	0.0303	−0.0051
B	0.0016	0.0021	0.0131	−0.0051	0.0114

Predictive performance: ROC curve



Predictive performance: calibration plot



Conclusions

- Strong predictive performance.
- Interpretable associations at the population level.
- No evidence of latent subgroups: population appears continuous.
- Flexible hierarchical Bayesian model providing an early risk flag and interpretable biomarkers.

References

- Park, T., & Casella, G. (2008). *The Bayesian Lasso*. Journal of the American Statistical Association, 103(482), 681–686. https://doi.org/10.1198/016214508000000337
- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). *The horseshoe estimator for sparse signals*. Biometrika, 97(2), 465–480. https://doi.org/10.1093/biomet/asq017
- Piironen, J., & Vehtari, A. (2017). *Sparsity information and regularization in the horseshoe and other shrinkage priors*. Electronic Journal of Statistics, 11(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI
- Yang, S., Rouder, J. N. Assessing Two Common Priors of Covariance in Hierarchical Designs. *OFS*. https://doi.org/10.31234/osf.io/jen65_v2
- Guglielmi, A. (2025). [Bayesian Statistics Slides]. WeBeep, Politecnico di Milano.

The complete source code and the scripts for this project are available on GitHub at the following repository:

`https://github.com/elisanordera/Bayesian-Statistics-Project`.