MSc. in Mathematical Engineering A.Y. 2025/2026
Project Report of Bayesian Statistics (052499) – Prof. Alessandra Guglielmi
Supervisor: Simone Colombara

Bayesian Modeling of Metabolic Syndrome Risk in AVIS Blood Donors

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

10774710, 10797483, 10710741, 10765748, 10813868, 10765931

February 19, 2026

Abstract

This study presents a Bayesian hierarchical framework for the early screening of Metabolic Syndrome (MetS) based on longitudinal data from AVIS blood donors. To address the complexity of real-world clinical data, including a high rate of missingness and data-entry inconsistencies, we integrate a robust preprocessing using Multiple Imputation by Chained Equations (MICE) with a two level predictive mean matching approach. Our methodology follows a two-stage workflow: first, dimensionality reduction via Bayesian shrinkage (Horseshoe, Lasso, and Ridge) to identify significant covariates; second, the implementation of a multivariate hierarchical model to predict the five metabolic components, enabling the derivation of personalized risk profiles for Metabolic Syndrome. A Dirichlet Process (DP) prior is employed to model subject-specific random effects, providing a flexible non-parametric structure capable of capturing latent heterogeneity. While the DP model achieves the highest predictive performance, the results indicate a continuous distribution of risk rather than distinct latent clusters. The final model achieves a solid discriminative performance (AUC 0.723), proving effective in identifying individuals at higher risk of developing Metabolic Syndrome. These findings demonstrate that routinely collected donor data can be successfully transformed into a proactive decision-support tool. Indeed this research could enable clinicians to prioritize early lifestyle interventions and longitudinal monitoring, shifting the medical focus from reactive diagnosis to true prevention.

1 Introduction

This project arises from a collaboration with AVIS (Associazione Volontari Italiani del Sangue) and aims to study the risk of Metabolic Syndrome (MetS) using information collected during routine check-ups and blood test results from blood donors.

From a clinical perspective, MetS is particularly relevant because it represents a set of metabolic and cardiovascular alterations strongly associated with an increased risk of developing type 2 diabetes and cardiovascular diseases. The five components characterizing it are: abdominal obesity, fasting glucose, triglycerides, HDL cholesterol, and blood pressure. In the literature, abdominal obesity is often assessed using waist circumference; however, in this study, it is replaced with `BMI` (Body Mass Index), defined as body mass divided by height squared, since weight and height are measured more consistently in the dataset and are therefore more reliable. In particular, the adopted threshold is `BMI` ≥ 30, interpreted as an indicator of clinically relevant obesity. Metabolic Syndrome is therefore diagnosed when at least 3 out of 5 components are altered with respect to the adopted clinical thresholds.

Component	**Men**	**Women**
BMI (kg/m^2)	≥ 30	≥ 30
Blood pressure (mmHg)	$\geq 130/\geq 85$	SAME
Fasting Glucose (mg/dL)	≥ 100	SAME
Triglycerides (mg/dL)	≥ 150	SAME
HDL cholesterol (mg/dL)	< 40	< 50

Table 1: Clinical thresholds used for the identification of Metabolic Syndrome

The main idea is to leverage these routinely collected data to develop a potential screening tool, with two primary goals: identifying predictive biomarkers for high-risk donors and supporting physicians in prevention and follow-up strategies. More specifically, the research question concerns whether routinely collected donor data can be used to build a practical and interpretable risk screening approach that helps physicians identify donors who are likely to meet the criteria for Metabolic Syndrome, even when the condition is not yet clinically evident. The intended use is decision support rather than diagnosis: the tool provides an early "risk flag" and highlights the most informative biomarkers, so that clinicians can prioritize follow-up, recommend targeted lifestyle interventions, and, when appropriate, suggest additional clinical evaluation. In preventive medicine, identifying high-risk profiles early enables faster clinical monitoring and more effective lifestyle interventions.

2 Dataset presentation

The dataset contains observations collected over the period 2009–2024. Each donor is uniquely identified by the variable `CAI` (Codice Anonimo Identificativo). Each row corresponds to a single visit, and the same donor may appear multiple times; therefore, the dataset has a longitudinal structure.

The initial dataset consists of 100 203 rows and 42 variables. Variables include body measurements (e.g., `Peso`), clinical parameters (e.g., `PMAX`), and laboratory metabolic biomarkers (e.g., `Glucosio`), as well as demographic variables (e.g., `Sesso`) and the visit date (`Data`).

Since age is an important determinant of metabolic risk, the covariate `Età` is created from the birth date (`DATA_NASCITA`) and the visit date (`Data`), after converting both to date format with the `pandas` function *pd.to_datetime*.

3 Exploratory data analysis

Before fitting any model, an exploratory data analysis (EDA) is conducted to assess data quality, detect missing values (NaN), and verify the internal consistency of repeated measurements over time. As a first step, histograms are produced for all numerical variables, also reporting the count of missing values for each variable. This approach helps to check the shape of the distributions and identify clearly impossible values. They are available on our github page.

3.1 Analysis of the “Altezza” variable

The most evident issue emerges in height measurements: in the full scatter plot of observations, clearly out-of-scale values appear, likely due to data-entry or measurement errors. To better visualize the realistic range, a second “zoomed” plot is created; even in this view, values incompatible with physiological ranges are found, motivating a targeted cleaning procedure.

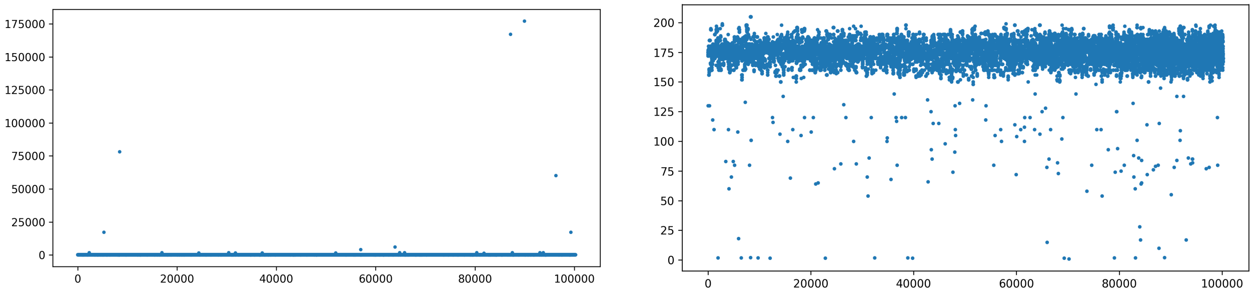


Figure 1: Scatter plots of the initial distribution of the height covariate.

Since each donor can appear in multiple visits, height should remain essentially constant over time for the same individual. Therefore, the variable `CAI` is used to assess within-donor consistency. The correction is performed in two steps:

- Cleaning based on a physiological range: for each donor, the mode of height is computed considering only values in the interval 145–205 cm. Then, for that same donor, both missing values and values outside the physiological range are replaced with the mode.
- Correction of isolated anomalous values with respect to the donor-specific mode: after the coarse correction, a second procedure is applied to capture values that are still inconsistent but not necessarily outside the physiological range. For each `CAI`, the mode is recomputed and values differing from the mode by more than 9 cm are flagged as deviant. To avoid excessive corrections in unstable cases, replacement is applied only to donors with a limited number of inconsistencies (at most 4). With this criterion, 285 height values are modified.

Overall, the procedure effectively stabilizes the `Altezza` variable against data-entry errors. A final scatter plot of height is produced to visually show the distribution after cleaning and compare it with the initial one.

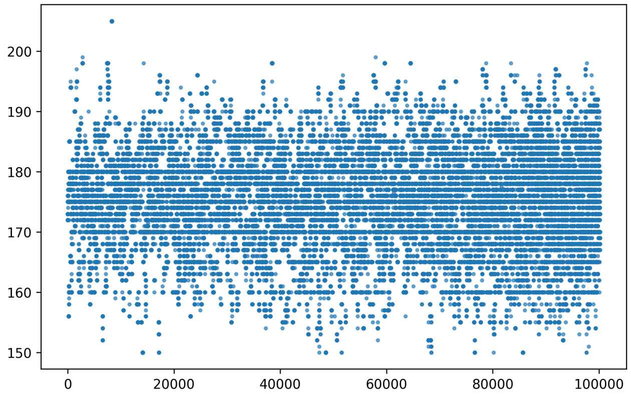


Figure 2: Scatter plot of the final distribution of the height covariate.

The remaining variables are not modified as they are less prone to human error; most of them are automatically entered by medical devices and, to the best of our knowledge, represent reliable clinical measurements.

4 Missing Values

A central issue in the dataset is the presence of a substantial amount of missing values. For some covariates, the proportion of missingness is very high (in some cases close to 80%), making it necessary to define a strategy for handling missing data before modeling. Before adopting a final approach, several imputation alternatives are evaluated:

- Missing values are replaced using within-patient median imputation for the variables `Altezza, Peso` and `Polso`, and then rows still containing missing values are removed. This solution exploits the longitudinal structure but remains limited because it applies only to a small number of covariates.
- Missing values are imputed across all columns using the median. This is a simple method that produces a complete dataset, but it may introduce distortions and tends to artificially reduce variability.
- KNN (K-Nearest Neighbors) imputation is tested, replacing missing data using the median computed from the most similar observations. To better account for correlation among covariates, the Mahalanobis distance is considered.

While these methodologies address the issue of missing values, we seek a more consistent approach to ensure the robustness and reliability of the data.

4.1 MICE

Multiple Imputation by Chained Equations (MICE) is selected as the final imputation strategy, as it is particularly well suited to preserving multivariate relationships among biomarkers, avoiding overly simplistic imputations, and handling high-dimensional datasets with complex missingness patterns.

As a first step, observations containing missing values in the target variables are removed from the dataset to prevent imputation-related uncertainty from directly affecting the outcomes of interest. Subsequently, covariates with more than 70% missingness are excluded, as imputing such variables would produce unreliable and strongly model-driven estimates. MICE is then applied only to the remaining covariates.

The imputation procedure is implemented using the `mice` package in `R`, with the algorithm explicitly specified to operate under a two-level predictive mean matching framework (`2l.pmm`). This choice is made to correctly account for the clustered structure of the data and to preserve within-patient dependencies. The variable `CAI` is defined as the multilevel grouping factor, allowing the imputation models to incorporate patient-specific random effects and appropriately model repeated measurements.

MICE imputes missing values by specifying a sequence of conditional models, one for each incomplete variable, where each variable is regressed on all others. Importantly, the algorithm does not replace missing values with simple regression predictions. Instead, for each missing observation, a regression model is used only to obtain predicted values, which are then employed to identify a set of observed cases with similar predicted means. The imputed value is randomly drawn from the actual observed values within this matched set. This predictive mean matching mechanism ensures that all imputations correspond to values that actually occur in the dataset, preserving the empirical distributions of the variables.

5 Preprocessing

A preprocessing phase is carried out to make variables more suitable for statistical modeling and to reduce instability due to different scales, strong skewness, and collinearity among covariates. A log-transformation is applied to all target variables and to covariates with high sample skewness (using $skewness > 0.5$ as an operational criterion). Subsequently, covariates are standardized.

High redundancy is observed among some variables (high collinearity). This can make estimated coefficients unstable, increase estimation variability, and complicate interpretation. For this reason, a manual reduction of covariates is performed based on correlation analysis:

- Among hematocrit (`Ematocrito_hct`), hemoglobin (`Emoglobina_hb`), and RBC (`Eritrociti_rbc`), a high correlation is found (e.g., $\rho_{\text{Hct, Hb}} \approx 0.88$ and $\rho_{\text{Hct, RBC}} \approx 0.80$).
  To avoid redundancy, only `Ematocrito_hct` is retained as the most reprsentative.

- Between MCH (`Emoglobina_massa_media_mch`) and Mean Corpuscular Volume (`Volume_medio`), a high correlation is observed ($\rho_{\text{MCH, MCV}} \approx 0.81$).
  Therefore, `Volume_medio` is retained and `Emoglobina_massa_media_mch` is removed.
- `Peso` and `Altezza` are removed because body measurement information is already summarized by `BMI`, which is more directly interpretable.

Finally the resulting dataset has 35 853 observations, 23 covariates and 5 target variables.
For all the modeling, a reduced dataset is generated to ensure computational efficiency, maintaining a total of approximately 1,000 observations. To remain consistent with the original population, we calculate the proportion of males and females in the complete dataset and perform a stratified selection.

Specifically, we extract 28 male and 7 female donors, selecting those with the highest number of recorded visits for each sex to maximize longitudinal available data while preserving the biological variability and the demographic structure of the original cohort.

The obtained dataset used to run our models is *df_balanced*.

6 Modeling Framework

A Bayesian mixed effects hierarchical model serves as the fundamental framework throughout our work, suitable for this scenario since it allows patient-specific random intercepts to account for within-subject correlation while modeling variation across individuals.

The workflow is conducted in two distinct phases:

- We implement four models by fixing the prior for random effects to a Gaussian distribution while varying the shrinkage prior for the fixed effects. The primary objective of this stage is to perform dimensionality reduction by identifying and removing non-informative covariates. *(see section 7)*
- We execute three models on the reduced dataset (excluding the covariates identified in the previous step). In this phase, we fix the prior for fixed effects to a Gaussian distribution and vary the priors for the random effects. This aims to enhance the predictive performance of the model and improve the clinical interpretability of the results. *(see section 8)*

Note: the underlying assumption, verified by external sources, is that observations can be considered conditionally independent, given the model parameters ($\boldsymbol{\beta}$ and $\boldsymbol{b_i}$). This implies that the random intercepts account for all within-subject correlation, leaving the residuals ($\boldsymbol{\epsilon_{ij}}$) as independent and identically distributed (i.i.d.) noise.

6.1 Technical Setting

`Python` and `Stan` models are the software used to implement the models. `R` is used as a support to implement MICE.

7 Covariate selection

To achieve the goal of accurately predicting the vector of the five target variables, the first step involves reducing the number of covariates. This dimensionality reduction aims to retain only the features that contribute most to the output variability while improving the computational efficiency of the models.

7.1 Methodology

In order to manage the 23 covariates obtained from the preprocessing (`Alanina_aminotransferasi_alt`, `Basofili_perc`, `Colesterolo_totale`, `Creatinina`, `Distribuzione_di_volume`, `Ematocrito_hct`, `Emoglobina_conc_media_mchc`, `Eosinofili_perc`, `Ferritina`, `Ferro_totale`, `Leucociti_wbc`, `Linfociti_perc`, `Monociti_perc`, `Piastrine`, `Polso`, `Proteine_totali`, `Volume_medio`, `Eta`, `CAI`, `Date`, `DATA_NASCITA`, `PMAX`, `Glucosio`, `Trigliceridi`, `Colesterolo_Hdl`, `BMI`, `SESSO`, `Rh`, `ABO`), a shrinkage approach is adopted, which identifies the variables showing the most influence on MetS.

For this purpose, different prior distributions on fixed effects are considered: Horseshoe, Regularized Horseshoe, Lasso, and Ridge. The Horseshoe prior is a hierarchical global–local shrinkage prior with a sharp spike at zero and heavy tails, allowing strong suppression of noise coefficients while preserving large signals. The Regularized Horseshoe prior adds a controlled slab width to stabilize shrinkage for large coefficients, combining adaptive sparsity with greater regularization control. In contrast, the Bayesian Ridge prior uses independent normal distributions centered at zero, imposing a global L_2-type shrinkage that pulls all coefficients toward zero without inducing sparsity. Finally, the Bayesian Lasso uses a double-exponential (Laplace) prior, which leads to stronger shrinkage near zero and can yield sparse estimates.

To model the residual and random effects' covariance matrices, an Inverse-Wishart distribution and an LKJ (Lewandowski-Kurowicka-Joe) distribution are explored to address computational issues. The Inverse-Wishart prior, as the classic conjugate prior for covariance matrices in multivariate normal models, simplifies computation but tends to link variances and correlations together in ways that can bias posterior estimates when variances are small or data are limited. In contrast, the LKJ prior is placed directly on the correlation matrix and used together with different priors on each standard deviation, effectively decomposing the covariance into scales and correlations and allowing more flexible and interpretable control of the covariance structure without imposing strong a priori dependency. The LKJ Cholesky decomposition with concentration parameter η follows the approach below:

$$\Sigma = LL^\top, \quad R \sim \text{LKJ}(\eta), \quad R = L_R L_R^\top, \quad L = \text{diag}(\sigma)L_R$$

Based on these considerations, different models are fitted under the same optimal parameter setting: 4 parallel chains, 1500 warmup iterations, 1000 sampling iterations, `adapt_delta` equal to 0.9, and a `max_treedepth` of 12. Moreover, the intercept is not included, and a jittering procedure is used to avoid numerical issues during sampling.

Before introducing the different models we test, we present the common hierarchical structure underlying all of them. This structure is a multivariate Gaussian linear model in which the mean component accounts for both the fixed effects, $\boldsymbol{\beta}$, and the random effects, $\mathbf{b}$:

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

$$\boldsymbol{\mu}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{b}_i$$

Here and for the rest of the report, K denotes the number of target variables (5) indexed by k, I is the number of donors indexed by i, and n_i represents the number of visits per donor indexed by j.

Note: $\boldsymbol{\beta}$ is a matrix since for each covariate, indexed by $p = 1, \dots, P$ (23), there are K (5) coefficients, one for each target.

Note: For the rest of the report $\mathbf{I}_K$ denotes the five-dimensional indicator matrix.

Model 1, Model 2, and Model 4 were run using, respectively: a regularized horseshoe prior for the fixed effects combined with an inverse-Wishart prior for the residual covariance (1); a regularized horseshoe prior for the fixed effects combined with an LKJ prior for the residual covariance (2); and a Bayesian ridge prior for the fixed effects combined with an LKJ prior for the residual covariance (4). Detailed specifications for these models are detailed in Appendix C. Model 3, which demonstrated the best performance, is reported below.

7.1.1 Bayesian Lasso LKJ (Model 3)

Residual covariance: LKJ

$$\boldsymbol{\Sigma} = \mathbf{L}_\Sigma \mathbf{L}_\Sigma^\top$$

$$\mathbf{L}_\Sigma = \text{diag}(\boldsymbol{\tau})\mathbf{L}_\Omega$$

$$\tau_k \overset{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5), \qquad \mathbf{L}_\Omega \sim \text{LKJ-Cholesky}(4)$$

Random effects:

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

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

$$\tau_{b_k} \overset{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5), \qquad \mathbf{L}_{\Omega_b} \sim \text{LKJ-Cholesky}(4)$$

Fixed effects: Bayesian Lasso

$$\beta_{pk} \mid \tau_{pk} \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau_{pk})$$
$$\tau_{pk} \overset{\text{iid}}{\sim} \text{Exponential}\left(\frac{\lambda^2}{2}\right)$$
$$\lambda \sim \text{Gamma}(1, 1)$$

7.2 Comparison and Comments

The selection of the most effective shrinkage model for variable selection is based on the *Expected Log Predictive Density.*

The ELPD is defined as the sum of the log predictive densities for each observation, under the posterior distribution:

$$\text{ELPD} = \sum_{i=1}^{n} \log \int p(y_i \mid \theta)\, p(\theta \mid y)\, d\theta$$

where y_i is the i-th observed data point, θ are the model parameters, $p(\theta \mid y)$ is the posterior distribution of the parameters given the data. To evaluate our models, it is estimated through Leave-One-Out cross validation.

The ELPD-LOO is a Bayesian measure of out-of-sample predictive accuracy. Unlike simple likelihood measures, it estimates how well the model predicts new data by iteratively training the model on $n - 1$ observations and evaluating it on the excluded point. Higher (less negative) values of ELPD indicate a better predictive performance, as they represent a higher probability assigned by the model to previously unseen data, while naturally accounting for the effective number of parameters and model over-fitting.

Using this criterion, the results can be summarized in the table below:

Model	Prior	ELPD
Model 1	Reg Horse	5399.057403
Model 2	LKJ + Reg Horse	failed
Model 3	LKJ + Lasso	5521.463091
Model 4	LKJ + Ridge	5502.184883

Table 2: Comparison of models.

Model 2 is deemed unsuitable because the run shows that nearly all iterations reach the maximum tree depth, leading to its exclusion.

Model 3 reveals itself to be the most suitable for the current problem. Based on that, incorporating insights from Model 1 and Model 4 as well, a variable selection study is performed by using the Hard Shrinkage criterion, analyzing the 95% Bayesian Credible Intervals for the fixed effects across all targets. Specifically, covariates are considered candidates for exclusion only if their intervals contain zero across all targets. Based on this analysis, the following variables are found to be non-significant: `Basofili_perc`, `Creatinina`, `Emoglobina_conc_media_mchc`, `Ferritina`, `Ferro_totale`, `Data`, `DATA_NASCITA`, `Rh`, `ABO`, `SESSO`.

7.3 Alternative Investigated Approaches

The final methodology is the result of an iterative exploration of several alternative approaches. Although some proved unsuccessful, they provided critical insightss. In particular, ee initially evaluate simple non-shrinking priors for the $\boldsymbol{\beta}$ coefficients (e.g., weakly informative gaussian prior). However, this approach introduced significant noise into the model due to the inclusion of non-significant covariates and high computational costs, highlighting the necessity of a sparse shrinkage prior.

8 Random effects Analysis

In the second stage of our analysis, we focus on the prior specifications for the random effects. To allow the data to drive the posterior estimates more freely, we replace high-shrinkage prior with weakly informative distributions, since covariance reduction has already been performed. Specifically, the fixed effects $\boldsymbol{\beta}$ and the residual covariance matrix Σ are governed by:

$$\boldsymbol{\beta}_k \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{0}, 2^2\mathbf{I}), \quad k = 1, \dots, K$$
$$\boldsymbol{\Sigma} = \mathbf{L}_\Sigma \mathbf{L}_\Sigma^\top, \quad \mathbf{L}_\Sigma = \text{diag}(\boldsymbol{\tau})\mathbf{L}_\Omega, \quad \mathbf{L}_\Omega \sim \text{LKJ-Cholesky}(4), \quad \tau_k \sim \mathcal{N}^+(0, 0.5^2), \quad k = 1, \dots, K$$

where $\mathbf{I}$ is the 13-dimensional identity matrix, since $\boldsymbol{\beta}_k$ is the 13-dimensional vector of independent components containing the fixed effects for the covariates remaining after the variable reduction.

The core of our comparative analysis lies in the modeling of the donor specific random effects $\mathbf{b}$. These are donor-specific random intercepts that account for similarities in the rows belonging to the same person. We implement and evaluate three distinct hierarchical structures, ranging from standard Gaussian assumptions to non-parametric Bayesian techniques. The final model selected for our analysis is illustrated in the following section, while the alternative specifications tested during the selection process—specifically the Normal-Normal LKJ, and the Normal-t LKJ are detailed in Appendix D. All these models are run on *df_reduced*, a dataset obtained by removing non-significant covariates from *df_balanced*.

8.1 Dirichlet Process Mixture Model (Model 7)

To account for potential heterogeneity and latent grouping in the random effects, we exploit a Dirichlet Process prior. In particular, we used a truncated Dirichlet Process mixture model with a Gaussian kernel. Following the hierarchical representation, the model is specified as follows:

$$\mathbf{b}_i \mid \boldsymbol{\theta}_i \overset{ind}{\sim} k(\cdot, \boldsymbol{\theta}_i), \quad i = 1, \dots, I$$
$$\boldsymbol{\theta}_i \mid P \overset{iid}{\sim} P \quad i = 1, \dots, I$$
$$P \sim DP(\alpha, P_0)$$

In this framework, $\boldsymbol{\theta}_i$ represents the latent parameter vector assigned to the i-th donor. The discrete nature of the Dirichlet Process P implies a positive probability of ties among the $\boldsymbol{\theta}_i$, which induces a clustering of subjects. Specifically, since P is discrete, the subject-specific parameters $\boldsymbol{\theta}_i$ take values in a finite set of M unique atoms, denoted as $\boldsymbol{\theta}_m^* = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, for $m = 1, \dots, M$. $M = 20$ is the truncated number of clusters adapted for computational efficiency on `Stan`. The random measure P is expressed via the truncated stick-breaking construction:

$$P = \sum_{m=1}^{M} \pi_m \delta_{\boldsymbol{\theta}_m^*}, \quad \boldsymbol{\theta}_m^* = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

where the weights π_m are determined by the concentration parameter α and the auxiliary variables $v_m \overset{iid}{\sim} \text{Beta}(1, \alpha)$ as follows:

$$\pi_1 = v_1, \quad \pi_m = v_m \prod_{j=1}^{m-1}(1 - v_j), \quad \pi_M = 1 - \sum_{j=1}^{M-1} \pi_j$$

In our implementation, the kernel $k(\cdot, \boldsymbol{\theta}_i)$ is a multivariate Gaussian density. Because each $\boldsymbol{\theta}_i$ maps to one of the unique cluster atoms $\boldsymbol{\theta}_m^*$, the marginal density of the random effects $\mathbf{b}_i$ results in a continuous mixture of Gaussian kernels:

$$f(\mathbf{b}_i) = \sum_{m=1}^{M} \pi_m \mathcal{N}_K(\mathbf{b}_i \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

The model is completed by the following prior specifications:

- Concentration parameter: $\alpha \sim \text{Gamma}(1, 1)$.
- Base measure P_0: it is the distribution from which the unique cluster parameters (atoms) $\boldsymbol{\theta}_m^* = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ are sampled. Each cluster has its own mean $\boldsymbol{\mu}_m \sim \mathcal{N}_K(\mathbf{0}, 2^2\mathbf{I})$, while the clsuter-specific covariance matrix $\boldsymbol{\Sigma}_m$ follows the LKJ prior distribution described also for the models above.

The idea behind using a Dirichlet Process Mixture Model is that instead of assuming a fixed number of groups or a normal distribution, it lets the data determine the number of groups and their weights, allowing similar random effects to share cluster-specific parameters. Importantly, even if the DP does not identify clear clusters in the data, it still serves as a robust prior, encouraging the sharing of information across random effects, which improves regularization and stabilizes estimates.

8.2 Comparison and comments

In the following table, we report the Expected Log Predictive Density (ELPD) for the three hierarchical models described above.

Model	Prior	ELPD
Model 5	Normal-Normal	5515.973349
Model 6	Normal-t-Stud + Reg Horse	5517.265821
Model 7	DP	5517.958127

Table 3: Comparison of models.

Finally, we select Model 7 as the definitive framework for our predictions and final analysis. The DP model achieves the highest ELPD-LOO value among the candidates; however, since all models show very similar predictive metrics, our preference for the Dirichlet Process approach is further justified by its distinct structural advantages. In particular, unlike standard Gaussian approaches, the DP prior allows the model to naturally capture the heterogeneous structure of the dataset by identifying latent clusters within the random effects, providing a more flexible and robust density estimation. Moreover, this approach is more generalizable, as it is better equipped to accommodate the addition of new data or a future expansion of the dataset without compromising the model's inferential integrity.

9 Model Interpretation

The following interpretations and conclusions are based on the test set obtained by removing the visits used in the training set from our original dataset.

9.1 Population-level associations across metabolic outcomes

We fit a multivariate hierarchical Bayesian model to jointly describe five log-transformed metabolic outcomes (`PMAX`, `Glucosio`, `Trigliceridi`, `Colesterolo_HDL`, `BMI`) as a function of standardized covariates and patient-specific random effects. Because covariates are standardized prior to model fitting, each fixed-effect coefficient represents the expected change in the log biomarker associated with a one standard deviation increase in the corresponding covariate, holding other predictors constant.

Posterior summaries of the fixed effects reveal systematic population-level associations between covariates and metabolic outcomes. Several predictors exhibit consistent directional effects across multiple biomarkers, indicating shared physiological pathways underlying different components of metabolic risk. The multivariate formulation allows information to be shared across outcomes, resulting in more stable and coherent estimates than separate univariate regressions.

The heatmap of posterior median coefficients highlights the global structure of associations, making it clear which covariates increase or decrease specific biomarkers. For individual outcomes, plots of the most influential covariates further quantify effect magnitudes and associated uncertainty through 95% credible intervals.

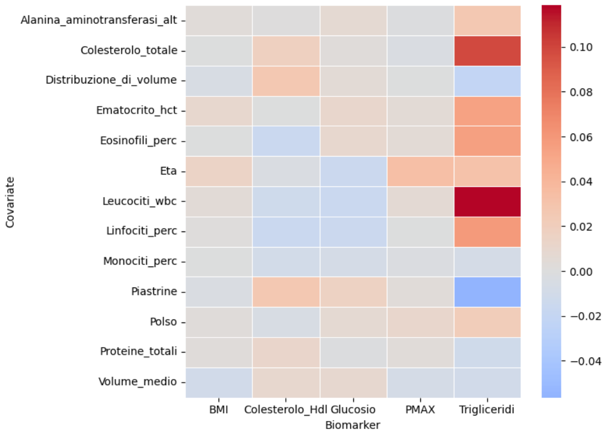


Figure 3: Heatmap of posterior median fixed effects (β). Colors represent the magnitude and direction of the impact of each covariate on the five target variables.

Most influential fixed effects

Inspection of the posterior median fixed effects, computed by aggregating results across the four chains, highlights a small number of covariate–outcome associations with comparatively large magnitude. These represent the most influential population-level effects in the model.

Triglycerides appear to be the biomarker most strongly influenced by covariates. In particular, `Leucociti_wbc`, `Colesterolo_totale`, `Linfociti_perc`, and `Ematocrito_hct` exhibit the largest positive posterior medians for `Trigliceridi`. This suggests that systemic inflammatory and hematological markers are positively associated with triglyceride levels at the population level.

Systolic blood pressure (`PMAX`) shows its strongest association with age (`Eta`), indicating a clear positive relationship between aging and blood pressure. Pulse (`Polso`) also exhibits a positive but more moderate association with `PMAX`. HDL cholesterol (`Colesterolo_Hdl`) is most strongly associated with total cholesterol (`Colesterolo_totale`) and `Distribuzione_di_volume`, reflecting shared lipid-related structure in the multivariate model. For body mass index (`BMI`), age and hematocrit display the most noticeable positive effects, although effect sizes are generally smaller compared to those observed for triglycerides.

Negative associations are comparatively modest in magnitude across outcomes. The most visible include a negative effect of `Piastrine` on `Trigliceridi` and a mild negative association between `Volume_medio` and `BMI`. However, the majority of effects are small in absolute value, reflecting substantial shrinkage induced by the hierarchical Bayesian structure. Taken together, the fixed-effect structure suggests that triglycerides and blood pressure are the outcomes most strongly influenced by the observed covariates, whereas other biomarkers exhibit weaker and more diffuse associations.

Regarding `Eta`, it emerges as the most influential and consistent predictor across all analyzed targets. Its high magnitude and narrow credible intervals confirm its role as a primary driver of metabolic decline, shifting all biomarkers toward the metabolic syndrome profile.

Furthermore, a finding of particular clinical relevance emerges from the analysis of the `Eosinofili_perc` variable, associated with adipose tissue inflammation. Unlike other covariates that exhibit heterogeneous or even protective effects across different parameters, eosinophils demonstrate a systematic consistency in promoting the metabolic syndrome phenotype: they are simultaneously associated with increased triglyceride and glucose levels and a reduction in HDL cholesterol. This evidence suggests that eosinophil counts may serve as an early driver of metabolic inflammation, simultaneously impacting multiple axes of metabolic risk.

9.2 Patient-level variability and Dirichlet Process structure

Patient-specific random effects capture residual heterogeneity not explained by observed covariates. These effects represent systematic deviations from the population mean trajectory and account for the repeated-measure structure of the data. While variability across patients is evident, most posterior credible intervals for individual random effects include zero, indicating moderate subject-level deviations after conditioning on covariates.

To flexibly model the distribution of random effects, we adopt a Dirichlet Process (DP) mixture prior, allowing for potential multimodality without pre-specifying the number of latent components. Posterior examination of cluster assignments indicates that effectively a single mixture component is occupied. All patients are assigned to the same latent component with high posterior certainty, and no evidence of distinct latent subgroups emerges.

The absence of multiple occupied components suggests that patient-level heterogeneity is continuous rather than multimodal. In this setting, the Dirichlet Process does not identify distinct patient phenotypes; instead, it behaves as a robust, flexible prior that reverts to a single-component random-effects model when the data do not support additional complexity.

9.3 Predictive performance for Metabolic Syndrome

Using posterior predictive simulations, we derive visit-level probabilities of Metabolic Syndrome (MetS), defined as the presence of at least three abnormal biomarker criteria. On the held-out test set, the model demonstrates moderate discriminative ability, with an ROC AUC of approximately 0.72. This indicates that visits with true MetS tend to receive higher predicted probabilities than non-affected visits, although the separation between the two groups is less pronounced than in the training data.

Calibration analysis reveals that predicted probabilities are not perfectly aligned with observed frequencies. In particular, the calibration curve lies systematically below the diagonal reference line, indicating that predicted risks are generally higher than the empirical MetS frequencies in the corresponding probability bins. This suggests a degree of overestimation of risk on the test set, which may reflect distributional differences between training and test patients as well as the additional uncertainty introduced when sampling random effects for previously unseen individuals.

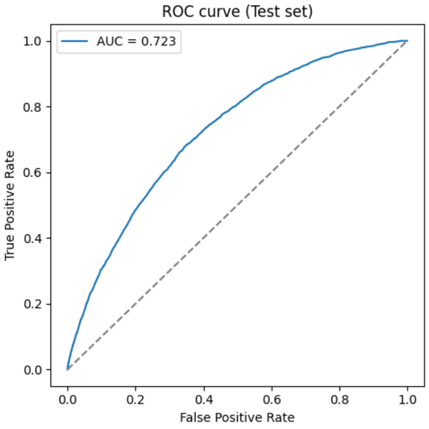


Figure 4: ROC curve

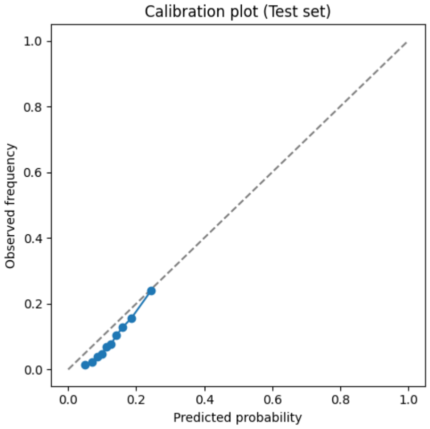


Figure 5: Calibration plot

In conclusion, the presented methodology integrates population-level covariate effects, patient-specific heterogeneity, and multivariate outcome dependence within a coherent Bayesian framework. Fixed effects capture systematic associations between routinely collected biomarkers and metabolic outcomes, while the Dirichlet Process prior flexibly models residual patient-level variability. Although no meaningful latent patient subgroups are supported by the data, the DP component enhances robustness without inducing artificial clustering.

Predictive performance on the independent test set confirms that the model retains meaningful discriminatory power, though with reduced accuracy and imperfect calibration compared to the training data. This behavior is consistent with realistic generalization performance and highlights both the strengths and the limitations of using routinely collected donor information for early MetS risk screening.

10 Covariate SESSO

As can be clearly observed (*see section 7, Model 3*) from the analysis of the credible intervals for the fixed effects of the variable `SESSO` across all targets, this covariate does not appear to be meaningful for the purpose of predicting the five target components, at least when using the Hard Shrinkage technique for covariate selection.

Despite this result, it is reasonable to assume that being male or female still plays a relevant role in predicting the values of the target variables; therefore, we propose a plausible explanation for our findings.

First, it is important to consider that we do not work on the full dataset originally provided for the analysis, due to computational limitations in running our models on such a large dataset (more than 100,000 observations). Indeed, when analyzing the scatter plot distributions of all pairs of target variables as well as the density plots in both the complete dataset and the reduced one, gender-related patterns are clearly visible in the full dataset, while they appear much weaker in the reduced sample. Consequently, different results might emerge if the models are trained on the complete dataset rather than on our reduced subset, which includes only 28 males and 7 females. The reduced variability and the strong imbalance between the two groups may limit the model's ability to detect gender-specific effects. This can be clearly seen from the densities of the covariate `Ematrocrito_hct` and the target `PMAX` shown above.

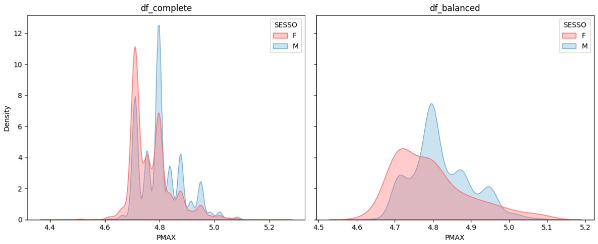


Figure 6: PMAX density plots colored by gender

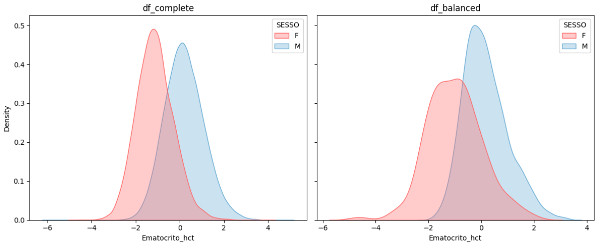


Figure 7: Ematocrito density plots colored by gender

Second, and perhaps more importantly, we hypothesize that some of the covariates already carry intrinsic information related to the patient's gender. In fact, females typically exhibit lower values of the target variables, but also consistently lower values in several of the covariates included in the model. This suggests that gender-related physiological differences are already indirectly encoded in these variables. As a consequence, the explicit inclusion of gender as an additional intercept may be redundant, since its effect is already implicitly captured through the relationships between the other covariates and the targets. As you can see in the plots below, for a lot of covariates and targets, the subpopulation of male donors exhibit higher values than the one of female donors.

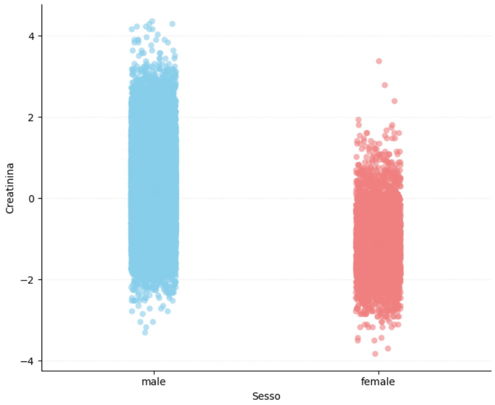


Figure 8: Strip plot of Creatinina

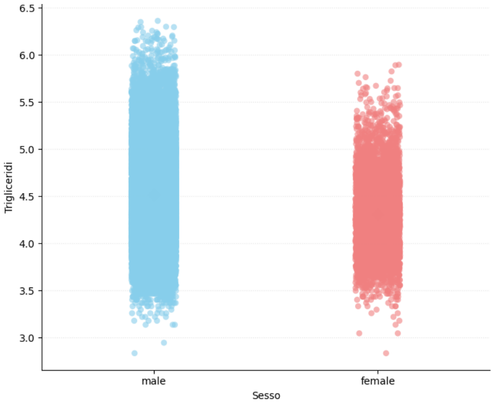


Figure 9: Strip plot of Trigliceridi

This interpretation is further supported by the clinical framework used to diagnose MetS, which relies on two distinct diagnostic reference tables for males and females. These tables adopt different threshold values for several clinical indicators, reflecting well-established biological differences between genders.

Our hypothesis, therefore, is that the model is already able to infer whether a subject is male or female from the pattern of the other covariates, without requiring this information to be explicitly included as an additional fixed intercept.

11 Conclusions and further developments

To translate our predictive model into a practical clinical tool, we must evaluate it beyond standard statistical metrics. Connecting the algorithm's results to real-world practice means rethinking how we set decision boundaries. Consequently, our final assessment shifts focus from conventional classification assumptions toward a threshold strategy that reflects the actual prevalence of the disease, ensuring the model functions effectively as a proactive screening instrument.

11.1 The *Prevention* Threshold

In standard binary classification tasks, a default decision threshold of $\tau = 0.5$ is typically employed to differentiate between classes. This assumes a relatively balanced dataset where the prior probabilities of the classes are comparable. However, our scenario presents a significant deviation from this norm: the prevalence of the syndrome in the training set is extremely low ($\sim 14\%$).

In such a highly imbalanced context, applying a standard 0.5 threshold would result in a conservative model that almost never predicts the positive class, effectively ignoring the rare but critical at-risk cases. Therefore, to align the model's behavior with the actual distribution of the disease, we adopted a threshold much closer to the population prevalence.

11.1.1 Optimizing for Recall

To transform these statistical outputs into a proactive clinical tool, we calibrated the model to function as a high-sensitivity screening instrument. We recommend a threshold in the 0.12–0.15 range. This decision is driven by the metric of Recall, which is paramount in medical screening.

Recall (or Sensitivity) is defined as the fraction of relevant instances that have been retrieved over the total amount of relevant instances:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

- TP (True Positives): At-risk patients correctly identified.
- FN (False Negatives): At-risk patients missed by the model.

By lowering the threshold to approximately 0.12, the Recall rises to $\sim 90\%$. This ensures that 9 out of 10 at-risk donors are identified. In our specific clinical setting, minimizing False Negatives is crucial: failing to identify a patient developing the syndrome implies missing the window for early intervention.

11.1.2 The Trade-off: Precision vs. Clinical Utility

Naturally, lowering the threshold increases the number of False Positives (FP), thereby reducing Precision ($\sim 35\%$). However, this trade-off is clinically justifiable based on the *cost* of an error:

- Cost of FN (Missed Case): High. The patient develops the syndrome without warning.
- Cost of FP (False Alarm): Low. The consequence is merely low-risk preventive counseling.

The Calibration Plot further validates this approach, confirming that in the 0.12–0.15 range, the model is perfectly calibrated. When the model predicts a 15% risk, the donor's actual empirical risk is indeed 15%, making the probability score a trustworthy metric for the clinician.

11.2 Final Interpretation: An Early-Warning Screening Tool

This threshold strategy reflects the core objective of our work: the model is not designed to be a definitive diagnostic tool for an existing condition, but rather an early-warning system for potential risk.

Unlike a binary diagnosis (Sick/Healthy), this tool places a *flag* on donors who, despite having values within standard reference limits, exhibit a multivariate profile that is suspicious. Specifically, it identifies individuals whose risk score is significantly higher than the average of the healthy donor pool used in the initial dataset.

This *warning* allows the physician to:

- Investigate further: Suggesting additional specific analyses that wouldn't normally be prescribed.
- Intervene proactively: Recommending dietary adjustments or lifestyle changes to prevent the onset of the syndrome.

By acting on these probability signals before the syndrome manifests, we shift the clinical approach from reactive treatment to active prevention, leveraging the model's ability to detect subtle deviations from the healthy population baseline.

11.3 Limitations and Future Developments

Given the single-semester timeline of this project, several avenues remain open for investigation. A primary goal for future iterations is scaling the model to a larger dataset. Supported by increased computational resources, this expansion would enable deeper hyperparameter tuning and the identification of more clinical clusters, improving the model's statistical robustness and generalizability across diverse populations.

12 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). *PriorElicitation_2025.pdf.* [Bayesian Statistics slides]. WeBeep, Politecnico di Milano.

Guglielmi, A. (2025). *MHGibbs_2025.pdf.* [Bayesian Statistics slides]. WeBeep, Politecnico di Milano.

Guglielmi, A. (2025). *LMM_GLMM_2025.pdf.* [Bayesian Statistics slides]. WeBeep, Politecnico di Milano.

Guglielmi, A. (2025). *CovariateSelection_2025.pdf.* [Bayesian Statistics slides]. WeBeep, Politecnico di Milano.

Guglielmi, A. (2025). *Clustering_BNP.pdf.* [Bayesian Statistics slides]. WeBeep, Politecnico di Milano.

13 Code Navigation Guide

The complete source code, and scripts for this project are available on GitHub at the following repository: `https://github.com/elisanordera/Bayesian-Statistics-Project`.

Below is a guide to the main directories and files:

13.1 Repository Structure

Folder

- Models: Contains all the Python notebooks used to run the Stan codes (`.stan`) of the models.

Preprocessing & Imputation Notebooks

- `mice.R`: Implementation of Multiple Imputation by Chained Equations (MICE) for handling missing values.
- `EDA.ipynb`: Exploratory Data Analysis to visualize distributions and initial correlations.
- `Datasets_balanced_and_reduced.ipynb` Script used to obtain *df_balanced* and *df_reduced*.

Model Selection & Inference Notebooks

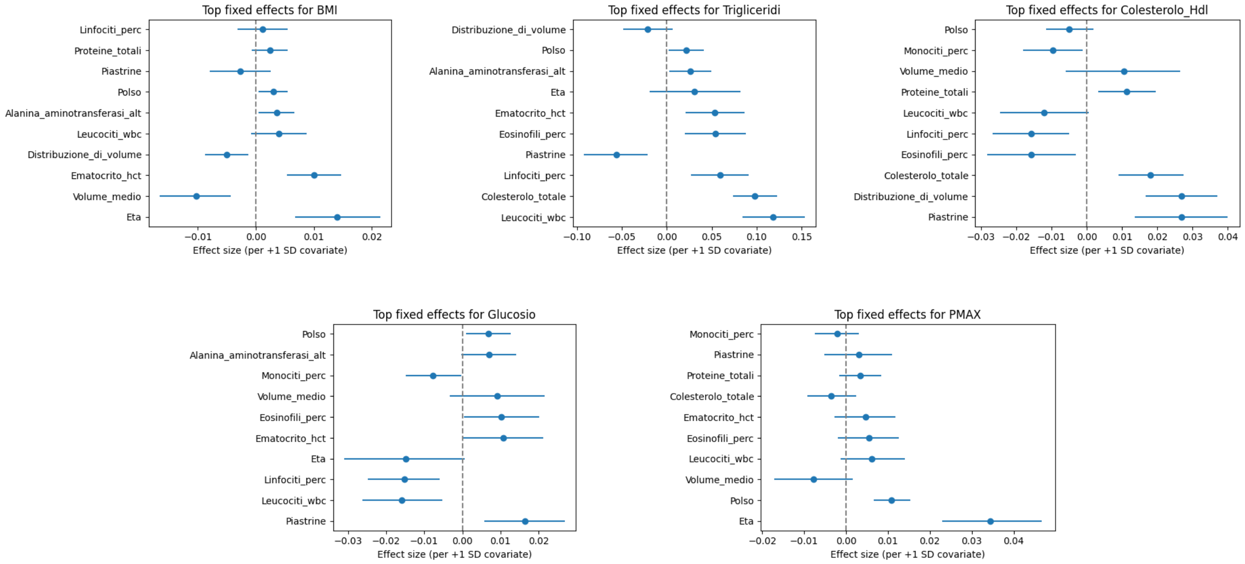
- `Covariates_selection.ipynb`: Analysis of 23 initial covariates to exclude non-significant predictors and define `df_reduced`.
- `Model_Selection.ipynb`: Comparison of different model specifications using Expected Log Predictive Density (`ELPD-LOO`).

Results & Interpretation Notebooks

- `Model_Interpretation.ipynb`: Final analysis, posterior predictive checks, and interpretation of the obtained results.

APPENDIX

A Additional Forest Plots



Forest plots of fixed effects by covariate where in each panel there is posterior estimates and uncertainty for the corresponding regression coefficients.

B Distributions of Covariates and Target Variables by Gender

The following density plots illustrate the distribution of the variables divided by gender.

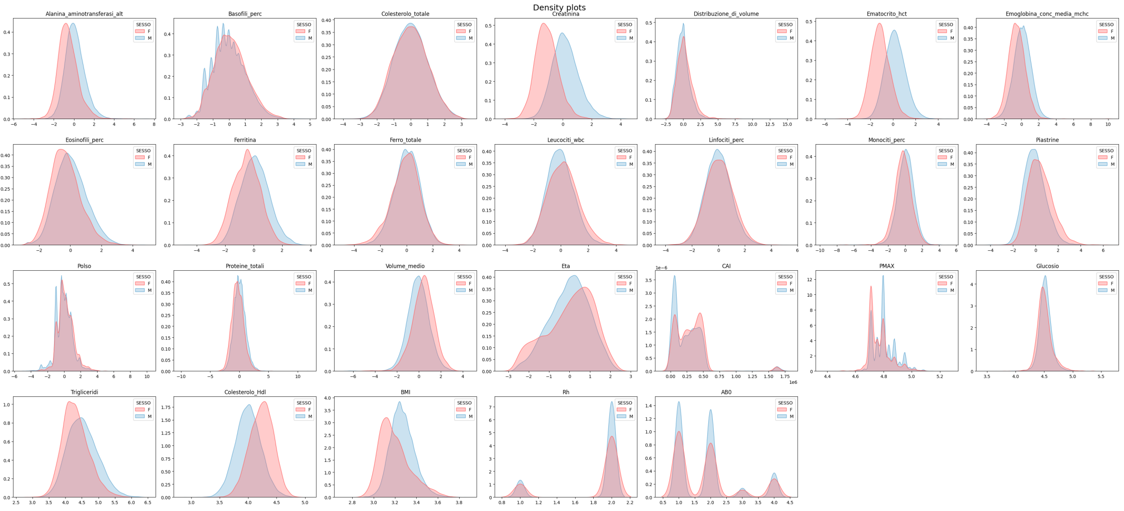


Figure 10: Density plots of the original complete dataset, stratified by gender.

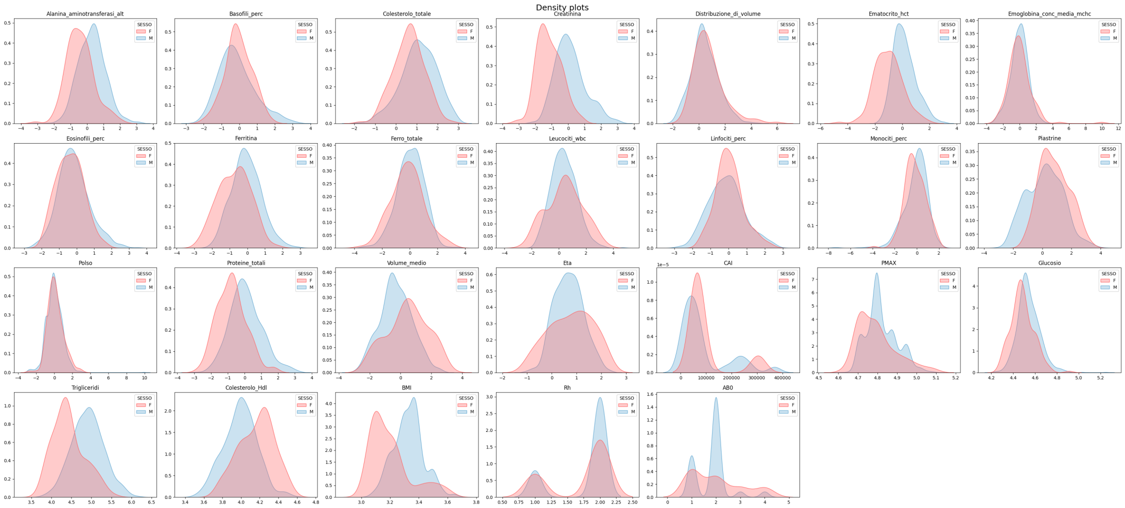


Figure 11: Density plots of the balanced dataset (`df_balanced`), stratified by gender.

C Additional Prior tested for Fixed Effect

C.1 Regularized Horseshoe (Model 1)

Residual covariance:

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(10,\ \mathbf{I}_K)$$

Random effects:

$$\mathbf{b}_i \mid \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b \overset{\text{iid}}{\sim} \mathcal{N}_K(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$$
$$\boldsymbol{\mu}_b \sim \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$
$$\boldsymbol{\Sigma}_b \sim \text{Inv-Wishart}(10,\ 0.5\mathbf{I}_K)$$

Fixed effects (Regularized Horseshoe):

$$\tilde{\beta}_{pk} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad \lambda_p \overset{\text{iid}}{\sim} \text{Cauchy}^+(0,1), \quad \tau \sim \text{Cauchy}^+(0, 0.002)$$
$$\tilde{c}^2 \sim \text{Inv-Gamma}(2,8), \quad \beta_{pk} = \tau\tilde{\lambda}_p\tilde{\beta}_{pk}$$
$$\tilde{\lambda}_p = \sqrt{\frac{\tau^2\tilde{c}^2\lambda_p^2}{\tau^2\tilde{c}^2 + \tau^2\lambda_p^2}}$$

C.2 Regularized Horseshoe LKJ (Model 2)

Residual covariance (LKJ):

$$\boldsymbol{\Sigma} = \mathbf{L}_\Sigma\mathbf{L}_\Sigma^\top, \quad \mathbf{L}_\Sigma = \text{diag}(\boldsymbol{\tau})\mathbf{L}_\Omega$$
$$\tau_k \overset{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5), \quad \mathbf{L}_\Omega \sim \text{LKJ-Cholesky}(4)$$

Random effects:

$$\mathbf{b}_i = \mathbf{L}_{\Sigma_b}\mathbf{z}_{b_i}, \quad \mathbf{z}_{b_i} \overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$
$$\mathbf{L}_{\Sigma_b} = \text{diag}(\boldsymbol{\tau}_b)\mathbf{L}_{\Omega_b}, \quad \tau_{b_k} \overset{\text{iid}}{\sim} \mathcal{N}^+(0, 0.5), \quad \mathbf{L}_{\Omega_b} \sim \text{LKJ-Cholesky}(4)$$

Fixed effects: Prior specification identical to Model 1.

C.3 Bayesian Ridge LKJ (Model 4)

Residual and Random effects covariance: Same LKJ structure as Model 2.

Fixed effects (Bayesian Ridge):

$$\beta_{pk} \mid \sigma_\beta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta)$$
$$\sigma_\beta \sim \text{Student-}t^{+}(3, 0, 2.5)$$

D Alternative Random Effects Specifications

This section details specifications for the random effects structure tested during the model selection process.

D.1 Normal-Normal LKJ Model (Model 5)

In this specification, we model the random effects $\mathbf{b}$ using a multivariate normal distribution with a non-centered parametrization to improve Hamiltonian Monte Carlo sampling efficiency. The model is defined as follows:

$$\mathbf{b}_i = \text{diag}(\boldsymbol{\tau}_b) \cdot \mathbf{L}_{\Omega_b} \cdot \mathbf{z}_{b_i}$$
$$\mathbf{z}_{b_i} \overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$
$$\tau_{b_k} \overset{\text{iid}}{\sim} \mathcal{N}^{+}(0, 0.5), \quad \mathbf{L}_{\Omega_b} \sim \text{LKJ-Cholesky}(4)$$

where I is the total number of donors (35), K is the target number (5), $\boldsymbol{\tau}_b$ represents the vector of scale parameters (standard deviations), and $\mathbf{L}_{\Omega_b}$ is the Cholesky factor of the correlation matrix $\boldsymbol{\Omega}_b$.

D.2 Normal-t-Student LKJ Model (Model 6)

As an alternative to the Gaussian specification, here the standard deviations of the random effects follow a truncated Student's t-distribution, allowing for heavier tails in the scale parameters. The LKJ parametrization is maintained:

$$\mathbf{b}_i = \text{diag}(\boldsymbol{\tau}_b) \cdot \mathbf{L}_{\Omega_b} \cdot \mathbf{z}_{b_i}$$
$$\mathbf{z}_{b_i} \overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$
$$\tau_{b_k} \overset{\text{iid}}{\sim} t^{+}(3, 0, 0.1), \quad \mathbf{L}_{\Omega_b} \sim \text{LKJ-Cholesky}(2)$$

The choice of 3 degrees of freedom for the t-Student prior on the scale parameters τ_{b_k} provides a weakly informative prior that is less sensitive to extreme observations compared to the Half-Normal distribution.