Finite mixture models for trajectory analysis of in-hospital routine laboratory values: application as biomarkers in spinal cord injury patients

Abel Torres Espín

Master en Bioinformática y Bioestadística

Abel Torres Espin

Area 2, aula 1

Daniel Fernández Martínez

Carles Ventura Royo

Lunes 30 de Junio del 2022

Abel Torres Espin

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Finite mixture models for trajectory analysis of in-hospital routinely laboratory values: application as biomarkers in spinal cord injury patients*
Nombre del autor:	*Abel Torres Espin*
Nombre del consultor/a:	*Daniel Fernandez Martinez*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	05/2022
Titulación:	*Máster universitario en Bioinformática y bioestadística UOC-UB*
Área del Trabajo Final:	*M0.167 Trabajo final de máster - Aula 1*
Idioma del trabajo:	English
Número de créditos:	15
Palabras clave	*Finite mixture models, trajectory analysis, spinal cord injury*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Finalidad y contexto. El diagnóstico y pronóstico temprano después de lesiones traumáticas de la médula espinal (SCI por sus siglas en inglés) presenta un gran reto debido a la complejidad patológica y la heterogeneidad de pacientes. Datos obtenidos durante la práctica médica habitual como las analíticas de laboratorio, pueden proveer información sobre los procesos patofisiológicos, y tienen potencial para ser usados como biomarcadores. Hipotetizamos que la evolución temporal de marcadores en sangre después de SCI se puede modelar, y que el resultado de estos modelos se puede usar como predictores de características de los pacientes.

Métodos. Hemos modelado los 20 marcadores en sangre más comunes en un cohorte de SCI y traumatismo espinal usando modelos finitos mixtos para determinar distintas trayectorias temporales. La probabilidad de pertenecer a un grupo de trayectorias se usó como predictores en modelos de *machine learning* para la predicción de características de los pacientes.

Resultados. Existen distintos grupos de trayectorias no lineales para la mayoría de los marcadores estudiados, y estas trayectorias están asociadas con distintas características de los pacientes. Además, la probabilidad de pertenecer a distintas trayectorias es predictivo de si el paciente va a morir en el hospital, si el paciente presenta una lesión medular, y de la severidad de la lesión.

Conclusiones. Datos extraídos de la práctica clínica rutinaria se pueden utilizar para modelar las trayectorias dinámicas de los marcadores sanguíneos después del SCI. Nuestro trabajo sugiere que los cambios temporales de marcadores en sangre pueden ser utilizados como biomarcadores para SCI.

Abstract (in English, 250 words or less):

Background: Early diagnostic and prognostication after acute traumatic spinal cord injury (SCI) is challenging due to pathology complexities and population heterogeneity. Routinely collected data during standard medical practice, such as laboratory analytes, can serve as surrogates of underlying pathophysiological processes and therefore be used as a biomarker. We hypothesized that distinct temporal trends of blood analytes can be modeled after SCI and that those would be predictive of patient characteristics.

Methods: Using real-world data from available electronic health records, we assembled a big-data asset and modeled distinct laboratory analytes measured over time during the early hospitalization after acute spine trauma with or without SCI. We fitted longitudinal finite mixture models (FMM) to determine distinct group trajectories over time on 20 blood analytes commonly measured in these populations. The probability of group trajectory membership was used in machine learning models to predict patient characteristics.

Results: We show non-linear heterogeneous temporal trends of blood analytes after spine trauma and SCI. These trajectories are associated with different patient characteristics. In dynamic prediction experiments, the probability of belonging to a specific analyte trajectory is predictive of whether the patient would die in hospital, the patient presented with an SCI, and SCI severity.

Conclusions: Routinely real-world data can be used to model blood analytes' dynamic changes after SCI with prediction validity for patient characteristics. Our work suggests that temporal blood trends are promising early predictors of SCI pathology. This work sets the bases for further developing dynamic biomarkers in neurotrauma and other neurological conditions.

Table of Contents

List of Figures

List of Tables

1 Summary

Background: Early diagnostic and prognostication after acute traumatic spinal cord injury (SCI) is challenging due to pathology presentation complexities and population heterogeneity. Identifying objective predictors of injury severity and neurological recovery is essential for efficient patient management. Routinely collected data during standard medical practice, such as laboratory analytes, can serve as surrogates of underlying pathophysiological processes and therefore be used as biomarker signatures for diagnostic and prognostication. However, these multivariate, dynamic, and heterogenous markers and their intricate relationship to SCI clinical pathology require advanced analytical approaches. We hypothesized that distinct temporal trends of blood analytes can be modeled after SCI and that those would be predictive of patient characteristics. We use longitudinal finite mixture models to study the multivariate temporal dynamics of a heterogeneous SCI population to investigate them as biomarkers for diagnostic and prognostication severity.

Methods: Using real-world data from available electronic health records, we assembled a big-data asset and modeled distinct laboratory analytes measured over time during the early hospitalization after acute spine trauma with or without SCI. Specifically, the MIMIC III and IV datasets were used to fit longitudinal finite mixture models (FMM) to determine distinct group trajectories over time on 20 blood analytes commonly measured in these populations. The resulting probability of group trajectory membership was then used in machine learning models to predict patient outcomes in an internal cohort as well as an external cohort from the TRACK-SCI study.

Results: The obtained FMM models with more than one class trajectory illustrate heterogeneous temporal trends of blood analytes after spine trauma and SCI. These trends are non-linear, and for most analytes, the trajectory is better modeled by non-Gaussian approximations of the error deviance. These trajectories present distinct temporal evolutions and are associated with different patient characteristics. In dynamic prediction experiments, the probability of belonging to a specific analyte trajectory is predictive of whether the patient would die in the hospital, present with an SCI, and SCI severity.

Conclusions: Daily real-world data can be used to model blood analytes dynamic changes after SCI with prediction validity for patient outcomes. Our work suggests that temporal blood trends are promising early predictors of SCI pathology. Here we presented a proof-of-concept to test our hypothesis.

Contributions of this work: This work expands on the current knowledge of blood changes early after SCI suggesting methodologies that better capture the non-linear heterogeneous dynamics of SCI pathophysiology. To our knowledge, this is the first time that finite mixture models have been used to that end. This work also establishes the utility of considering the prediction of time trends as potential biomarkers for SCI. We offer the code and the trained models to facilitate future development. We hope that his work sets the bases for further developing dynamic biomarkers in neurotrauma and other neurological conditions.

2 Introduction

2.1 Context and Justification

This work discusses the analysis of longitudinal biomedical data, specifically in determining homogeneous groups of patients with similar temporal patterns, also referred to as trajectories. Biological processes, including pathological ones, are dynamic, meaning that they change over time. Therefore, pathobiological developments underlying medical conditions are also dynamic. Single discrete measures (at a given time point) are often used as proxies of the current stage of these processes, limiting the understanding of previous and future states. Longitudinal studies collecting the same measures over time are more informative than discrete ones because they can describe the temporal evolution of a given process. However, an issue with longitudinal studies is that they increase the complexity of their analysis due to the higher number of measures and interdependencies. In the last decades, a growing body of research has been dedicated to the statistical treatment of longitudinal data. A specific way to work with longitudinal measures is to determine different temporal trends or trajectories (Nagin, 2014; Ram & Grimm, 2009; van der Nest et al., 2020) to categorize patterns by a single categorical variable. A different trajectory pattern then describes a group of homogeneous patients with similar temporal progression in a medical context. Therefore, finding those groups of trajectories is the problem of clustering (i.e., finding entities with high intra-group similarities and inter-group dissimilarity) and dimensionality reduction as time variables are reduced to single categorical labels or a small set of parameters defining the time trend. This work analyzes temporal laboratory data from a specific medical context, spinal cord injury (SCI), using different group-trajectory analytical frameworks to determine their potential validity as biomarkers. The following paragraphs introduce SCI, the issue of objective diagnosis of SCI damage and its prognostic, and the current use of routinely blood markers for SCI prediction. Chapter 3 summarizes the state-of-the-art of class trajectory analysis concerning this work.

2.1.1 Brief introduction to spinal cord injury

Traumatic spinal cord injury (SCI) causes permanent autonomic (e.g., blood pressure dysregulation), sensory (e.g., pain), and motor (e.g., paralysis) dysfunction because of the direct damage to spinal cord tissue. It was estimated that there were around 930.000 new SCI cases in the world in 2016, with a calculated prevalence of 27 million people worldwide (James et al., 2019). Although the number of patients is relatively low compared to other medical conditions, SCI has a tremendous personal impact on those suffering from it and their families; SCI has an average estimated lifetime cost of millions of dollars per patient (Merritt et al., 2019). There are no repair treatments for SCI, but an improvement in acute medical management has increased patient survival in the last few decades (Kumar et al., 2018). In addition, the treatment of chronic symptoms has improved patients' lives. Nevertheless, neurological recovery after injury is limited partly because of the difficulty of accurately determining and assessing acute injury characteristics (diagnostic) and predicting patient progression (prognostication) (Albayar et al., 2019; Jogia, Kopp, et al., 2021).

SCI is a complex medical condition (Fig. 1). The initial damage to the spinal cord kills neurons, glia, and vascular cells (Ahuja et al., 2017; Alizadeh et al., 2019). It triggers many secondary pathophysiological processes like inflammation, excitotoxicity, necrosis, and apoptosis, expanding tissue damage for days to weeks and months after injury (Ahuja et al.,

2017). In addition, factors such as injury location (where the cord is damaged), initial severity (how much tissue is destroyed), type of injury (e.g., blunt, penetrating), demographics, medical history, comorbidities, among others, greatly contribute to differences between patients' pathology (i.e., clinical presentation)(Failli et al., 2012; Fouad et al., 2021; Jogia, Kopp, et al., 2021; Liebscher et al., 2022). These factors create severe population heterogeneity, complicating patient diagnostic, prognostication, and ultimately acute patient management and clinical research. Furthermore, SCIs are dynamic, with symptoms and pathological processes rapidly changing as secondary pathophysiology progresses.

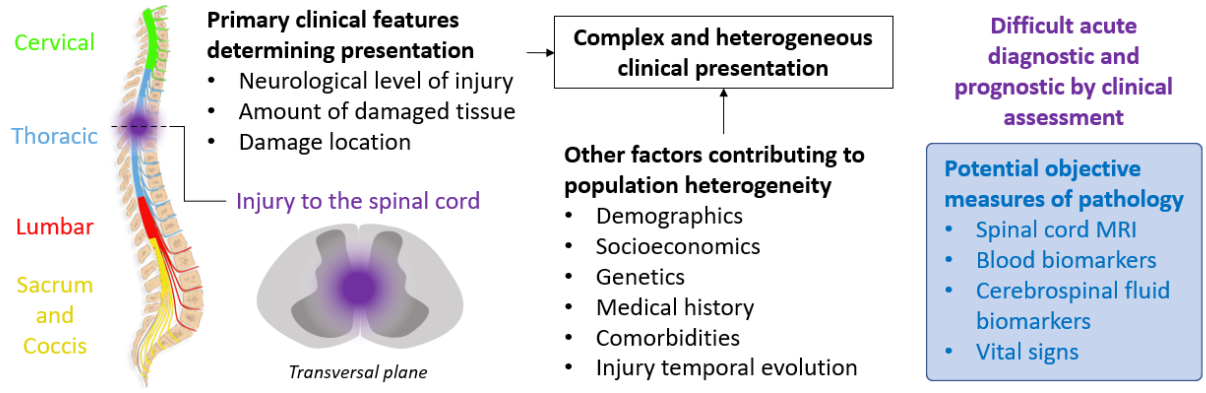


Figure 1. Spinal cord injury complexity. Spinal cord injury clinical presentation is primary determined by the neurological level of injury along the spinal cord, the amount of damaged tissue and the specific tissue that is damaged. Other factors contributing to clinical presentation are demographics and socioeconomics, genetics, past medical history, concurrent comorbidities, and injury temporal processes. All these cause difficulties on determining proper diagnostic and prognostic by the only use of clinical assessments. In the blue box, a list of suggested objective measures for SCI.

2.1.2 The issue of acute SCI diagnostic and prognostic

The complexity of interrelated factors that affect patients' pathology makes early SCI diagnostic and prognostication challenging (Albayar et al., 2019; Jogia, Kopp, et al., 2021). Indeed, individuals with similar injury characteristics can have different recovery trajectories (Khorasanizadeh et al., 2019). This is important because the acute diagnostic determines medical management and predicts patients' recovery. In addition, poor determination of patients' pathology hinders clinical research and trials by introducing unknown heterogeneity. Therefore, reliable diagnostic and prognostication are critical for patients' care.

The most common form to determine the location and severity of the injury, two definitory features of SCI symptomatology, is through neurological assessment (Jogia, Kopp, et al., 2021). However, the results of these neurological assessments early after injury are unreliable, even with standardized tests such as the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI) exam (Betz et al., 2019). The issue is that neurological assessments rely on patients' motor and sensory responses, but those can be difficult or impossible to measure depending on whether the patient is responsive. Several factors can affect patient responses, including patient intoxication (drugs, alcohol), patient consciousness, polytrauma, and the presence of spinal shock; an acute transient depression of neurological function below the level of injury (Jogia, Kopp, et al., 2021; Ruiz et al., 2017). Moreover, it is common to use these early neurological measures to gauge future patient

recovery. Thus, finding objective measures for early patient diagnostics is vital for medical decision-making in current management and translational clinical research.

The use of quantitative methods that do not depend on patients' responsiveness for diagnostic and prognostication are being investigated (Fig. 1). These include MRI neuroimaging (Haefeli et al., 2017; Talbott et al., 2015), time-series physiological measures such as blood pressure (Hawryluk et al., 2015; Squair et al., 2017; Torres-Espín et al., 2021), and fluid biomarkers (Brown et al., 2020; Harrington et al., 2021; Jogia, Lübstorf, et al., 2021; Kwon et al., 2010, 2019; Kyritsis et al., 2021; Leister et al., 2021). Although these are showing promising results, these are often time-consuming methods, require highly specialized analytical approaches, and are not broadly available for their use, which makes them unpractical and not generalizable, at least for the time being (Jogia, Kopp, et al., 2021). In response to these limitations, there is an increasing interest in using in-hospital routinely collected data as part of the regular patients' medical management as biomarkers for SCI, such as blood laboratory values (Brown et al., 2020; Harrington et al., 2021; Leister et al., 2021). These have the advantage that they are broadly available, highly standardized across clinical centers, are easy and fast to collect, and may reflect real-world clinical scenarios more closely than designed data collection studies.

2.1.3 Routinary Blood analytes as biomarkers in SCI

Secondary spinal cord damage triggers pathophysiological cascades of measurable events in the blood (Bourguignon et al., 2021; Kyritsis et al., 2021). This is not unique to SCI since the blood values of different cells and molecules are indicators of global organ function, homeostasis, nutrition, pathophysiological events such as inflammation, and others. Thus, the level of different blood markers at different time points after SCI can be proxies for the underlying pathophysiology. Then, the question is how well these blood markers relate to SCI pathology diagnostic and progression and their utility as biomarkers (Fig. 2). There is a growing interest in the SCI research community to investigate this matter following the success of other neurological conditions such as Alzheimer's (Chen et al., 2017; Dong et al., 2019).

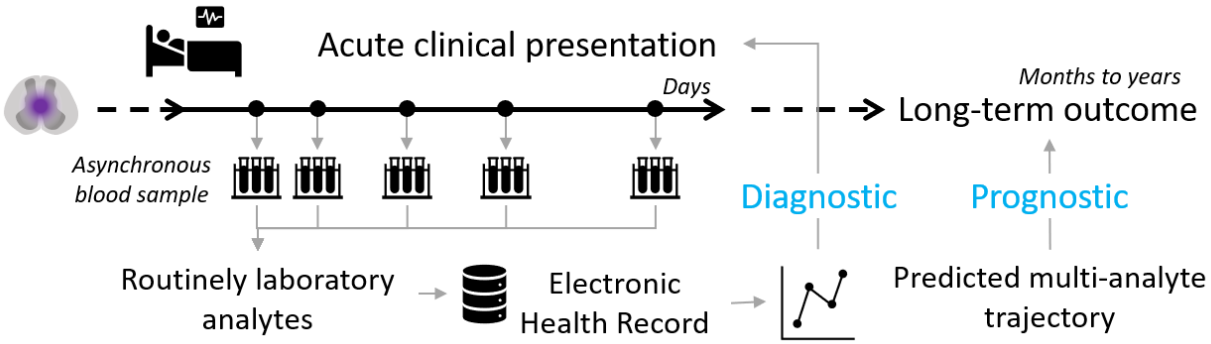


Figure 2. Schematic diagram of the use of routinely laboratory analytes for SCI patient diagnostic and prognostication. Fluid samples such as blood are routinely collected early after injury during hospitalization as part of the clinical management process, resulting in asynchronous (non-constant intervals) measures of several laboratory analytes. The results of these are collected in electronic health records that can be then used for objective diagnostic of acute injury, as well as long-term outcome prognostication.

Furlan and colleagues described hematologic abnormalities during the first week after SCI in isolated cervical injury patients compared to traumatic controls without SCI. They found blood analytes correlating with clinical metrics of injury severity (Furlan et al., 2006). Bourguignon and colleagues associated changes in blood markers up to a year after injury with injury severity, with more complex injuries associated with more abnormal blood values (Bourguignon et al., 2021). Changes in blood analyte levels also relate to developing secondary conditions highly prevalent in SCI patients, such as pressure sores (Gurcay et al., 2009). Further evidence of the relationship between routinely blood analytes and SCI pathology is found in studies aiming to investigate the prediction power of distinct blood metrics for patients' neurological recovery. The levels of distinct blood analytes from the first few days to a month after injury correlate with neurological outcomes at different posterior times up to 12 months, and their use in prognostic models in combination with clinical characteristics improves prediction performance (Brown et al., 2020; Harrington et al., 2021; Leister et al., 2021). In a detailed analysis, Leister and colleagues found that temporal changes in different blood metrics from a few hours to weeks to a year after SCI can predict whether patients will walk one year after injury (Leister et al., 2021). Overall, there are convincing pieces of evidence suggesting that routinely collected blood analytes are candidate biomarkers for SCI diagnostic and prognostication.

Besides the excitement of the latest reports, there is a significant challenge in analyzing blood analytes as biomarkers for SCI: a given analyte's relation with injury severity, location, and neurological recovery is poor (Brown et al., 2020; Harrington et al., 2021). For instance, Brown and colleagues showed that the higher association between a single blood analyte and neurological assessment scores had a Kandal's tau of 0.352. That effect size can be considered moderate at best and, given the variability of blood analytes, probably insufficient to be useful as a biomarker. Indeed, the same authors observed that the initial neurological status is the most significant predictor of neurological function at 3 and 12 months after injury, with an R^2 ranging from 0.5 to 0.766, and that blood analytes increased R^2 at best by 0.043 (Brown et al., 2020). Although significant in their study, that performance improvement may not justify the clinical utility of blood analytes in isolation.

One potential explanation for this limited predictive performance is that blood analytes are not necessarily direct measures of tissue damage but sensors that integrate the effect of different pathophysiological cascades. Kyritsis and colleagues explored this idea using advanced transcriptomic analysis, showing that distinct populations of white blood cells early after SCI integrate complex transcriptomic signals related to injury severity (Kyritsis et al., 2021). This may indicate the need to consider multiple analytes and their relationships to capture complex interconnected processes. Another possibility is that blood analytes change rapidly over time, as well as the initial SCI pathophysiology. Those changes may not be synchronous, making it difficult for a single measurement or several measurements that are too far away apart to capture SCI pathophysiological dynamics. Moreover, single measurements of blood analytes may be more affected by changes induced by other physiological reasons and not related to SCI pathophysiology. Another option explaining the poor predictivity of blood analytes is that studies on this matter may not consider the plethora of potential confounding factors (Jogia, Kopp, et al., 2021) that can generate unobserved population heterogeneities, hampering pinpointing the associations between markers and outcomes. Thus, the noted three features affecting the robustness of associations between blood analytes and outcomes in SCI can be summarized as multidimensionality, temporality, and heterogeneity. Here we use data-driven approaches tailored to address those issues at once.

2.1.4 Multidimensionality, temporality, and heterogeneity of blood analytes in SCI

The previous section described three potential reasons why blood analytes as biomarkers in SCI could be challenged: multidimensionality, temporality, and heterogeneity. Previous works have anecdotally dealt with those problems as part of their data methodology but without much focus on them and in isolation.

In the case of multidimensionality, Brown and colleagues used principal component analysis (PCA) to reduce the number of blood analytes into a small set of composed variables or principal components (PCs) (Brown et al., 2020). Each retained PC then reflects different correlations among all analytes in independent directions, such that the first PC explains the maximal variance in the data, the second the maximal variance left unexplained by the first, and so on (Jolliffe & Cadima, 2016). Then, the PCs can be used as a composing biomarker (Huie et al., 2019). While this strategy can be helpful in some situations, the application of PCA is limited by their assumption that the correlation patterns arrive from a single population, that the biological processes proxied by the PCs are independent, and the difficulty in managing non-independent observations (e.g., temporal data) (Jiang & Eskridge, 2000). While extensions of PCA for dealing with those limitations exist, such as multiple factor analysis (Abdi et al., 2013), their utility in SCI biomarker discovery is still to be studied. Prior reports in SCI have also analyzed longitudinal changes of different analytes in one form or another. Furlan and colleagues compared measures and analytes using one-way analysis of variance (ANOVA) with subsequent pairwise contrast between days (Furlan et al., 2006). This approach ignores the intra-subject correlation due to repeated measures and considers time a discrete measure (see chapter 3.2). Bourguignon and colleagues and Leister and colleagues improved the longitudinal analysis of analytes by using linear mixed models to account for subject time dependencies and model time trends as continuous (Bourguignon et al., 2021; Leister et al., 2021). Some limitations of these two previous works are that the analysis only considered random intercepts and was limited to linear trends, which may not capture the non-linear evolutions of blood analyte changes. In addition, both methods used above (ANOVA and linear mixed model) were conducted in a univariate form, limited to a single analyte at the time, and do not account for unobserved heterogeneity.

A potential analytical approach considering multidimensionality, temporality, and heterogeneity is using a family of statistical technics known as multi-class trajectory modeling. These methods aim at discovering previously unknown homogenous groups of patients with similar temporal profiles of different longitudinal variables (Nagin, 2014; Nagin et al., 2018; Ram & Grimm, 2009; van der Nest et al., 2020).

2.2 Objectives

This work's overarching goal is to analyze in-hospital routinely collected blood analytes early after SCI to discover different groups of trajectories and their use as biomarkers. In that context, the three following aims are defined.

- **Aim 1:** To determine group-trajectory for different laboratory analytes using growth mixture models and their associations to patient characteristics in an SCI population
    - **Aim 1.1:** To construct a dataset from available electronic health record databases containing daily laboratory values.

    - **Aim 1.2:** To summarize all laboratory, clinical, and demographic data through exploratory data analysis. This includes summary statistics and visualizations for decision-making on data preparation.
    - **Aim 1.3:** To model the trajectory for each one of the laboratory analytes and discover potential distinct trajectory groups
    - **Aim 1.4:** To uncover the relationship of each trajectory with patient clinical characteristics
- **Aim 2:** To determine multi-trajectory groups across laboratory analytes using group-based multi-trajectory analysis and their associations to patient characteristics
    - **Aim 2.1:** To model multi-trajectories and potentially discover groups across all laboratory analytes
    - **Aim 2.2:** To uncover the relationship of the multi-trajectory groups with patient clinical characteristics
- **Aim 3:** To build a predictive model of outcomes considering group trajectory in data from an observational study
    - **Aim 3.1:** To classify a set of new patients from an observational study into predicted trajectory groups
    - **Aim 3.2:** To incorporate the predicted trajectory groups in a predictive model of patient neurological outcome

2.3 Summary of the Methods

This section offers a summary of the methods used in this work. Chapter 4 provides a more detailed description of the methodology.

Data: Three databases are used, two for building trajectory models and one for the classification of new patients. For trajectory modeling, we used two different epochs of MIMIC (Medical Information Mart for Intensive Care), an extensive single-center database with EHR of patients admitted to the Beth Israel Deaconess Medical Center in the USA. The two epochs are the MIMIC-III, with 46,520 patients from 2001 to 2012 (Johnson et al., 2016), and the MIMIC-IV, with 382,278 patients from 2008 to 2019. Both databases are accessible through a data use agreement (DUA) through the PhysioNet project (Goldberger et al., 2000). Both databases are de-identified by the data providers, with no risk for patient identification nor the requirement of ethical approval. Both MIMIC databases share a core schema structure and contain daily in-hospital laboratory analytes and data on clinical presentation, diagnostics, procedures, medications, and vitals. A spine trauma and SCI patients cohort is constructed using ICD9/ICD10 diagnostic classification codes. For the classification of new patients, we used data from 137 patients enrolled in the Transforming Research and Clinical Knowledge in Spinal Cord Injury (TRACK-SCI) study (Tsolinas et al., 2020), a longitudinal observational cohort study at the Zuckerberg San Francisco General Hospital and at the University of California San Francisco. TRACK-SCI collects highly granular in-hospital and post-hospitalization data, including laboratory assays and long-term neurological outcomes.

Exploratory analysis: Exploratory data analysis (EDA) is performed to summarize the data. Temporal spaghetti and marginal density plots are generated to understand the amount of available data, the underlying distribution, the potential presence of outliers, and non-linearities. We used this information to curate the data, as well as to make decisions before modeling. Latent Class Growth Analysis (LCGA) was performed as part of the exploratory analysis. LCGA is part of the model-based finite mixture model family, as explained in chapter 3, with a restrictive set of parameters that can be used as the first approximation of trajectory groups' presence (Nagin, 2014; van der Nest et al., 2020).

Trajectory modeling: A growing body of methods and approaches to determine group (class, cluster) trajectories, also known as longitudinal clustering. Here we used model-based trajectory analysis through longitudinal finite mixture models (van der Nest et al., 2020) of the type of growth mixture models (GMM), also known as the latent class mixture model (Proust-Lima et al., 2017; Ram & Grimm, 2009; van der Nest et al., 2020). GMM model the changes of a response variable over time as a latent (unobserved) continuous function in subgroups of subjects, where observations at a given time are realizations of that latent function with noise (see Chapter 3). GMM is well set for the problem since it allows for aperiodic timepoints (observations might be performed at different times across subjects), which is the case for in-hospital laboratory values. In addition, we used group-based multi-trajectory (GBMT) analysis (Nagin et al., 2018) to determine groups of patients that share similar trajectories on all considered analytes at the same time. Models are constructed from the MIMIC-III/IV datasets. Several models are fitted to explore the parameter space, the number of mixture components, non-linear transformations, and link functions. We used the previously described 2-step workflow, and the decision process will systematically search for the best (i.e., most parsimonious) model (van der Nest et al., 2020). Different likelihood-based information criteria (i.e., ICL, BIC) and similar metrics of parsimony will be used to determine the number of trajectories, known as class enumeration. After class enumeration, parameters of non-linearity would be tuned. Once the model is decided, trajectory membership assignment of a subject is done by the trajectory class with a higher posterior probability for that subject. Trajectory membership of TRACK-SCI subjects (unseen by the model) is predicted and assigned for the trajectory class with a higher probability for each subject.

Trajectory characterization: Patients' trajectories were characterized based on clinical and demographic features extracted from the databases: age, gender, ethnicity, cohort group (SCI with vertebral fracture, SCI without vertebral fracture, spine trauma with no SCI), length of hospital stay, whether the patient died in hospital, and the number of ICD diagnostics. Differences between trajectory groups were analyzed using ANOVA or t-test for continuous variables and Fisher exact test for categorical variables. For each one of the analytes, p-values were adjusted for false discovery rate (FDR) by the Benjamini-Hochberg method, and q-value is provided. The level of significance was set at $q < 0.05$.

Predictive outcome modeling: We designed different experiments to determine the analyte trajectories' predictive utility as a biomarker for different outcomes. Since determining the best approach for modeling outcomes is out of the scope of this work, a single Machine Learning model type was used.

Software: All this work was performed in R (R Core Team, 2021). For fitting GMM and LCGA, we used the R package lcmm (Proust-Lima et al., 2017). For GBMT, we used the R package gbmt (Magrini, 2021/2022). Linear models with regularization for prediction are fitted using the glmnet and caret R packages (Friedman et al., 2010; Kuhn, 2021; Microsoft and Hong Ooi, 2019). Several other packages, such as the meta-package tidyverse (including dplyr, ggplot2) are used for data wrangling and EDA.

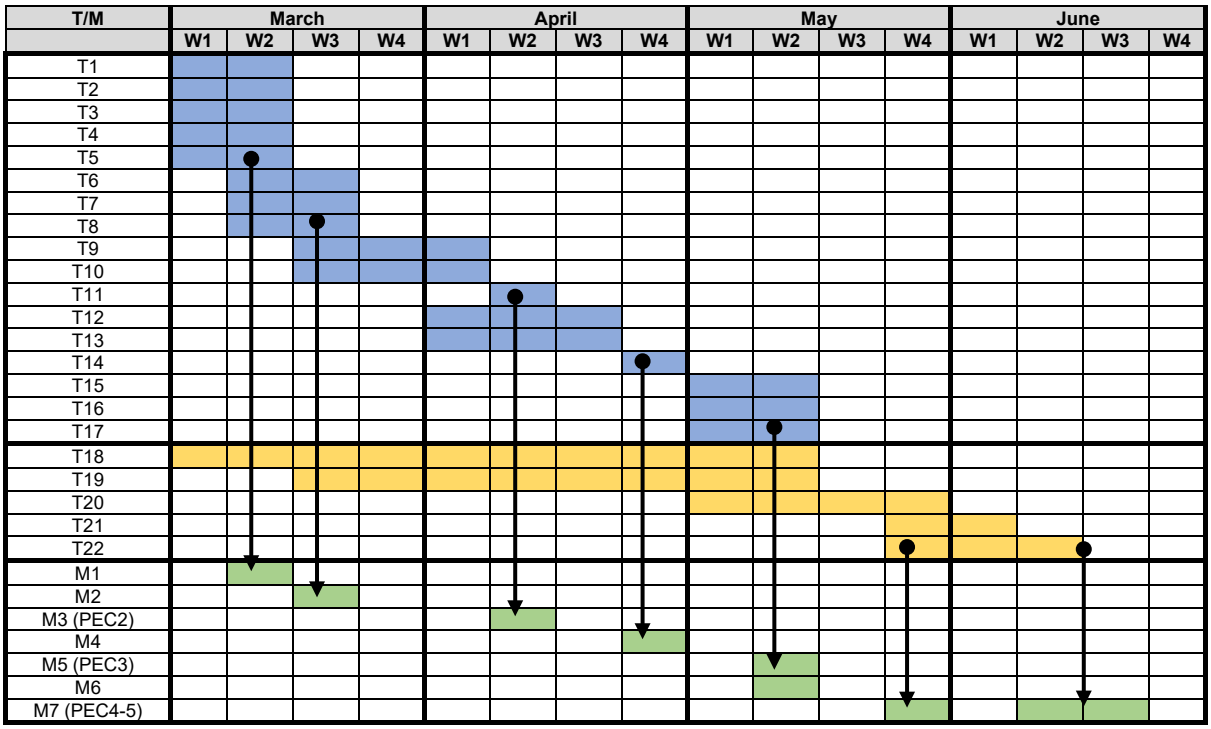
2.4 Planning

This section describes the task, milestones, and calendar for the realization of the TFM.

Tasks and Milestones

Aim 1: To determine group-trajectory for different laboratory analytes using growth mixture models and their associations to patient characteristics.
Aim 1.1: To construct the dataset from available EHR databases
Task 1 (T1): To download the database files for MIMIC-III/MIMIC-IV
T2: To select the group of patients from MIMIC-III/IV using ICD9/10s for spine trauma
T3: To extract, annotate and prepare for analysis the clinical characteristics, demographics, and laboratory data from MIMIC-III/IV
T4: To select the laboratory analytes to use given the available data
T5: To format the TRACK-SCI dataset for analysis
Milestone 1 (M1): To have the data ready for analysis
Aim 1.2: To summarize all data through exploratory data analysis
T6: To summarize the patient characteristics and demographics on a table
T7: To plot each analyte over time, as well as its marginal density, and perform data preparation for analysis
T8: To determine the time window for analysis, and the modeling parameters
M2: To have a good understanding of the data at hand and its modeling requirements
Aim 1.3: To model the trajectory for each one of the laboratory analytes and discover potential distinct trajectory groups
T9: To fit LCMM models for each one of the laboratory analytes, determining the proper number of groups and modeling parameters
T10: To describe the final models based on their group trajectories, posterior probabilities, uncertainties, and assigned membership
Aim 1.4: To uncover the relationship of each trajectory with patient clinical characteristics
T11: For each one of the laboratory analytes, to determine the group descriptive summaries of patient characteristics and demographics and compare between groups
M3 (PEC2): To have trajectory groups and the underlying models
Aim 2: To determine multi-trajectory groups across laboratory analytes using group-based multi-trajectory analysis and their associations to patient characteristics.
Aim 2.1: To model multi-trajectories and potentially discover groups across all laboratory analytes
T12: To fit GBMT models and select the best fitting one
T13: To describe the multi-trajectory group based on their analyte trajectories
Aim 2.2: To uncover the relationship of the multi-trajectory groups with patient clinical characteristics
T14: To determine the multi-trajectory group descriptive summaries of patient characteristics and demographics and compare between groups
M4: To have multi-trajectory groups and the underlying models
Aim 3: To build a predictive model of outcomes considering group trajectory in data from an observational study.
Aim 3.1: To classify a set of new patients from an observational study into predicted trajectory groups
T15: To obtain the predicted trajectory group for TRACK-SCI for the individual LCMM as well as the multi-trajectory GBMT models
Aim 3.2: To incorporate the predicted trajectory groups in a predictive model of patient neurological outcome
T16: To fit an LM/GLM for neurological outcome prediction, including trajectory groups as features
T17: To assess model performance with and without trajectory groups
M5 (PEC3): To describe trajectory groups of unseen data and test their performance in patient prognostication
Dissertation (writing and presentation)
T18: Introduction and methodology
T19: Exploratory analysis
T20: Results and discussion
M6: To have the first draft
T21: To finish the dissertation
T22: Slideshow
M7: Dissertation closure and defense

Task and Milestones schedule (Gantt chart)



2.5 Summary of contributions and outputs

The present work adds to the body of evidence suggesting blood analyte values collected in routine medical practice as biomarkers for SCI. It applies modern temporal trend modeling methods (e.g., finite mixture models) and discovers non-linear heterogeneous groups of trajectories predicting patient outcomes. In addition, it illustrates the utility of using real-world data from the in-hospital daily activity for SCI research. As a result of this work, the code and fitted models are provided to facilitate future development.

2.6 Summary of the rest of this document

The rest of the document is organized as follows: Chapter 3 offers a background on the state-of-the-art longitudinal data and class trajectory analysis. Chapter 4 describes the specific methodology used for this work. Chapter 5 presents the results. Chapter 6 discusses the results and puts them into context. Chapter 7 summarizes the work with a conclusion and future work. Chapter 8 contains a glossary with definitions of the most relevant terms. Chapter 9 lists the references. Chapter 10 has extra annexed material left out from the main chapters, including the R code reproducing this work.

3 State of the art

3.1 Longitudinal data analysis, latent growth curve models, and class trajectories

Data of the type of multiple time point measurements or observations per subject is said to be longitudinal with intra- or within-subject repeated measures over time. Studies collecting longitudinal data can provide information on the evolution or progression of biomedical processes over time (trajectory) at the individual and the population level. These studies introduce dependencies between observed values as intra-subject correlations, which need special treatment during statistical analysis (Fitzmaurice & Ravichandran, 2008; Schober & Vetter, 2018; Van Der Leeden, 1998). Another consequence of longitudinal studies is that variance may change over time (Fitzmaurice & Ravichandran, 2008). These two characteristics break two fundamental assumptions that are the basis for standard analytical technics such as linear regression models: independence of observation and homogeneity of variance. Thus, the analysis of longitudinal data requires special considerations. A traditional specialized form of analysis for longitudinal studies with more than two measurements from the same subjects is through partitioning variance techniques such as repeated measures analysis of variance (RM-ANOVA). These methods perform mean differences for each time point with respect to an across-time grand mean, focusing on the pooling subjects within-time and the differences between-time (Bock, 1979; Gueorguieva & Krystal, 2004). Therefore, RM-ANOVA does not model a time trend per se, but time is considered a categorical variable, with each time point being its levels (Fig. 3). RM-ANOVA works well in well-designed, relatively simple longitudinal experiments; however, they have several limitations. Among others, RM-ANOVA requires that all subjects have measurements at all time points (it cannot handle missing temporal data), does not consider time as a continuous process, and is limited to Gaussian distributed errors (Gueorguieva & Krystal, 2004; Krueger & Tian, 2004; Singh et al., 2013).

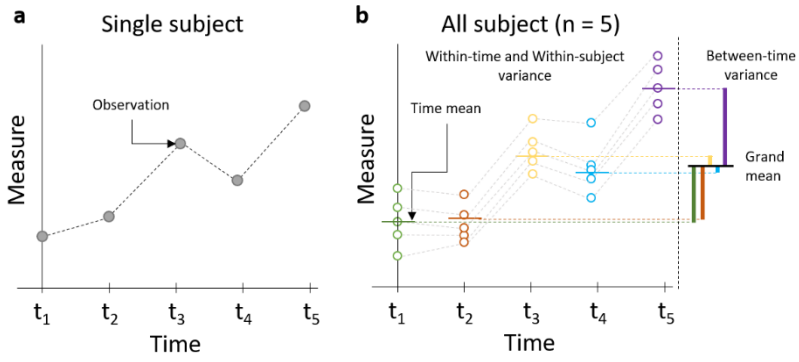


Figure 3. Schematic representation of one-way repeated measures ANOVA. (**a**) Representation of the repeated measures of a single subject. (**b**) Representation of the variance partition for between-time, within-time, and within-subject. The dotted color lines represent the time mean for each timepoint respect to the grand mean.

A more flexible approach to longitudinal data analysis is through model-based techniques that focus on within-subject variation across time, such as latent growth curve models (LGMs) (Van Der Leeden, 1998; van der Nest et al., 2020). Several names are used in the literature for the same concept, such as growth models, latent trajectory models, curve models, or multilevel analysis. Here we use the full name to describe the set of models broadly included in the same analytical framework. Longitudinal LGM is different from the traditional mean difference methods in that they consider temporal evolution as a latent (unobserved) process and the observed measurements as the realization of that process with noise (Fig. 4 and 5). These models account for changes over time represented as time trends (a.k.a latent

trajectories or growth curves)(van der Nest et al., 2020). Two general analytical frameworks are commonly used for the specification of these models: the linear regression model framework (Nagin, 2014; van der Nest et al., 2020), and the Structural Equation Modeling (SEM)/factor analysis framework (Ram & Grimm, 2009; Rovine & McDermott, 2018). This work focuses on using LGM through linear regression models.

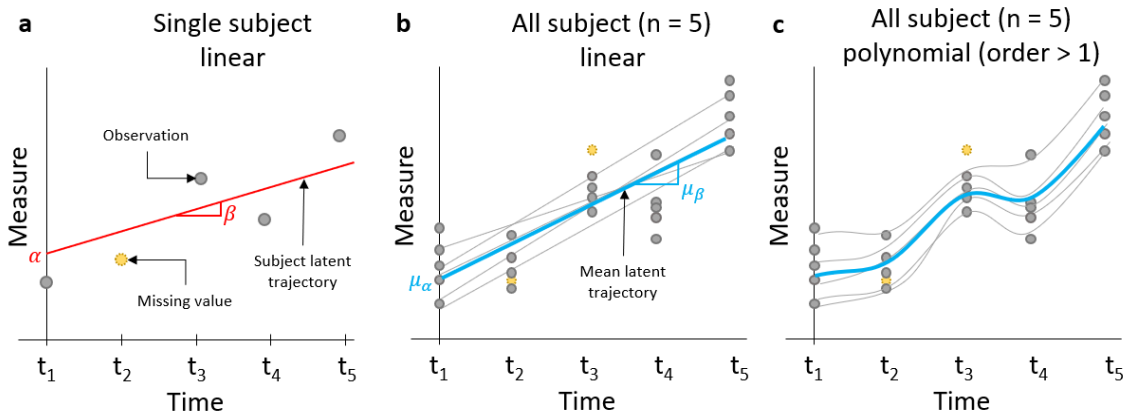


Figure 4. Schematic representation of growth curve modeling through a linear mixed model. (**a**) Repeated measures of a single subject modeled as subject-specific linear trend (red line) parametrized by a subject intercept (α) and a subject slope (β). This approach allows for missing time data and data collected at different time points across subjects. (b) Full sample average trajectory (blue line) parametrized by a population mean intercept (μ_α) and a population mean slope (μ_β). (c) Example of polynomial fit to adjust to non-linear trends.

In addition to modeling the temporal trend as a latent process, extensions of LGM can consider more than one unknown class (latent class) of homogenous groups of subjects that share their temporal trajectories on a heterogenous sample. For a single class, the use of LGM is then concerned with modeling the mean trajectory of a single population through the probability of the observed values (of the entire sample) conditional to time (Table 1 Eq. 1) (Laursen & Hoff, 2006; Van Der Leeden, 1998). Single-class LGM (a.k.a growth curve models; GCM) can be extended to known subgroups of subjects (e.g., gender, drug treatment). The separated growth trajectories per each subgroup can be modeled by including the subgroup and the interaction with time as predictors. When the objective is to identify two or more unknown classes of trajectories, the goal is to find unobserved distinct groups of subjects that share temporal trends and model their class-specific mean trajectory and the subject-specific trajectory. This is the problem of temporal clustering. In LGM, this can be achieved by a mixture of conditional probabilities to time where the number of mixture components is a fixed parameter (specified by the analyst, Table 1 Eq. 2), thus modeling the mean trajectory per class or mixture (Fig. 6)(van der Nest et al., 2020). Furthermore, like other regression models, LGM can be extended to include both time-invariant and other time-variant predictors, generalized models through link functions, and non-linearities through polynomials (of order greater than 1) and splines (Muthen & Asparouhov, 2008; van der Nest et al., 2020). Finally, LGM can also handle missing temporal data (Fig. 4), and the subject's data can be measured at different timepoints as time can be subject-specific.

This work is concerned with the finding of trajectory classes. The following few sections will explain mixture LGM through a family of models called finite mixture models (FMM). Since these models build upon more simple models such as single-class LGM, we provide an overview of their specifications and assumptions. For completeness, we also included model specifications for one-way RM-ANOVA. We are only considering univariate and univariable models for simplicity, with no other within-subject predictors than time. Furthermore, there is

an extensive list of non-parametric approaches to longitudinal clustering that are not discussed here; see (Genolini et al., 2015; Teuling et al., 2022) for reference.

Table 1. The conditional probability of Latent Growth Models (LGM)

General Notation
y: single observed measure of the response variable to model
Y: the vector of observed measure of the response variable to model
i: index of the subject. $i \in \{1, \dots, n\}$ where n is the total number of subjects in the sample.
j: index the occasion (timepoint) of measurement. j $\in \{1, \dots, J\}$ where J is the total number of timepoints per subject. Note that this allows for each subject to be measured at different timepoints.
k: index of mixture component. $k \in \{1, \dots, K\}$
π: mixing or class membership probability for k, where $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$

	Model	Subject probability function
Eq. 1	Single-class LGM	$P(Y_i \mid time_i) = \prod_{j=1}^{J} p(y_{ij} \mid time_{ij})$*
Eq. 2	Multi-class (Mixture) LGM	$P(Y_i \mid time_i) = \sum_{k=1}^{K} \pi_k \cdot \left[\prod_{j=1}^{J} p(y_{ij} \mid time_{ij}; k)\right]$*

*Note that these probabilistic models assume conditional independence of the realization of y_{ij} to the subject-specific random error over time.

3.2 Single-class univariate linear mixed models

3.2.1 One-way RM-ANOVA model

RM-ANOVA can be specified as a particular case of a mixed effect linear model (Table 2, eq. 3). The subject is introduced as a random effect, and time is considered a fixed effect categorical variable with T times levels. RM-ANOVA assumes that subjects are independent draws from a single underlaying population. In addition, the model assumes that random errors (residuals) are normally distributed with a mean of 0 and equal variance. It also assumes that the random effect of the subject is a normally distributed i.i.d variable with mean 0 and equal variance and that both residuals and random subject effects are independent. A complete treatment of RM-ANOVA and its assumptions is out of the scope of this work.

3.2.2 Univariate Growth Curve Models (GCM; single-class LGM)

In the regression framework, GCM is specified as a linear mixed effect model where a linear model is estimated for each subject with time as a continuous predictor (Table 2, eq. 4). These models are also known as multilevel or hierarchical because the model's parameters vary at more than one level; in the case of longitudinal data, at the population level trajectory, and at the individual subject time trajectory (Van Der Leeden, 1998). Thus, random effects (intercept and rate of change coefficients) represent the subject's linear deviation from the average time trend (fixed effects). Random errors represent the measurement error or intra-subject residuals to the subject-specific model. Like RM-ANOVA, the model assumes that subjects are independent draws of a single population. Random effects and errors are assumed to be normally distributed, with a mean of 0, and have their own variances and covariances among them. Depending on the covariance structure specified, GCM can model correlated errors over time (e.g., autoregressive, where an error is correlated with past errors) (van der Nest et al., 2020).

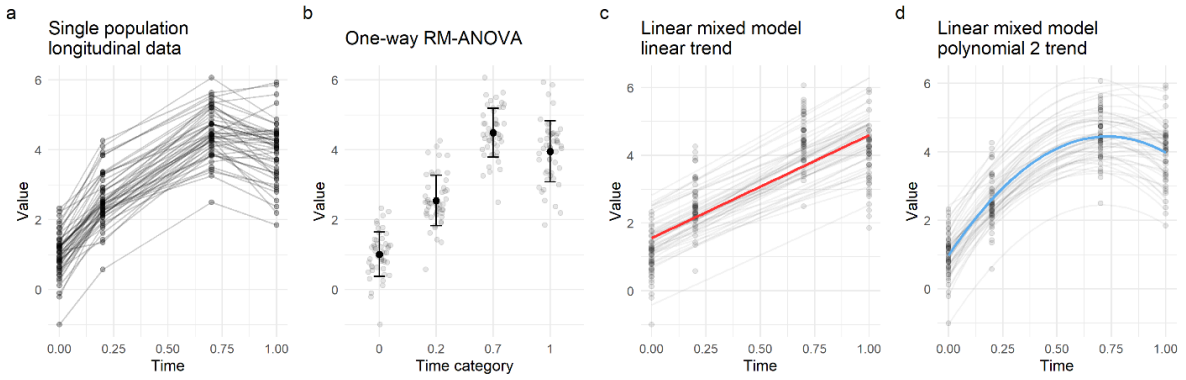


Figure 5. Single class longitudinal study simulated data. (a) Spaghetti plot of the simulated data per subject, note that time measurements are asynchronous (taken at non-constant intervals). A homogeneous curve trajectory can be observed. (b) Analysis of the simulated data by a one-way RM-ANOVA where time is considered a discrete categorical variable and it is not modeled as a time trend. (c) Application of linear mixed model to the simulated data with subject-specific random effects (intercept and slope), defining subject-specific trends (grey lines) and population trend (red line). (d) Application of linear mixed model with polynomial time transformation (of second order) to fit a quadratic trend (blue line).

Table 2. Single-class longitudinal data regression models for univariate analysis

General Notation
y: observed measure of response variable to model
t: the time predictor
i: index of the subject. $i \in \{1, \dots, n\}$ where n is the total number of subjects in the sample
j: index the occasion (timepoint) of measurement. j $\in \{1, \dots, J\}$ where J is the total number of timepoints per subject. Note that this allow for each subject to be measured at different timepoints.
ϵ: random error (residuals)

	Model Name	Scalar model specification	Assumptions
Eq. 3	One-way RM-ANOVA* (linear mixed effects model, time is categorical)	$y_{ij} = \mu_j + S_i + \varepsilon_{ij}$* where μ_j is the fixed effect of time j (the average at each timepoint); S_i is the random effect of subject i; ϵ_{ij} is the random error for the subject i and time j	$S_i \sim N(0, \sigma_s^2)$ $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ $S_i \perp \varepsilon_{ij}$ (independence) Compound symmetry of covariance matrix
Eq. 4	GCM** (linear mixed effects model, time is continuous with random intercept and slope)	$y_{ij} = a_i + \beta_i t_{ij} + \varepsilon_{ij}$** $a_i = \mu_\alpha + b_{0i}$ $\beta_i = \mu_\beta + b_{1i}$ where a_i is the intercept for subject i; β_i is the slope of subject i; and t_{ij} is the time predictor for subject i at time j. μ_a is the average population intercept (fixed effect); b_{0i} is the subject i intercept deviation from the population (random effect); $\mu_\beta$is the average population slope (fixed effect); b_{1i} is the subject i slope deviation from the population (random effect).	$b_{0i} \sim N(0, \sigma_{b0}^2)$ $b_{1i} \sim N(0, \sigma_{b1}^2)$ $\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon i}^2)$ $cov(b_{0i}, b_{1i}) \neq 0$ (random effects can covariate) Covariance matrix can assume different covariance structure

*Note that this model can be extended for two or more fixed terms (aka factors, variables) (two-way, three-way,…) and their interactions.
**Note that this model can be extended with other time-invariant, time-variant, and higher-order time predictors.

3.3 Finite Mixture Models for multi-class trajectory analysis

Identifying different unobserved classes of trajectory from the same data can be seen as the problem of clustering; finding homogeneous groups of patients with share characteristics unknown to exist in a heterogeneous population. In the case of trajectory, this can be achieved using a family of longitudinal finite mixture models (FMM), where the model assumes that the underlaying population is a mixture of latent trajectory classes (Lai et al., 2016; Ram & Grimm, 2009; van der Nest et al., 2020). Thus, FMM is an extension of LGM to multi-class by incorporating a mixture of conditional probabilities on time (Table 1, Eq. 2); a *post hoc* determination of classes with no prior knowledge of different trajectory subpopulations. There are distinct model specifications and assumptions for longitudinal FMM that receive different names (Nagin, 2014; Proust-Lima et al., 2017; van der Nest et al., 2020). In this work, we use two of these particular cases of longitudinal FMM: Latent Class Growth Analysis (LCGA) and Growth Mixture Models (GMM) (Jung & Wickrama, 2008; Muthén, 2004; Muthen & Asparouhov, 2008; Nagin, 2014; Ram & Grimm, 2009; van der Nest et al., 2020).

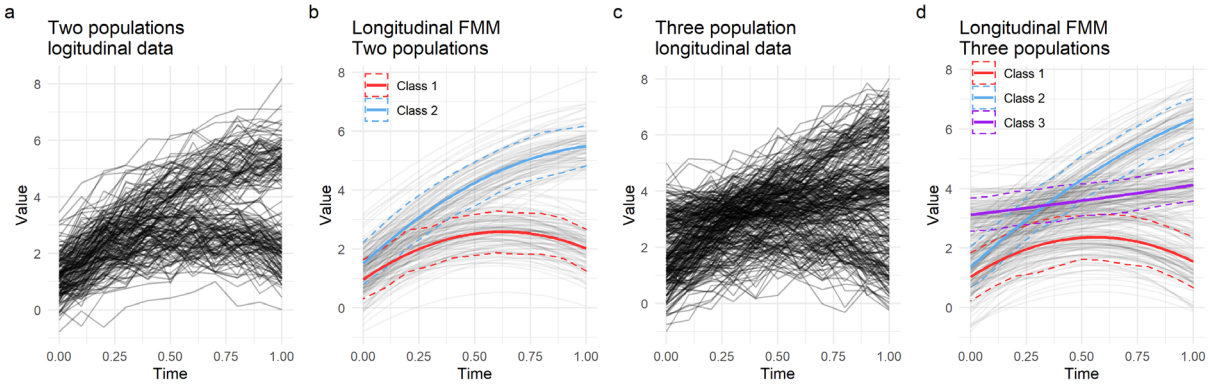


Figure 6. Application of FMM to simulated data of samples from different populations with distinct longitudinal trends. (a) Simulated longitudinal variable for heterogeneous sample composed of two different populations. Each line is a subject. (b) Application of longitudinal FMM with non-linear transformation to the two samples with intra-class subject-specific parameters (random effects). Subject trajectories are shown in gray, population level mean trajectory (solid line) and confidence interval (dotted line) for each class are shown in red and blue. (c) Simulation of three distinct populations to illustrate a less obvious trend choice. Each line is a subject. (d) Application of longitudinal FMM with non-linear transformation for three different classes.

3.3.1 Latent Class Growth Analysis (LCGA)

In LCGA, models assume that all subjects within a class follow the estimated mean trajectory for that class (no random effect), but these can be different between classes. This means that within-class between-subject variations on trajectory are not modeled and treated as the residual error for the trajectory of a given class. A typical particular case of LCGA is Group-based trajectory models (GBTM). These are a special case of LCGA because it assumes that the error variance is the same for all classes and time points (Nagin, 2014), which constrains the models even further. Table 3 Eq. 5 shows the specifications of LCGA models. It is appreciable in the equation that LCGA does not model between-subject variability since the lack of subject-specific parameters (random effects). This makes these models more tractable than other FMMs as they usually require a small set of parameters to be estimated, which helps model small datasets (Jung & Wickrama, 2008; van der Nest et al., 2020). As a trade-off, they usually find more trajectories than models that account for within-class subject variability. The obtained trajectories might not represent the subject-level trajectories but the

population class mean (Jung & Wickrama, 2008; Nagin, 2014). Some authors have argued that when there is certainty of model convergence and data is big enough, LCGA may serve as an initial exploratory and modeling step before using more complex models such as the ones discussed below (Jung & Wickrama, 2008; Van Der Leeden, 1998; van der Nest et al., 2020).

3.3.2 Growth Mixture Models (GMM)

These models are a mixture of GCM. Table 3 Eq. 6 shows the specifications for GMM, where both fixed and random effects are modeled. GMM then finds different latent trajectory classes through a mixture of linear mixed effect models, allowing for subject-specific variation within a class, and capturing differences between subjects (Jung & Wickrama, 2008; Ram & Grimm, 2009; Teuling et al., 2022). These models have higher flexibility than LCGA due to the incorporation of the random effects and the possibility to model different variance and covariance structures, adapting to complex designs and data generation processes. Thus, two main advantages of GMM over LCGA are the modeling of the individual subject and the differences between subjects and the major flexibility to accommodate different error structures (e.g., temporal autocorrelations). Moreover, once a GMM has been estimated, the model can also be used to predict the individual trajectory of an unseen patient and its deviation from the mean class trajectory. With the advent of precision medicine, GMM may be more appealing than LCGA for subject-specific prediction. These advantages come at the cost of increasing model complexity, with a higher number of parameters to estimate, which requires bigger sample sizes and computational costs.

Table 3. Multi-class longitudinal data regression models for univariate analysis (constrained to linear trends)

General Notation
y: observed measure of response variable to model
t: the time predictor
k: index of mixture component (class). $k \in \{1, \dots, K\}$
i: index of the subject. $i \in \{1, \dots, n\}$ where n is the total number of subjects in the sample
j: index the occasion (timepoint) of measurement. j $\in \{1, \dots, J\}$ where J is the total number of timepoints per subject
ε: random error (residuals)

	Model Name	Scalar model specification	Assumptions
Eq. 5	LCGA* (fixed effect model per class)	$y_{ij}^k = \alpha^k + \beta^k t_{ij} + \varepsilon_{ij}^k$* where α^k is the intercept for class k and β^k is the slope of for class k.	$\varepsilon_{ij}^k \sim N(0, \sigma_{\varepsilon kj}^2)$
Eq. 6	GMM* (mixed effect model per class)	$y_{ij}^k = \alpha_i^k + \beta_i^k t_{ij} + \varepsilon_{ij}^k$* $\alpha_i^k = \mu_\alpha^k + b_{0i}^k$ $\beta_i^k = \mu_\beta^k + b_{1i}^k$ where α_i^k is the intercept for subject i in class k and β_i^k is the slope of subject i in class k. μ_α^k is the average population intercept (fixed effect) for class k; b_{0i}^k is the subject i intercept deviation from the population (random effect) for class k; μ_β^k is the average population slope (fixed effect) for class k; b_{1i}^k is the subject i slope deviation from the population (random effect) for class k.	$b_{0i}^k \sim N(0, \sigma_{b0k}^2)$ $b_{1i}^k \sim N(0, \sigma_{b1k}^2)$ $\varepsilon_{ij}^k \sim N(0, \sigma_{\varepsilon kj}^2)$ $cov(b_{0i}^k, b_{1i}^k) \neq 0$ (random effects can covariate) Covariance matrix can assume different covariance structure per class

*Note that this model can be extended with other time-invariant, time-variant, and higher-order time predictors.

3.4 A note on non-linear trajectories

It is often the case that the latent growth process of a longitudinal variable is not a line but a curve. All models considered so far are linear models, meaning that they are specified as linear combinations of variables and parameters, assuming a latent linear trend for the trajectories. This can cause the mischaracterization of very curvy trajectories not well represented by a line. Curve trajectories can be modeled from the linear framework by specifying non-linear transformations on the time variable, such as polynomials of order higher than one and splines (Hastie et al., 2009). In the context of the multi-class problem, each trajectory class may follow a different curve shape that can be modeled by considering non-linear transformations over time. Figure 7 illustrates this using a synthetic example where three distinctive classes of curve trajectories have been simulated. While a linear trend is acceptable for the third class, a line does not reflect the actual changes over time for the first and second classes. The model can be improved by considering polynomial transformations. For example, Eq. 6 can be extended to a quadratic form (polynomial of order 2) of time by the equation $y_{ij}^{k} = \alpha_i^k + \beta_{1i}^k t_{ij} + \beta_{2i}^k t_{ij}^2 + \varepsilon_{ij}^k$ where β_{2i}^k represents the unknown fixed effect for the square of time. Not much gain can be observed for the third class, but the curved model trend adjusts to the observed data much better than a line for the first and second class.

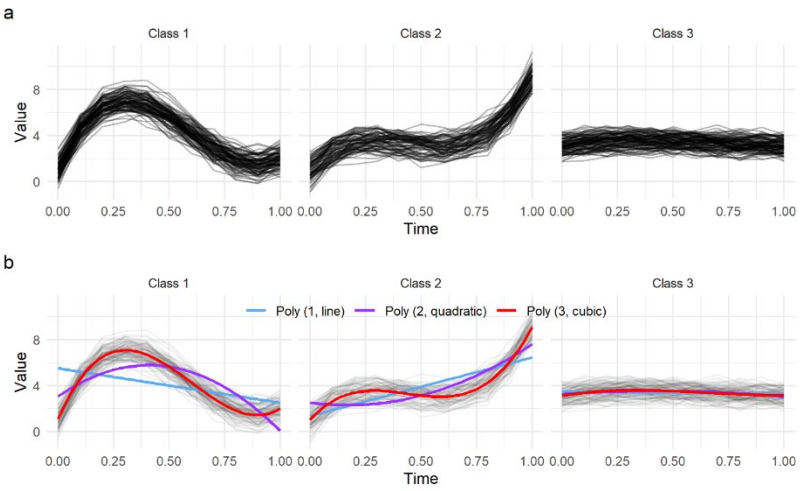


Figure 7. Demonstration of polynomial fits for a three population simulated data with distinct trajectories. (a) Simulated data for each one the population with distinctive curve growth. Each line is a subject. (b) Three different polynomial class-specific longitudinal fits for orders 1, 2 and 3. Poly 1: linear, Poly 2: quadratic, Poly 3: cubic.

Besides the evident flexibility of modeling a curve rather than a line, this comes at the expense of increasing model complexity since additional parameters need to be estimated to describe the curved shape of the subject-specific and population levels. Given that all these transformations use the same underlying model and are estimated through the same method, a search for the most parsimonious model can be done through well-described model selection methodologies and goodness-of-fit metrics (see model estimation and selection section). Thus, model selection can limit unnecessary model parametrization that can cause issues such as overfitting, non-convergence, and low precision on the parameter estimates.

3.5 Multivariate trajectories

So far, we have considered modeling the trajectory of a single measure either in a single-class trajectory, analyzing the difference in the trajectory of known subject subgroups (i.e., covariates), or discovering unknown trajectory classes. Either way, those approaches are

univariate since the trend of a single variable is modeled. However, biological and medical processes are interconnected, and several events can change over time. For example, the numbers of red blood cells and white blood cells may change following similar or distinct trends. Thus, it is often the case that studies capture more than one variable measured over time for the same group of subjects. In those cases, one may want to analyze trajectories across those multiple variables and their dependencies, in other words: a joint multivariate analysis of longitudinal variables. There are a few methods that we can take into consideration here depending on the research question, whether the variables of interest are continuous, categorical, or a mix, if the data is balanced on time or not, and a few other factors and assumptions about the data (Genolini et al., 2015; Nagin et al., 2018; Verbeke et al., 2014).

For completeness, an option is the analysis of variance approach. Similar to the univariate one-way repeated measures ANOVA, a one-way repeated measures multivariate analysis of variance (RM-MANOVA) can be specified (Krzysko et al., 2014). Similar limitations then those described above for RM-ANOVA apply for RM-MANOVA, notably that they are limited to balanced data and gaussian measures. Other options are from the latent variable framework. We can generally categorize these methods on whether the joint modeling assumes a latent structure on the time trend or for all variables at each time point (Verbeke et al., 2014). For instance, this latter approach can be achieved through factor analysis or principal component analysis frameworks assuming balanced designs for dimensionality reduction, followed by a longitudinal analysis of the latent factors or principal components (Verbeke et al., 2014). Thus, these methods do not model time on the original variable space but on the latent constructs per time point. For methods that model the latent time trajectory, the assumption is that the observed longitudinal variables trend are realizations of a single longitudinal latent variable growth curve (Nagin et al., 2018; Verbeke et al., 2014). These can be specified through, for instance, multivariate extensions of LGM with joint conditional probabilities that postulate the interrelationship between longitudinal variables (Nagin et al., 2018; Verbeke et al., 2014). An in-depth treatment of this topic is out of scope; Verbeke and collaborators have published a comprehensive review for joint trajectory modeling (Verbeke et al., 2014).

Multivariate longitudinal analysis can also be extended to the multi-class model to determine unobserved homogeneous groups of subjects with similar temporal evolution on more than one longitudinal variable. One option is an extension of the joint multivariate LGM methods stated above by incorporating a mixture of distributions. Here, different longitudinal observed variables are jointly analyzed in the context of finding different groups of subjects with similar latent evolutions over a single or multiple joint latent trajectories (Lai et al., 2016; Proust-Lima et al., 2017). An alternative approach for the multi-class multi-variable problem is the group-based multi-trajectory modeling (GBMT) that Nagin and his colleagues described (Nagin et al., 2018). The goal is to find subgroups of patients that share similar trajectories across different longitudinal variables. Thus, while the multi-class joint-trajectory approach finds subgroups of patients with similar multi-variable latent trends, in the GBMT, subjects are grouped based on their similarities to each observed longitudinal variable. Bellow, I discuss these two major approaches for multi-class multivariate trajectory analysis.

3.5.1 Multi-class Joint-trajectory LGM modeling

For the joint-trajectory modeling, multivariate mixture LGM can be considered, allowing for unbalanced data with missing values over time, mixed types of variables (e.g., one variable can be continuous, another count), and non-linear trends. Joint-trajectory LGM is a

generalization of univariate LGM for the multivariate case, where the goal is to model and analyze the interrelationship of two or more variables that follow similar trends or relate to each other in some form over time (whether or not the variables are observer over the same period) (Verbeke et al., 2014). The joint model estimates trajectories for the considered temporal variables (either for a single class or for a multi-class model) and links them with tables of joint conditional probabilities. These tables capture the probability of, for example, following a trajectory for variable two, given that the subject is following a trajectory for variable 1 (Nagin et al., 2018). Therefore, joint-trajectory models can analyze the link between trajectories of different variables over time that are thought to be realizations of a common underlying process (e.g., temporal inflammation changes after injury; not to confuse with the single variable latent process, nor the latent class trajectory). Different model assumptions can be considered, namely the shared parameters and random-effects models (Verbeke et al., 2014). The former assumes the subject-specific parameters to be the same (shared) across the different longitudinal variables. This often imposes unrealistic constraints on the relationship between longitudinal variables (Verbeke et al., 2014). These assumptions can be relaxed by considering the subject-specific parameters (the random effects) to be unique for each longitudinal variable. The major drawback of these models is that they become more intractable as the number of considered variables to model increases due to the rapid increase of pairwise conditional probability that must be calculated and stored during estimation.

3.5.2 Multi-trajectory modeling

Like joint-trajectory models, multi-trajectory models are designed as a multivariate approach by simultaneously considering the trajectory of more than one variable over time in a multi-class context. The difference is that multi-trajectory models do not estimate the linkage between variables in pairwise conditional probability tables, but they estimate trajectory classes for multiple variables instead of a single latent growth curve (Nagin et al., 2018). This generates classes or groups of subjects that follow similar trajectories over different measured variables, independently of whether those variables follow the same trend or not. This approach is of recent development as an extension of GBTM (Nagin et al., 2018). The subject-specific likelihood of multivariable conditional on time is given by a multivariate extension of Eq. 2 for V longitudinal response variables column vector (Eq. 7). This indicates that GBMTs are estimated as the conditional probability of the multivariable longitudinal response over time.

$$P\left(Y_i^1, Y_i^2, \dots Y_i^V \middle| time_i\right) = \sum_{k=1}^{K} \pi_k \cdot \left[\prod_{v=1}^{V} P_v(Y_i^v | \, time_i, k)\right]$$

$$\text{Where} \quad P_v(Y_i^v | time_i, k) = \prod_{j=1}^{J^v} p_v(y_{ij}^v, k) \qquad \text{(Eq. 7)}$$

Note that the v subscript over the J timepoints on the second part of the equation indicates that the longitudinal variables can be measured over different times. These models assume conditional independence at the individual level, which means that conditional on membership to group *k,* Y_i^V are independently distributed.

3.6 Model estimation and selection

The models specified above (GCM, LCGA, GMM, joint-trajectory, GBMT) are estimated using maximum likelihood approaches (Nagin, 2014; Nagin et al., 2018; van der

Nest et al., 2020) of a model-specific likelihood function. As such, they share known characteristics of maximum likelihood estimates, such as their robustness and the fact that the estimates are normally distributed asymptotically. The general form of the subject-specific likelihood function for GCM is shown in Eq. 1 and for FMM in Eq. 2. Specific forms of the conditional probability density function can be considered depending on the different nature of the outcome metrics. For instance, for Gaussian distributed outcomes, y_i^k (the longitudinal sequence for subject i and class k) can be assumed to follow a multivariate normal distribution with mean μ^k and variance Σ^k [$y_i^k \sim MVN(\mu^k, \Sigma^k)$] (van der Nest et al., 2020). For count outcomes, the zero-inflated Poisson distribution can be used; for censored ones, the censored normal distribution, and for binary ones, the logit distribution (Nagin, 2014). The goal then is to estimate a set of parameters θ that maximizes the likelihood of y_i^k where θ are specified by the assumed distribution function. In FMM, the mixing probability for the class k (π_k) also needs to be estimated. Thus, the shape and probability of membership to the trajectories are determined by the specific parameters of the model (Nagin, 2014).

In general, if the likelihood function is differentiable, derivative methods for finding function maxima can be applied. In most cases, numerical iterative methods are necessary to determine or approximate the maximum of the likelihood function, such as gradient descent, the Newton-Raphson method, or the Expectation Maximization method (EM) (Jennrich & Sampson, 1976; Redner & Walker, 1984). Mixed Mixture models can be maximized using EM or Newton-Raphson methods (Proust-Lima et al., 2017; Redner & Walker, 1984). The specifics of estimation procedures to find the maximum likelihood estimates are out of scope for this work. Commonly, their use depends on their implementations of different software packages. The following section focuses on the estimation methods implemented in the software used during the realization of this work.

3.6.1 A brief note on software and model estimation

Based on the complexity of FMM approaches, all model options and estimation procedures are unlikely to be available in a general-purpose statistical package. Usually, estimating FMM would require specialized software; therefore, model specifications, estimation algorithms, and model selection methods may be limited by their availability and implementation. Van der Nest and colleagues provide a comprehensive list of different software for performing FMM available up to 2020, including a summary of their characteristics (van der Nest et al., 2020). In this work, we used the R programming language (R Core Team, 2021) for being a mature scripting language designed explicitly around statistics and data science but with extended capabilities for being general-purpose. The specifics of R packages used can be found in chapter 4. The main modeling packages used in this work are the *lcmm* R package (Proust-Lima et al., 2017) for fitting univariate multi-class LCGA and GMM and the *gbmt* R package (Magrini, 2021/2022) for fitting GBMT models.

The *lcmm* package. The package provides modeling and utility tools for estimating latent class mixed models (lcmm) in its linear mixed form and its extension to latent process and joint modeling. Maximum likelihood estimation is performed through an iterative procedure extended from the Marquardt algorithm, a Newton-Raphson method of finding the solution of a function by linear approximations of its gradient. These methods iteratively update the set of parameters to estimate using the gradient (partial derivatives for each parameter) of the log-likelihood function until some criteria are reached, the point at which the algorithm is considered to have converged. *Lcmm* updates the vector of parameters θ for the iteration $l + 1$ using the equation:

$$\theta^{l+1} = \theta^l - \delta\left(\tilde{H}^l\right)^{-1}\nabla(L(\theta^l))$$

Where δ is the step, $\widetilde{H}$ is a diagonal-inflated Hessian matrix (containing all second partial derivatives), θ^l is the set of parameters at iteration l and $\nabla(L(\theta^l)$ is the gradient of the log-likelihood function at iteration l (Proust-Lima et al., 2017). The package uses three criteria for convergence: parameter stability, where the parameters estimate of one iteration has changed with respect to the previous iteration no more than a given threshold (a.k.a. tolerance); log-likelihood stability, where the change of log-likelihood between two consecutive iterations is smaller than a threshold; and size of derivatives, where the size of the gradient at a given iteration is smaller than a threshold. *Lcmm* requires these three criteria for convergence. In addition, the variance-covariance matrix of the maximum likelihood estimates is estimated by the inverse of the Hessian matrix. The log-likelihood is given by the sum of the logarithm of each subject likelihood function (Eq. 2). The shape of the specific likelihood function in Eq. 2 is set depending on whether a linear mixed model or a latent mixed process is used. For linear mixed models, a multivariate normal density function is used. In the case of latent process models, the individual conditional expectation to time is modeled as an extension of a linear mixed model through a latent function with different link functions depending on the nature of the variable to model. In the case of this work, and given the implementation by the *lcmm* package, we use the Beta density and quadratic I-splines as link functions to model non-gaussian continuous variables (see chapter 4). Moreover, the package defines π_k by a multinomial logistic model of class membership considered a discrete random variable. See (Proust-Lima et al., 2017) for more details on model estimation.

The *gbmt* package. This package has been recently released to CRAN (March 2022), and its implementation is not described in an article. Part of the methodology used by the author is described here (Magrini, 2022). The package implements an EM algorithm for the maximum likelihood estimation of group-based multi-trajectory models as specified in Eq. 7. The EM algorithm is useful and robust for estimating models in case of incomplete data. This allows for the *gbmt* package to estimate multi-trajectory models with missing temporal data and values taken at different timepoints across subjects, which is a characteristic of the longitudinal laboratory data analyzed in this work. In order to be able to compare and derive trajectory groups across different variables, scaling is commonly done. The package incorporates four scaling procedures: centering, standardization (centering and dividing by standard deviation), division by the sample mean, and logarithmic division by the sample mean. Then, the EM algorithm is employed as a maximum likelihood estimator initialized from random values for missing data and parameters. It follows an iterative alternation between expectation (E) and maximization (M) steps until convergence (Magrini, 2022; Redner & Walker, 1984). In the case of GBMT, the E-step consists of computing the posterior probability of each group for each subject after computing the likelihood by:

$$P(K_i = k \mid y_i) \equiv \pi_{ik} = \frac{\hat{\pi}_k \prod_{t=1}^{T} \emptyset\left(y_{i,t} \mid \hat{\beta}_k, \hat{\Sigma}_k\right)}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{t=1}^{T} \emptyset\left(y_{i,t} \mid \hat{\beta}_k, \hat{\Sigma}_k\right)}$$

Where the hat symbol marks the current step estimate of the parameters and $\emptyset$ is the multivariate normal density function. The M-step obtains the maximum likelihood estimate of the parameters. The probability for group *k* is obtained by averaging the posterior probabilities of group *k* across all subjects: $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \pi_{i,k}$. Missing values under Missing at Random (MAR) assumption (Rubin, 1976) are imputed during the M-step based on the iteration estimates of the parameters and the observed values (Magrini, 2022). This process continues until the convergence of the likelihood. To prevent suboptimal solutions by local maxima, the EM algorithm is randomly initiated several times, and the estimates from the highest likelihood are retained.

3.6.2 Model selection and goodness-of-fit metrics

In the multi-class context, the number of classes (number of mixtures, k) is *a priori* unknown, and it is a fixed parameter provided by the analyst. Finding the correct number of classes, often called class enumeration (van der Nest et al., 2020), is a common problem in clustering methods of identifying the number of effective groups in our data. One of the advantages of using model-based approaches for clustering is that we can use all tools available in probabilistic modeling to determine adequate models, turning the problem of the number of classes of a problem into model selection (van der Nest et al., 2020). In addition to class enumeration, model selection can be used to determine the best non-linear specification for the trajectories (see the previous section). Several model selection procedures and metrics are proposed for longitudinal FMM; Van der Nest and colleagues provide a comprehensive summary (van der Nest et al., 2020). Generally, these metric-based statistics of goodness-of-fit have different properties and can be categorized in log-likelihood-based metrics (such as information criteria), subject classification performance, and distribution properties of the data. There is no single best method since their performance depends on the data characteristics at hand (e.g., subpopulation heterogeneity, sample size, outcome distribution), and therefore, the ultimate selection of the best model may require more than one approach. In this work, we used the following different metrics and criteria:

Log-likelihood information criteria. These are metrics of fit that balance the log of the maximum likelihood with the number of parameters to estimate. This balance is needed because the maximum likelihood is monotonic to the increase of model parametrization. Therefore, a penalty is introduced to select the model with the best fit with the minimum parametrization possible. An example of this form of metric for model fit is the Bayesian Information Criterion (BIC) (Schwarz, 1978): $-2\log[L(K)] + \log(n)\,[m(K)]$, where $L(K)$ is the maximum likelihood of the model with *K* classes, n is the sample size, and $m(K)$ the number of independent parameters for the model with *K* classes.

Subject classification performance. These metrics are based on the classification performance of the model for the subject data used to estimate the model (model training in machine learning lingo). In FMM, subjects are classified (assigned to the class) by the highest posterior probability (pp_i) for that subject, where $pp_{ik} = P\,(ky_i) = \frac{\hat{\pi}_k \hat{P}^k(y_i)}{\sum_{h=1}^{K} \hat{\pi}_h \hat{P}^h(y_i)}$. The closest pp_{ik} for a given class is to 1 indicates better classification, and therefore better models. There are several metrics based on classification performance (van der Nest et al., 2020). Here we use APPA (average posterior probability of assignment) and the ICL-BIC (integrated classification likelihood, BIC approximation), a hybrid metric that combines both likelihood-based criteria and classification performance. $APPA_k = \frac{1}{n_k}\sum_{i=1}^{n_k} pp_{ik}$ and $ICL - BIC = -2\log[L(K)] + \log(n)\,[m(K)] + 2E(K)$, where ***E(K)*** is the entropy of the *K* class model. Entropy measures classification uncertainty in class assignment, where higher values imply higher classification uncertainty. Therefore, its addition to BIC in the ICL-BIC acts as an extra penalty term based on classification performance. $E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} pp_{ik}\log\,[pp_{ik}] \geq 0.$

4 Methodology

4.1 Data

4.1.1 MIMIC data: electronic health records for modeling

Two different epochs of the MIMIC (Medical Information Mart for Intensive Care) dataset were used for trajectory modeling. MIMIC is an extensive single-center database with EHR of patients admitted to critical care units of the Beth Israel Deaconess Medical Center in the USA. The two epochs are the MIMIC-III, with 46,520 patients from 2001 to 2012 (Johnson et al., 2016)(Johnson et al., 2016), and the MIMIC-IV, with 382,278 patients from 2008 to 2019. Both databases were accessed under a data use agreement (DUA) through the PhysioNet project (Goldberger et al., 2000). Data was downloaded from physionet.org. Both MIMIC databases (DB) are relational DB structured in tables. Documentation about the DB schema can be found at mimic.mit.edu/docs.

Cohort selection: For cohort selection, we used the International Statistical Classification of Disease (ICD) codes, version 9/10 (www.who.int/classifications/classification-of-diseases). Table 4 shows a description of the tables downloaded for each DB. The variables used from those tables can be seen in the annexed table 1. The spine trauma and traumatic SCI patients cohort was constructed using ICD9/ICD10 diagnostic classification codes. The inclusion criteria for the patient search were: adults (>= 15 years), acute patients with ICD diagnostic codes with a specification for SCI or spine trauma (vertebral fracture), and admitted to the hospital as an emergency. These include all codes of the series ICD9: 952, 953, 806 and 805, and ICD10: S120, S121, S122, S123, S124, S125, S126, S128, S129, S140, S141, S142, S220, S240, S241, S242, S320, S321, S340, S341, S342, S343. The full list of codes and descriptions can be seen in the respective files in the GitHub repository. From the MIMIC diagnostics table, we selected the first hospital admission per patient that shows in the DB, and that has a presence of selected ICD9/10 codes in their hospital stay.

Table 4. DataBase tables accessed for each MIMIC version

MIMIC-III table	MIMIC-IV table	Description
PATIENTS	core/patients	Demographic information for each patient
ADMISSIONS	core/admissions	Information for each unique hospitalization for each patient
DIAGNOSES_ICD	hosp/diagnoses_icd	Hospital assigned diagnoses ICD codes
PROCEDURES_ICD	hosp/procedures_icd	ICD code for procedures
LABEVENTS	hosp/labevents	Laboratory events and values for each patient
D_LABITEMS	hosp/d_labitems	Dictionary for each laboratory assay
D_ICD_DIAGNOSES	hosp/d_icd_diagnoses	Dictionary for each ICD diagnoses
D_ICD_PROCEDURES	hosp/d_icd_procedures	Dictionary for each ICD procedures

There is a potential overlap of patients between both versions of MIMIC. In the transition to MIMIC-IV, subjects' identifiers were not preserved (see MIMIC documentation). Both DB has overlapping periods of catchment (MIMIC-III from 2001 to 2012; MIMIC-IV from 2008 to 2019), and there is the potential for the same patient represented in both datasets. In order to prevent patient duplication, we excluded MIMIC-IV patients with the same sequence of ICD diagnostics (order sequence of code and number) with the assumption that the same sequence of diagnostics in two different patients is improbable. Although this filter might exclude some non-overlapping patients, considering duplicated entries as independent poses a higher risk for modeling than the potential of reducing patient catchment. A total of 515 form MIMIC-IV patients were excluded for the risk of duplication, all of them admitted during the overlapping years of catchment of both DB. Then we harmonized patient demographics data in a single dataset from both MIMIC epochs (harmonization strategy can be seen in the

annexed code). Finally, we selected patients admitted to the hospital in an emergency. A flow diagram and demographics table can be seen in the next chapter.

Extraction of laboratory values: Laboratory values are stored in the *labevents* table in each DB, and the metadata information for each laboratory assay can be found in the *d_labitems* table. The laboratory values for the selected cohort and hospital stay were extracted, and the time of laboratory sample collection from admission date and time was calculated. Laboratory values with missing LOINC codes were excluded.

Extraction of demographics and stay characteristics: Demographics included age, gender, ethnicity, and insurance type. Stay characteristics included length of stay (calculated in days), number of ICD diagnostics, admission type, admission location, and discharge location. MIMIC-III does not provide patient age directly, but it can be computed. Dates for each patient are shifted for privacy reasons, but temporal consistency is maintained. Age at hospital admission was calculated by subtracting admission date (ADMITTIME) from birth data (DOB). MIMIC-IV does not provide age, and it is not computable. A range of 3 years can be obtained through an “anchor age” variable (anchore_age). Gender did not require any data cleaning. Ethnicity, insurance, admission type, admission location, and discharge location were harmonized between DBs by collapsing categories with modeling information lost (see annexed code). Only patients with admission type specified as an emergency were included. Length of stay was calculated by subtracting discharge date and time (DISCHTIME) from admission date and time (ADMITTIME).

4.1.2 TRACK-SCI: prospective patients for prediction

For the classification of trajectories in new patients, we use data from 137 patients enrolled in the Transforming Research and Clinical Knowledge in Spinal Cord Injury (TRACK-SCI) study (Tsolinas et al., 2020), a longitudinal observational cohort study at the Zuckerberg San Francisco General Hospital and at the University of California San Francisco. TRACK-SCI collects highly granular in-hospital and post-hospitalization data, including laboratory assays and long-term neurological outcomes. Data was received de-identified. Harmonization of the laboratory names was conducted to pair the analytes in MIMIC with those in TRACK-SCI. The dynamic range for each analyte was compared to ensure that the values were in the same scale. In addition, the ASIA (American Spinal Injury Association) Impairment Severity (AIS) grade was extracted as an outcome metric (Betz et al., 2019; Roberts et al., 2017). AIS grade measures the level of neurological impairment after SCI on a 5-point ordinal scale (A to E).

4.2 Laboratory analyte exploratory data analysis and data cleaning

Exploratory data analysis (EDA) is performed to summarize the data. Temporal spaghetti and marginal density plots are generated to understand the amount of available data for the laboratory analytes, the underlying distribution, the potential presence of outliers, and non-linearities. This information is used to curate the data, as well as to make decisions before modeling. Summary tables for the amount of laboratory data are provided. Of 415 unique laboratory analytes extracted, the 20 most common were present above 80% of the selected subjects from MIMIC. Therefore, only these 20, referred to as the modeling set, will be used for further analysis (see results). Some spikes on the temporal trends for of the laboratory

analytes were observed from the spaghetti plots and as extreme values in the marginal distributions. These values can constitute anomalies in the data. In order to determine if these extreme values were unlikely given subject-specific distribution, we applied two filters. First, 0 values were excluded as these are unprovable in any of the modeling set of analytes. Next, we applied a variation of John Tukey's rule for outlier determination as proportional to the interquartile (IQR) range (Tukey, 1977). In this case, we filtered out extreme values with $Lower\ limit = quantile_{20} - 1.5IQR$ and $Upper\ limit = quantile_{80} + 1.5IQR$ on the subject-specific marginal distribution for each analyte. Note that 20% and 80% quantiles instead of Tukey's 25% and 75% were used to be more permissive, especially for skewed distributions where Tukey's can be too restrictive (Seo, 2006).

4.3 Trajectory modeling

Univariate multi-class trajectory modeling: Following van der Nest and colleagues' recommendations (van der Nest et al., 2020), an initial exploratory analysis for each one of the modeling set analytes was conducted by LCGA, a particular case of GMM which restricts intra-class parameters to be invariant across subjects (no random effects); see chapter 3. They suggest using restricted models such as LCGA to initially explore the heterogeneity in the trajectories, approximating the number of classes, finding the proper model specifications, and performing initial exploration on non-linearities. For each analyte, we conducted a linear search for the number of classes ranging from 1 to 5. We investigated linear trajectories and polynomial transformations over time for orders of 2 (quadratic) and 3 (cubic). Finally, given the nature of the different analytes, we explored the use of three different link functions for the latent process: a linear link (i.e., linear mixed model), considering the analyte as a continuous Gaussian; the Beta density function, and a quadratic I-spline with 3 knots on the tertials of the data (Proust-Lima et al., 2017). These last two were considered to model continuous variables with potentially not Gaussian distributions. A total of 900 models were specified. These models were fitted in R using the *lcmm::lcmm()* function with no random effects. BIC, the ICL(van der Nest et al., 2020), and the APPA were calculated for each model. Both BIC and ICL were used as the primary decision criteria for model selection, where models with lower values were considered to provide a better fit. APPA was used as a secondary decision model selection tool, where only models with APPA > 0.7 were considered. In addition, the percentage of subjects attributed to each class was used, where at least 1% of the subject were represented in each class. Given the computational cost of this procedure, the process was parallelized using the *parallel* R package in 15 CPU cores and 36Gb of RAM machine (Kuhn, 2021).

After model exploration, plausible models were selected for each analyte following the specified criteria. Next, GMM specifying random effects were fitted for such models. In order to relax the symmetry constraints of polynomial transformations, natural splines instead of polynomials were used to consider a non-linear basis for the time trajectories. The number of degrees for the splines was chosen from the number of degrees selected during the LCGA exploratory analysis. A total of 122 models were specified. The posterior probability of class membership for each analyte and subject was calculated, and class membership was assigned based on the highest posterior probability.

Multi-trajectory modeling: to determine groups of patients that share similar trajectories across different analytes, we used GBMT analysis (Nagin et al., 2018). Several models were fitted using the *gbmt* R package to explore the parameter space: the number of mixture components k = {1, ..., 8}, and polynomial transformations d = {1, …, 4}. BIC and APPA were used for model selection criteria. Since several analytes are correlated, GBMTs

were fitted with glucose, hematocrit, platelet count, RDW, and white blood cell count as predictors. As in the case of univariate analysis, class membership was assigned based on the maximal posterior probability.

Trajectory membership probability of TRACK-SCI subject (unseen by the model) was calculated, and class trajectory was assigned for the trajectory class with a higher probability for each subject.

4.4 Predictive modeling experiments

We designed three predictive modeling experiments to study trajectory membership for each analyte as a biomarker. Experiment I was set to predict whether a patient would die in-hospital, Experiment II was set to predict whether a spine trauma patient had an SCI, and Experiment III was set to predict SCI severity calculated by the latest known in-hospital AIS grade for the TRACK-SCI cohort. For each experiment, the posterior probability for each analyte class trajectory was calculated and used as predictors in an ElasticNet model. In other to simulate dynamic predictions, each ElasticNet model was restricted to use posterior probabilities of trajectories calculated from data up to a time cutoff = {1, 3, 7, 14, 21} days. Thus, for example, models with a cutoff of one day would use the posterior probabilities for each class trajectory for each analyte calculated with in-hospital data from hospital arrival up to day 1. Each experiment type and cutoff combination were run 25 times with changes in random seed for the repeated runs but with fixed seed across cutoff for comparability.

ElasticNet models were fitted with the “glmnet” model in the R caret package (Kuhn, 2021) with a five cross-validation setup for tunning alpha and lambda hyperparameters. Best hyperparameters were selected as specified by default in the package. For Experiments I and II, model performance was evaluated for the sample used for training (in-train) and a sample left aside for testing (out-train) with a split 80/20 for training and testing. Given the imbalances between the binary classes in Experiment I and II, down-sampling of the majority class was used for training to improve model performance. Thus, both categories of the target variable had equal prevalence in the training dataset. For the test sample, the original proportion between classes was maintained, reflecting the prevalence of the specific target to predict in a real-world scenario. Experiment III had no train/test split given the small sample size. The generalizability of out-of-sample prediction was estimated using leave-one-out validation. As performance metric, for all models, a smooth ROC curve and area under the curve were calculated using the pROC R package (Robin et al., 2011).

4.5 Statistics and software

Patients’ trajectories were characterized based on clinical and demographic features extracted from the databases: age, gender, ethnicity, cohort group (SCI with vertebral fracture, SCI without vertebral fracture, spine trauma with no SCI), length of hospital stay, whether the patient died in hospital, and the number of ICD diagnostics. Differences between trajectory groups and patient cohorts were analyzed using ANOVA or t-test for continuous variables and Fisher exact test for categorical variables. P-values were adjusted for false discovery rate (FDR) by the Benjamini-Hochberg method, and q-values were reported. The level of significance was set at $q < 0.05$.

This work was performed in R (R Core Team, 2021). For fitting GMM and LCGA, we used the *lcmm* R package (Proust-Lima et al., 2017). For GBMT, we will use the *gbmt* R package (Magrini, 2022). ElasticNet models for prediction experiments were trained using the caret R package (Kuhn, 2021). The details of the complete code reproducing this research can be found on GitHub (https://github.com/ATEspin/UOC-TFM).

5 Results

5.1 Data building for trajectory modeling

5.1.1 Cohort extraction

The SCI and spinal trauma patients cohort was built from the MIMIC-III and MIMIC-IV databases. Figure 8 shows a flow diagram of the construction of the dataset; a total of 1194 and 4429 unique patients were found for MIMIC-III and MIMIC-IV, respectively. An issue of using both III and IV epochs of MIMIC is the potential for overlapping patients since both databases had overlapping years of the catchment (MIMIC-III from 2001 to 2012; MIMIC-IV from 2008 to 2019), and the authors of MIMIC did not respect the subject identifier from MIMIC-III to IV. These patients were filtered from MIMIC-IV previous harmonizing and merging datasets. After filtering, a total of 5208 patients were identified. Of those, 2615 patients were filtered out, either for not being admitted to the hospital as an emergency or given the results of the exploratory analysis for the laboratory values as described in the next section. A final cohort of 2615 patients was available with processed laboratory analyte data for further analysis and modeling.

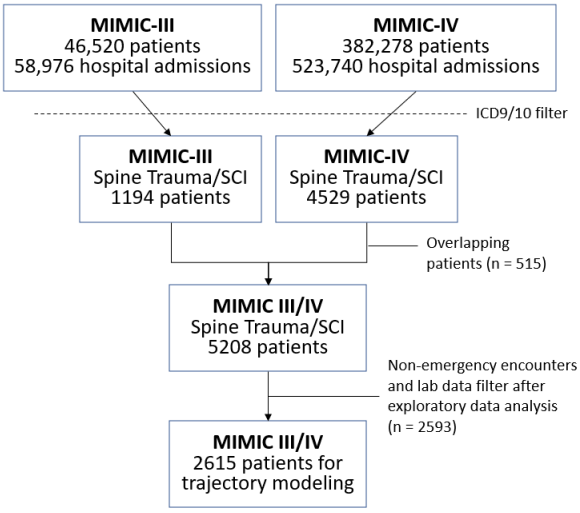


Figure 8. Flow diagram of cohort build. Subjects from MIMIC-III/IV where first filtered based on their ICD9 and ICD10 diagnostic codes. Then, potential overlapping patients were filtered from MIMIC-IV. After laboratory analyte data extraction and data cleaning, patients with less than 3 measures for any of the 20 most common analytes where excluded. The data from a total of 2599 patients from both MIMIC databases were used for modeling.

5.1.2 Laboratory analyte exploratory data analysis

A total of 413 distinct laboratory analytes were found in the data. Table 5 cross-tabulates the number of analytes per category and fluid sample. Of the 413, 157 were categorized as hematology, 161 as chemistry, and 25 as blood gases. Regarding the fluid sample, the fluid with more distinct analytes was blood, with blood chemistry and hematology being the two categories with more analytes.

Table 5. Cross-tabulation of laboratory analytes per category and fluid sample

Fluid sample \ Category	Blood Gas	Chemistry	Hematology	Total
Ascites	0	9	13	**22**
Blood	24	104	71	**199**
Cerebrospinal Fluid (CSF)	0	3	13	**16**
Joint Fluid	0	1	11	**12**
Other Body Fluid	1	9	14	**24**
Pleural	0	8	15	**23**
Urine	0	27	20	**47**
Unknown	0	0	0	**70**
Total	**25**	**161**	**157**	**413**

Between 90 and 98% of the patients presented the 20 most common analytes, as shown in Annexed Table 2 (the complete list can be found in the code repository); all were analytes measured in blood from the hematology and chemistry categories. Of the remaining analytes, the proportion of patients they were obtained from dropped to <70%, and 342 out of the 413 analytes were present for only 10% of the patients or less. We then selected the 20 most common analytes to consider for analysis, which we refer to as the modeling set.

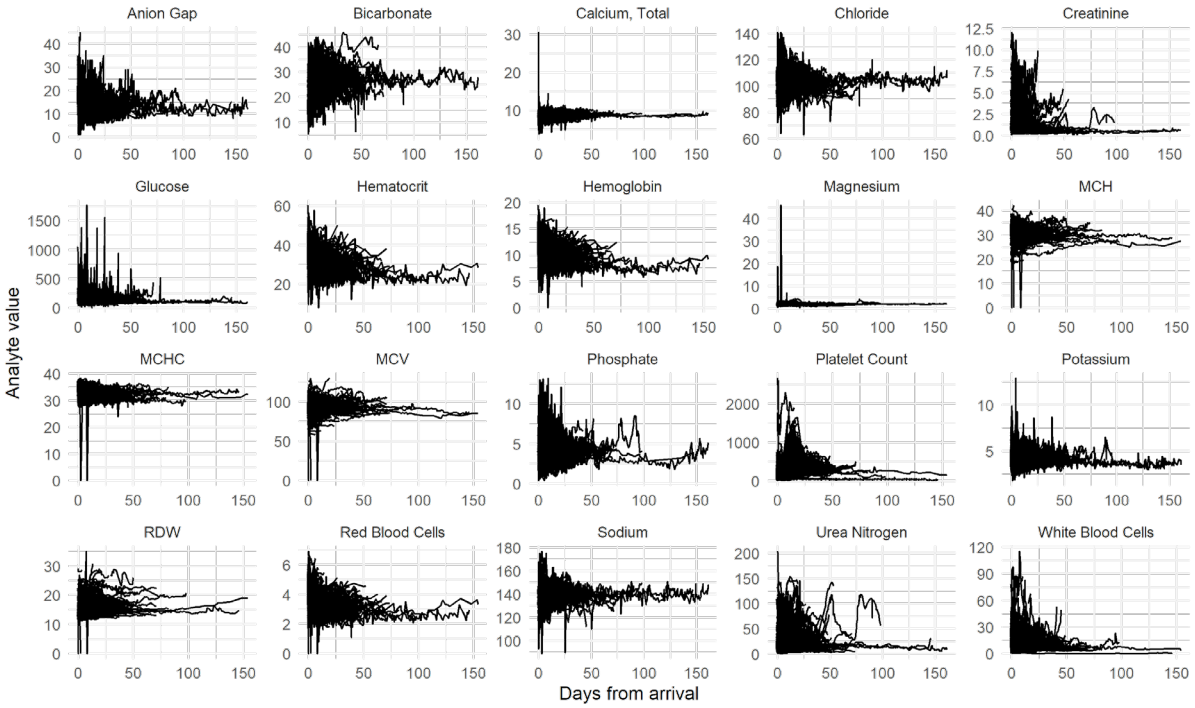


Figure 9. Spaghetti plots for the raw data of the modeling set of analytes (20 most common). Note that unexpected spikes are observed, probably indicative of data errors. Each line represents a subject.

Figure 9 shows spaghetti plots of the analyte modeling set over time from the date of hospital arrival. As expected by the nature of the data, time is asynchronous, meaning that analytes were obtained for each patient at different timepoints in non-regular intervals. No apparent trends are observable from the plots. We can observe fluctuations over time, with, in general, a high dynamic range early after admission that reduces as time progresses (Fig. 9). The number of subjects with data in a given analyte also reduces over time, with very few subjects with data beyond 50 days (Annexed Figure 1). This can also be confirmed by the distribution of length of stay, where the median time is 4.6 days (first quartile: 2.09 and third quartile: 9.07), and 96.4% of the cohort was discharged before 28 days in the hospital (Fig. 10).

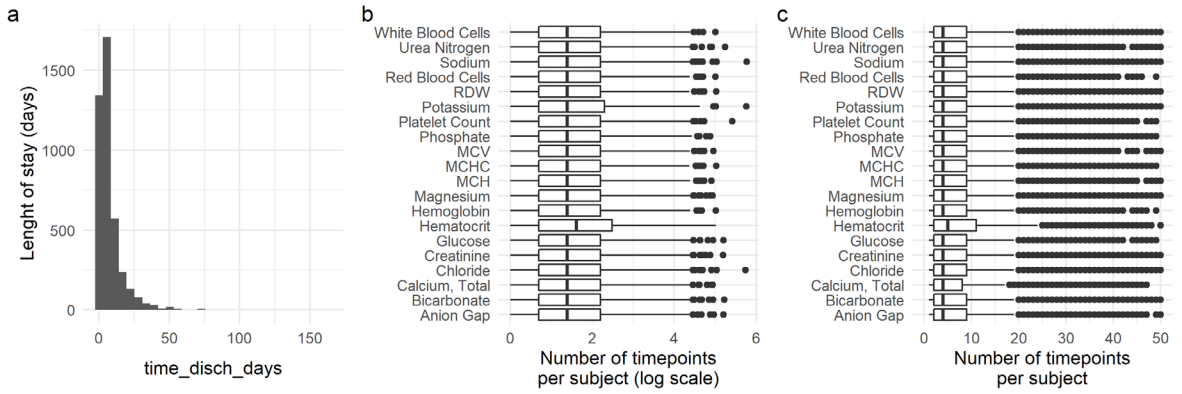


Figure 10. Length of stay distribution. (a) Histogram of the length of stay in days for the cohort. (b) Logarithmic scale count of the number of measurements per subject and the modeling set of analytes. (c) Count of the number of measurements per subject and the modeling set of analytes.

We can also observe subtle spikes in the data from the spaghetti plots, probably caused by errors in the original MIMIC dataset. These spikes create extreme values in the marginal distributions for each analyte in the modeling set (Annexed Figure 2). Using the measurement range values, we can confirm that some of these values are outliers. For example, values of 0 in MCH, as we see in the plot, are not possible (as MCH is the ratio of hemoglobin to red blood cells). Although the number of spikes is small, we performed a data filter to detect and discard those highly likely observations of outliers (see methods), resulting in less extreme values per analyte (Fig. 11 and 12). Finally, given the high drop in the number of patients with data available beyond the first few weeks, we decided to limit the set of analytes for modeling up to 21 days from hospital admission (Fig. 11).

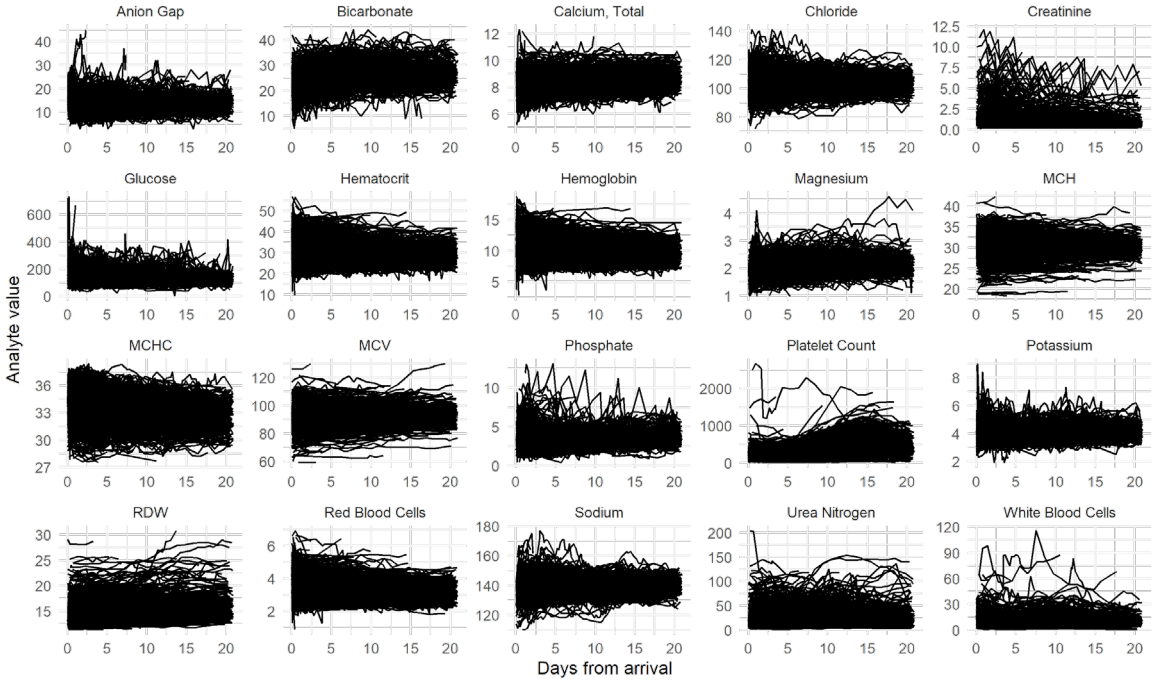


Figure 11. Spaghetti plots for the outlier-cleaned modeling set of laboratory analytes for the first 21 days after admission.

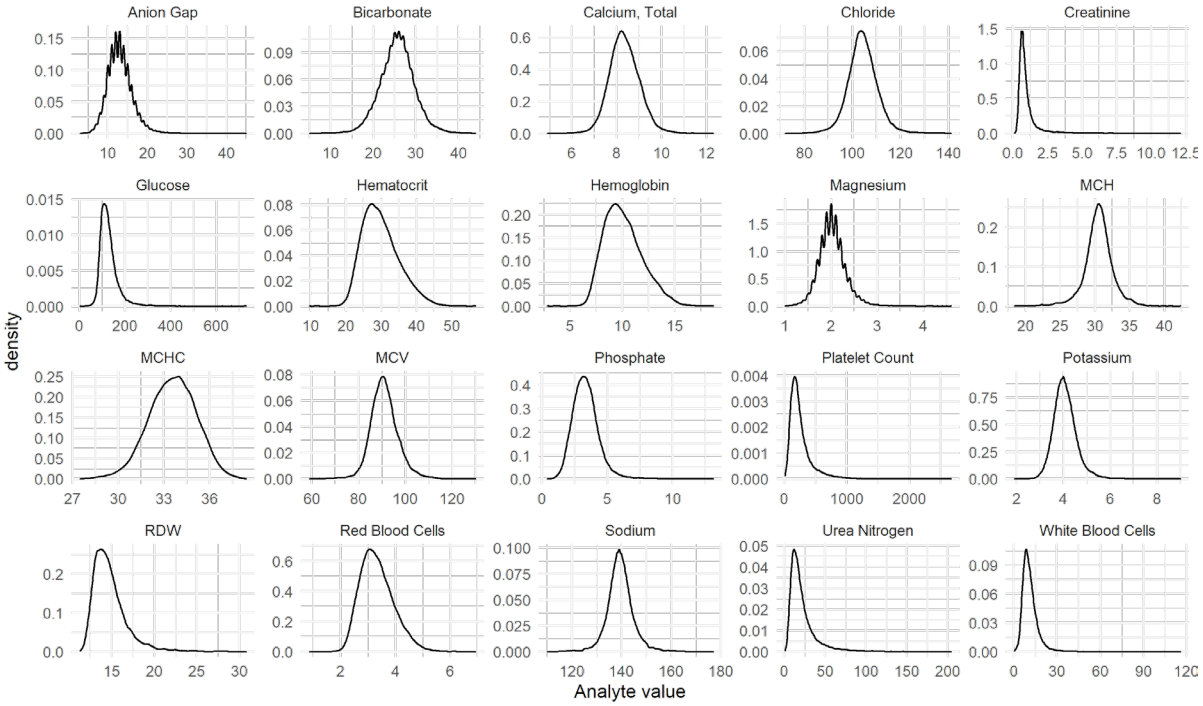


Figure 12. Marginal distributions for the outlier-cleaned modeling set of laboratory analytes for the first 21 days after admission

5.1.3 Cohort characteristics

Table 6 shows demographic variables for the final selected cohort for MIMC datasets. After adjusting p values, we can observe statistical differences in age, gender, insurance type, ethnicity, and dataset (i.e., MIMIC III or IV) across the three cohort groups. The SCI Fracture cohort is younger and with higher proportions of males than the spine trauma patients. Notably, the proportion of males and females in the spine trauma group was almost 1 to 1, while in both SCI groups, the proportion was above 2 males for each female. Higher proportion of patients in the SCI Fracture group came from MIMIC-III database (60%) while in the SCI noFracture and spine trauma groups ~60% of patients came from MIMIC-IV database. This may reflect changes in coding and charting practices from MIMIC-III version to version IV.

Table 6. Demographics for the MIMIC cohorts

Characteristic	**SCI Fracture N = 382[1]**	**SCI noFracture N = 125[1]**	**Spine Trauma N = 2,108[1]**	**p-value[2]**	**q-value[3]**
Age	55 (39, 71)	57 (44, 72)	65 (44, 81)	<0.001	**<0.001**
Gender				<0.001	**<0.001**
F	101 (26%)	37 (30%)	924 (44%)		
M	281 (74%)	88 (70%)	1,184 (56%)		
Insurance				0.011	**0.011**
Medicaid	37 (9.7%)	13 (10%)	154 (7.3%)		
Medicare	118 (31%)	48 (38%)	865 (41%)		
Other	221 (58%)	62 (50%)	1,064 (50%)		
Other Government	6 (1.6%)	2 (1.6%)	25 (1.2%)		
Ethnicity				<0.001	**<0.001**

ASIAN	6 (1.9%)	0 (0%)	43 (2.3%)		
BLACK/AFRICAN AMERICAN	14 (4.4%)	23 (20%)	88 (4.8%)		
HISPANIC/LATINO	14 (4.4%)	9 (7.9%)	77 (4.2%)		
MULTI RACE/ETHNICITY	1 (0.3%)	0 (0%)	5 (0.3%)		
OTHER	17 (5.4%)	5 (4.4%)	84 (4.6%)		
WHITE	264 (84%)	77 (68%)	1,540 (84%)		
Unknown	66	11	271		
Dataset				<0.001	**<0.001**
MIMIC-III	231 (60%)	49 (39%)	826 (39%)		
MIMIC-IV	151 (40%)	76 (61%)	1,282 (61%)		

[1] Median (IQR); n (%)
[2] Kruskal-Wallis rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)
[3] False discovery rate correction for multiple testing

Table 7 shows the hospital stay characteristics for the MIMIC extracted patients. After adjusting p values, we observe statistical differences in length of stay, and discharge location. On average, the SCI Fracture patients stayed in hospital 3 and 4 more days than the spine trauma and SCI noFracture group. Regarding discharge location, SCI Fracture had higher mortality rate (12%), and more proportion of patients were discharged to rehabilitation (54%) than the other two groups. Consequently, we observe a reduction in home/hospice as well as skill nursing facility discharge in these patients.

Table 7. Hospital stay characteristics for the MIMIC cohorts

Characteristic	**SCI Fracture N = 382[1]**	**SCI noFracture N = 125[1]**	**Spine Trauma N = 2,108[1]**	**p-value[2]**	**q-value[3]**
Length of stay (days)	11 (7, 19)	8 (5, 13)	7 (4, 11)	<0.001	**<0.001**
Unknown	1	0	0		
Admission location				0.079	0.10
CLINIC REFERRAL	55 (14%)	22 (18%)	251 (12%)		
EMERGENCY ROOM	301 (79%)	89 (71%)	1,730 (82%)		
TRANSFER FROM HOSP	20 (5.3%)	12 (9.6%)	95 (4.5%)		
TRANSFER FROM SNF	0 (0%)	0 (0%)	3 (0.1%)		
WALK-IN/SELF REFERRAL	4 (1.1%)	2 (1.6%)	20 (1.0%)		
Unknown	2	0	9		
Discharge location				<0.001	**<0.001**
ACUTE HOSPITAL	4 (1.0%)	0 (0%)	9 (0.4%)		
AGAINST ADVICE	0 (0%)	2 (1.6%)	11 (0.5%)		
DIED	46 (12%)	6 (4.9%)	141 (6.9%)		
HOME/HOSPICE	56 (15%)	46 (37%)	687 (34%)		
ICF	1 (0.3%)	0 (0%)	0 (0%)		
LONG TERM CARE	22 (5.8%)	5 (4.1%)	88 (4.3%)		
REHAB	208 (54%)	45 (37%)	528 (26%)		
SHORT TERM CARE	3 (0.8%)	0 (0%)	8 (0.4%)		

SKILLED NURSING FACILITY	35 (9.2%)	17 (14%)	544 (27%)		
TRANSFER TO OTHER	7 (1.8%)	2 (1.6%)	29 (1.4%)		
Unknown	0	2	63		
Number of ICD diagnostics	13 (9, 18)	13 (8, 21)	14 (9, 20)	0.10	0.10

[1] Median (IQR); n (%)
[2] Kruskal-Wallis rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)
[3] False discovery rate correction for multiple testing

5.2 Individual laboratory analyte class trajectory models

5.2.1 Laboratory analyte univariate exploratory trajectory modeling

Researchers have previously recommended a staged workflow because of the computational time it takes to run these models and the difficulty of prioritizing model selection. It starts by running a search of the parameter space using LCGA, discarding the models that are not a good fit, and adjusting other parameters before incorporating random effects. We followed such a strategy, performing a grid search for each one of the modeling analytes for the number of classes (1 to 5), the degree of polynomials (1 to 3), and three distinct link functions: linear or identity for Gaussian continuous responses, and beta and I-spline to model non-Gaussian continuous responses. For model selection, we used BIC, ICL, and APPA (see methods). A total of 900 models were specified. The results of the models for each analyte can be seen in the annexed tables 3 to 22). Annexed figures 3 and 4 show the model selection criteria for the hematology and the chemistry analytes, respectively.

Figure 13 illustrates the results from analytes with different model fit patterns. An interpretation of the model fit selection would go as follow. For example, for red blood cell count, the polynomial degree is the dominant determinant of better model fit by both BIC and ICL. In these cases, we conclude that non-linear trends are a better fit and that the type of link function is almost irrelevant. This informs that assuming a Gaussian response is acceptable for analytes with this pattern of model fit. Regarding the number of classes, for red blood cell count, BIC reduces progressively with the increase of classes, while ICL presents a more prominent drop at two classes but with a more stable value afterward. APPA has a more linear reduction as the number of classes increases. Given that APPA reached acceptable levels in most cases, we used BIC and ICL as the main decision-making tools for this initial step, leaving APPA as a secondary factor. Therefore, we concluded that a linear link and a polynomial of order 3 are best for this variable, selecting between 2 to 4 trajectory classes for modeling red blood cell count. In a different example, we can look at Potassium, where both the polynomial order and the shape of the link function are important for model fit. In this case, we selected models with polynomials of order 3, link function of type beta or spline, and 2 or 3 trajectory classes for further exploration. The results of this decision-making for all variables can be found in the annex. Next, these models will be narrowed down to find the best model given the selected workflow, incorporating random effects into the models.

In summary, all analytes show that at least 2, with a median of 3, trajectory classes are needed to model the data, indicative of heterogeneity in the selected cohort. In addition, most models presented a better fit with a polynomial transformation of degree 2 or 3, indicative that non-linear trends are present in the data. Finally, with the exception of bicarbonate, chloride, hematocrit, hemoglobin, MCHC, and red blood cell count, all other analyte data were better

fitted to models with non-linear link functions (beta or I-spline), which suggests that blood analytes are better modeled by non-Gaussian conditional distributions.

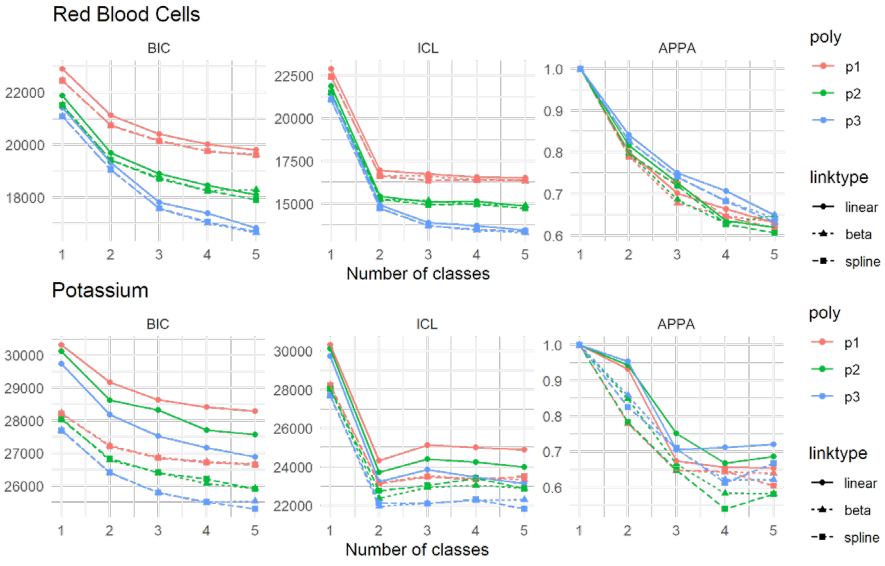


Figure 13. Example of model fit plots for two different types of model selection "patterns". Top row shows the BIC, ICL and APPA (mean APPA across classes) for Red Blood Cells. It can be observed that polynomial degree has a higher effect on BIC and ICL than the type of link function. For Potassium (bottom row), the effect of polynomial degree and type of link function is compounded. The higher drop in ICL for both analytes from 1 to 2 classes suggest that the major gain in model fit happens when more than 1 class is considered, illustrating the need for modeling heterogeneous populations.

5.2.2 Laboratory analyte trajectories

From the previous analysis, we selected those models that showed a better fit based on BIC, ICL, and APPA (Annexed Table 23). There was a clear best choice for some analytes, while the selection was more subjective for others. In those cases, we selected different models for the next step. The LCGA models helped determine the need for non-linear link functions, the maximum number of classes to consider, and whether the linearity of trajectory is a valid assumption. We then when through a similar process but, in this case, using GMM through the specification of the random effects. In addition, instead of polynomial transformations for the time variable, in order to relax some of the polynomial geometrical constrictions, we used a natural spline to model non-linear trends, with degrees determined through the exact search as before.

The final selected model for each analyte can be seen in Table 8. The predicted trajectories for each analyte are shown in figure 14. Creatinine models did not converge even after allowing for a large number of estimation iterations and was not considered any further. With the exception of bicarbonate and calcium, models with more than one class were selected as the best fit. This reaffirms the presence of heterogeneity in the population as seen for LCGA models. Nonetheless, in practice, some models showed a highly unbalanced membership distribution when subjects were classified into classes. Anion gap, chloride, creatinine, magnesium, MCH, MCV, phosphate, platelet count, potassium, sodium, urea nitrogen, and white blood cells selected models presented with a class containing above 90% of the subjects, with often classes close to 1% membership. Contrary, glucose, hematocrit, hemoglobin, MCHC, RDW, and red blood cells presented a more balanced distribution of subjects across classes. Except for glucose, all other analytes were better modeled by non-linear transformations of time, suggesting that trajectories of these analytes do not follow a linear trend. Moreover, 12 of the 20 analytes were better modeled by the use of non-Gaussian link functions.

Table 8. Final selected GMM models

Analyte	k	link	np	d	BIC	ICL	APPA	%class	%class	%class	%class	%class
Anion Gap	2	beta	22	3	104781.7	99681.87	0.98	2.00	98.00	NA	NA	NA
Bicarbonate	1	linear	15	3	111559.0	111559.0	1.00	100.00	NA	NA	NA	NA
Calcium, Total	1	linear	10	2	24868.36	24868.36	1.00	100.00	NA	NA	NA	NA
Chloride	2	beta	22	3	127905.2	122725.3	1.00	0.85	99.15	NA	NA	NA
Glucose	2	beta	11	1	222867.9	218145.3	0.91	10.46	89.54	NA	NA	NA
Hematocrit	3	linear	25	3	138572.1	134284.5	0.82	31.93	13.51	54.56	NA	NA
Hemoglobin	3	linear	25	3	64672.17	60404.33	0.82	26.74	11.37	61.89	NA	NA
Magnesium	2	spline	23	3	-4861.69	-9859.52	0.97	1.95	98.05	NA	NA	NA
MCH	3	linear	18	2	46805.77	41775.40	0.97	94.97	1.84	3.19	NA	NA
MCHC	2	linear	14	2	59109.67	54145.65	0.95	96.54	3.46	NA	NA	NA
MCV	2	beta	16	2	97687.94	92518.59	0.99	98.92	1.08	NA	NA	NA
Phosphate	2	beta	16	2	50457.54	45400.07	0.99	2.74	97.26	NA	NA	NA
Platelet Count	2	beta	22	3	248904.5	246232.0	0.51	35.65	64.35	NA	NA	NA
Potassium	2	beta	22	3	23972.63	18936.72	0.97	3.11	96.89	NA	NA	NA
RDW	3	beta	20	2	34238.71	29400.81	0.93	8.99	87.09	3.92	NA	NA
Red Blood Cells	3	linear	25	3	13818.27	9586.07	0.81	18.94	10.95	70.11	NA	NA
Sodium	2	beta	16	2	129149.1	124014.2	0.99	97.35	2.65	NA	NA	NA
Urea Nitrogen	2	beta	16	2	146168.7	141216.9	0.95	7.65	92.35	NA	NA	NA
White Blood	3	beta	27	3	109955.1	104834.1	0.98	0.85	98.08	1.08	NA	NA

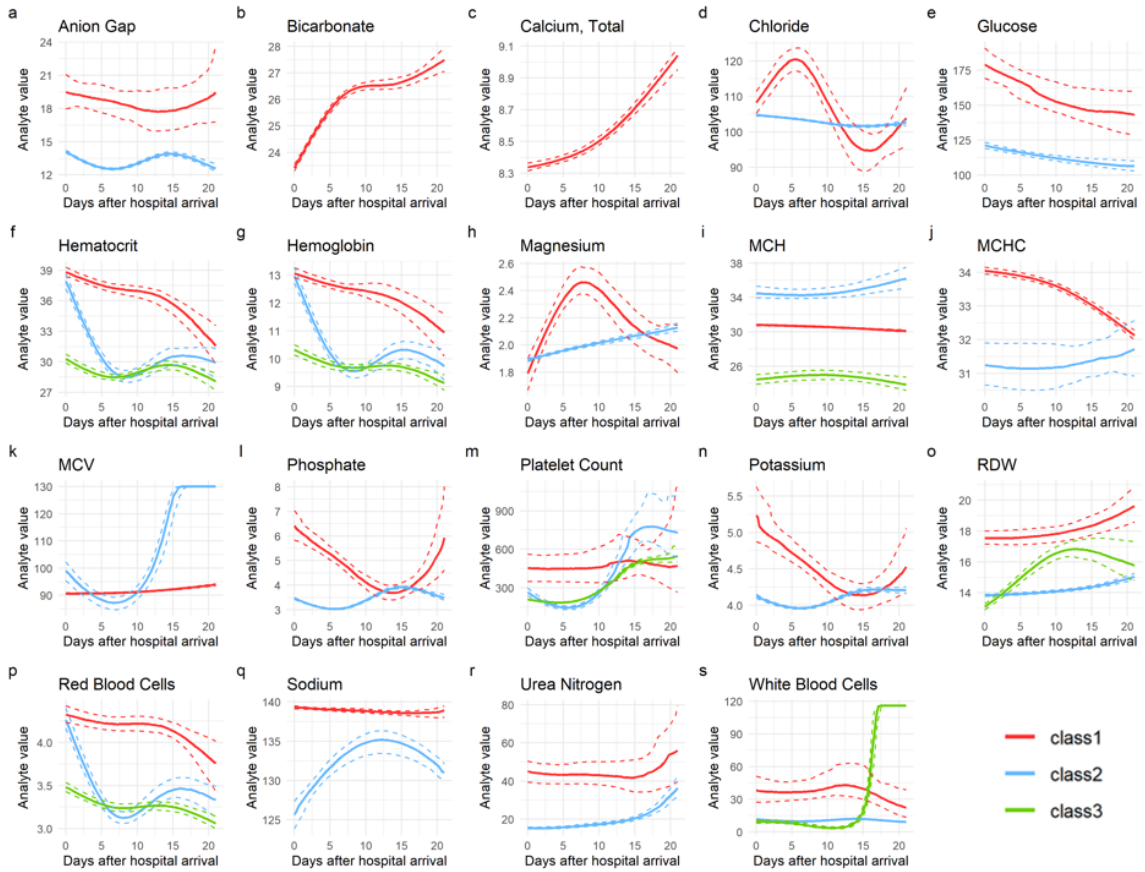


Figure 14. Predicted mean trajectories per analyte and class for the selected models. The mean trajectory for each one the classes for each analyte model is shown, together with the 95% CI. The differences in 95%CI width are explained by the huge differences in sample size between classes. Classes with more subjects assigned to it presents with tider CI ribbons. Note that although colored the same for visualization, these are univariate models and therefore classes might not be constituted by the same subjects across analytes.

5.2.3. Patient trajectory characteristics

We performed a univariate analysis comparing class trajectories for each analyte for different demographics and clinical characteristics. These analyses can be found in the annex (Annexed Tables 24 to 32). A general pattern that emerges from the analysis is significant differences in age, the proportion of patients dead during hospitalization, and the number of diagnostics for most chemical analytes. The class trajectories with lower subject numbers in anion gap, chloride, glucose, magnesium, phosphate, potassium, sodium, and blood urea nitrogen are associated with higher in-hospital mortality rates and a higher number of diagnostics. Except for sodium, those smaller class trajectories are generally characterized by higher values of these analytes early after hospital arrival, with major non-linear dynamic changes over the first three weeks compared to the class with the most patients (Fig. 14). For example, in the minority class of patients on Potassium trajectory with a higher mortality rate (potassium class 1), the analyte levels of those patients are higher early after hospitalization, with a rapid drop after toward levels of class 2. Of the chemical analytes, only glucose and sodium showed differences in the proportion of males and females between trajectory classes.

In the case of hematology analytes, hematocrit, hemoglobin, MCV, RDW, red blood cell counts, and white blood cell counts presented different in-hospital mortality rates across

class trajectories and the number of diagnostics. Class 1 for hematocrit, hemoglobin, and red blood cell counts showing higher initial analyte values with a posterior drop was associated with a lower mortality rate and the number of diagnostics than classes 2 and 3. Length of stay differences can be found between classes of glucose, hematocrit, hemoglobin, platelet count, and red blood cells. Finally, the proportion of patients in each cohort group (SCI Fracture, SCI noFracture, and Spine Trauma) varied across classes for glucose, hematocrit, hemoglobin, platelet count, red blood cell counts, and white blood cell counts.

5.3 Multi-trajectory modeling

5.3.1. Selected GBMT model

The previous models were univariate; we independently derived class trajectories for each analyte. However, blood analytes are interrelated, and distinct patterns may emerge when the different analytes are considered together. To do so, we fitted group-based multi-trajectory models (GBMT). Using the information gained in the univariate modeling exercise, we focused on GBMT to determine groups of subjects that follow the same trajectories across hematocrit, glucose, white blood cells, RDW, and platelet count. A model search was performed as before, with the results of the model selection process shown in the Annexed Table 33 and Figure 15. A final model of 3 groups of trajectories with a third-order polynomial was selected based on model fit metrics. The predicted mean trajectories with 95%CI for the selected model can be seen in figure 16. Class 1 compressed most of the subjects (90.55%) while the other classes were tiny in size (6.62% for class 2 and 2.83% for class 3). All considered polynomial models presented a similar distribution, with a single class collecting the vast majority of the subjects. This differed in the models with degree 1 (linear trend), with better distribution across at least two classes. Nonetheless, the linear trend models had a worse fit than the polynomial models.

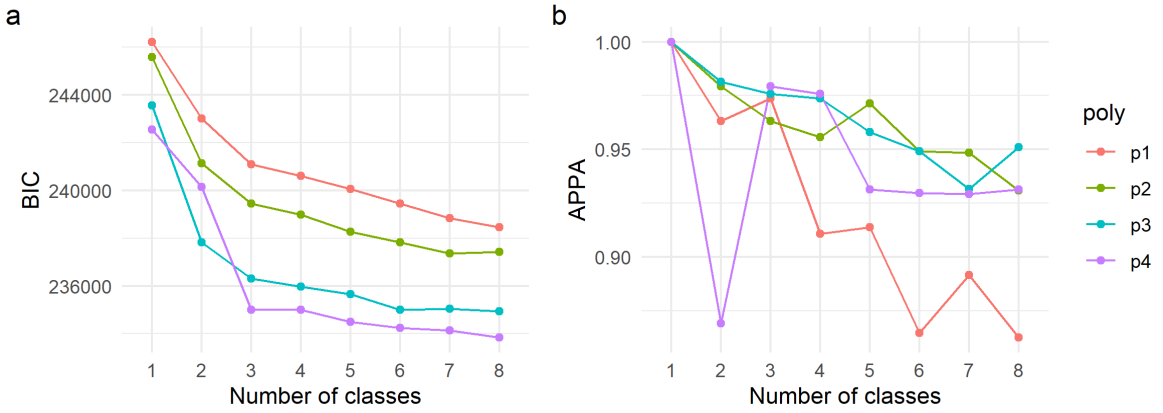


Figure 15. Model fit metrics for GBMT. GBMT models were fitted and BIC and APPA used as model selection criteria.

5.3.2. Multi-analyte trajectory characteristics

The three selected multi-analyte trajectory classes are very similar in their trajectories, with a few exceptions (Fig. 16). The three classes present similar temporal evolutions on glucose and white blood cell counts. Regarding hematocrit, the major differences are between classes two and three. There is a noticeable initial drop and slight recovery for class two, while class three is constantly low. The major differences between classes are perhaps for RDW, where for class one there is a constant linear increase, for class two a linear increase with lower initial values and higher change rate than class one, and for class three a rapid increase with a pick at day five and a posterior progressive decrease and plateau. This suggests that RDW may be the primary driver of the differential multi-analyte trajectory class definition.

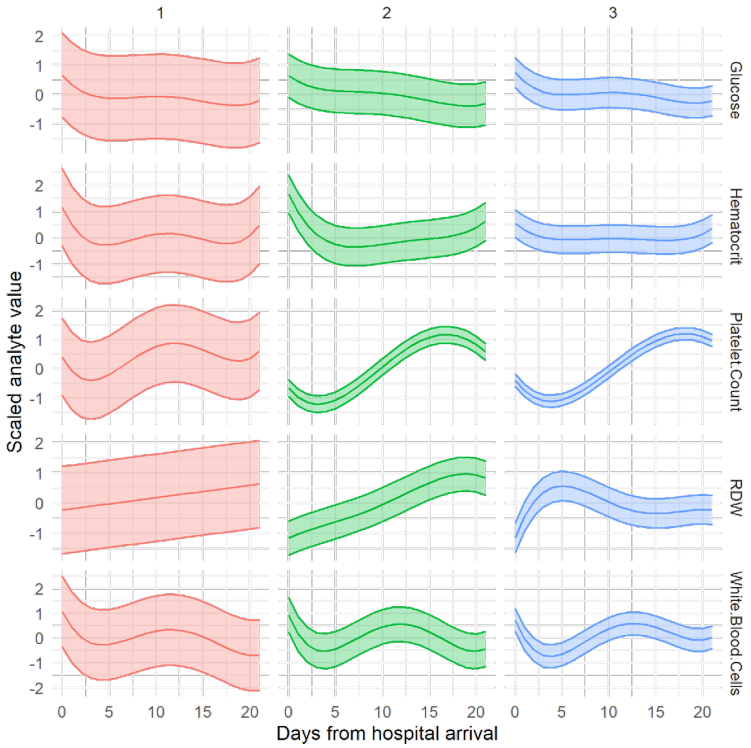


Figure 16. Predicted mean trajectories for GBMT selected model. The mean trajectory and 95% CI for the 4 class GBMT selected model is shown. Given these trajectories where jointly modeled, each group (1 to 4) is constituted by the same subjects across the 5 considered analytes.

Table 9 shows the univariate analysis of different patient characteristics between the assigned classes. Age, gender, cohort type, length of stay, and the number of diagnostics show significant differences between classes. On average, class one consisted of older patients, a higher proportion of females, a higher proportion of spine trauma patients, a lower number of days in the hospital, and a lower number of diagnostics than classes two and three. Class two and three were very similar in all studied univariate associations, with a slightly higher proportion of SCI Fracture patients in class three than in class 2. This may point out that these two minority classes are unnecessary partitions or that the underlying blood trajectory differences cannot be associated with the studied patient characteristics.

Table 9. Multi-trajectory class univariate analysis

Characteristic	Class 1 N = 2368	Class 2 N = 173	Class 3 N = 74	p-value	q-value
Age	63 (44, 80)	54 (32, 73)	54 (44, 70)	<0.001	**<0.001**
Gender				<0.001	**<0.001**
F	990 (42%)	49 (28%)	23 (31%)		
M	1,378 (58%)	124 (72%)	51 (69%)		
Ethnicity				0.8	0.8
ASIAN	44 (2.1%)	2 (1.5%)	3 (4.5%)		
BLACK AFRICAN AMERICAN	115 (5.6%)	7 (5.2%)	3 (4.5%)		
HISPANIC LATINO	90 (4.4%)	7 (5.2%)	3 (4.5%)		
MULTI RACE ETHNICITY	5 (0.2%)	1 (0.7%)	0 (0%)		
OTHER	95 (4.6%)	7 (5.2%)	4 (6.1%)		
WHITE	1,718 (83%)	110 (82%)	53 (80%)		
Unknown	301	39	8		
Cohort				<0.001	**<0.001**
SCI Fracture	307 (13%)	49 (28%)	26 (35%)		
SCI noFracture	117 (4.9%)	4 (2.3%)	4 (5.4%)		
Spine Trauma	1,944 (82%)	120 (69%)	44 (59%)		
Length of stay (days)	7 (4, 10)	24 (19, 31)	26 (22, 33)	<0.001	**<0.001**
Unknown	1	0	0		
Died in hospital	179 (7.6%)	8 (4.6%)	2 (2.7%)	0.15	0.2
Number of diagnostics	13 (9, 19)	20 (14, 30)	17 (11, 25)	<0.001	**<0.001**

5.4 Dynamic prediction modeling

We set three dynamic prediction modeling experiments (Fig. 17). Since the GBMT model showed a single class constituted by more than 90% of the subjects, we performed these experiments using the univariate trajectory models per each analyte. Experiment I was designed to test whether blood analyte trajectories can be used to predict in-hospital mortality for the spine trauma and SCI population. Experiment II aimed to study whether blood analyte trajectories can predict if a patient with a spine trauma also presented an SCI. Experiment III used the TRACK-SCI data as an external cohort of SCI patients for which analyte trajectories are predicted. These predictions were used to build a prognostication model for SCI severity (motor complete vs. motor incomplete). We run each experiment at different cutoff days from hospital arrival to simulate dynamic predictions as data get available. The probability of class membership (i.e., posterior probabilities) per analyte was calculated for each patient, and those probabilities were used as predictors in an ElasticNet model. For Experiments I and II, the model with 21 days of data can be considered the model with the best prediction for class membership as trajectories were derived from data with up to 21 days of analyte values.

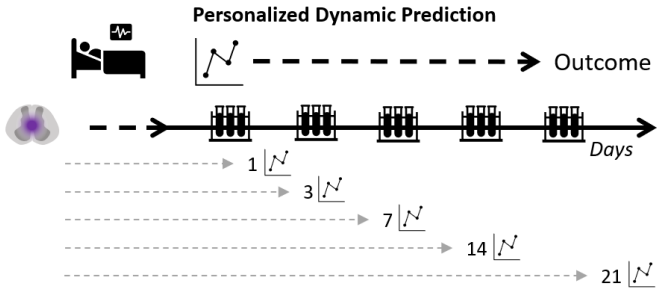


Figure 17. Dynamic prediction experiments. The experiments simulate a real-time dynamic prediction of outcomes where the predicted blood analyte trajectory is calculated with increased available laboratory data from EHR. To simulate this process, we build models with cutoff days of blood data up to 1, 3, 7, 14 or 21 days after hospital arrival.

5.4.1. Experiment I. Predicting death

Experiment I results are shown in figure 18. All models performed well with high AUC on the in-train sample for all the cutoff days, including when the model was trained with up to 1 day of laboratory data. The model reduced performance with the out-train sample (i.e., test sample) for all cutoff days. Nonetheless, the AUC for all out-train samples was considerably high, even for predicted trajectory membership using data for up to the first day after hospital arrival. This highly suggests that it is possible to predict whether a spine trauma or SCI patient will die in the hospital by predicting their blood analyte trajectories as soon as 1 day after arriving at the hospital, and that prediction improves as more data becomes available.

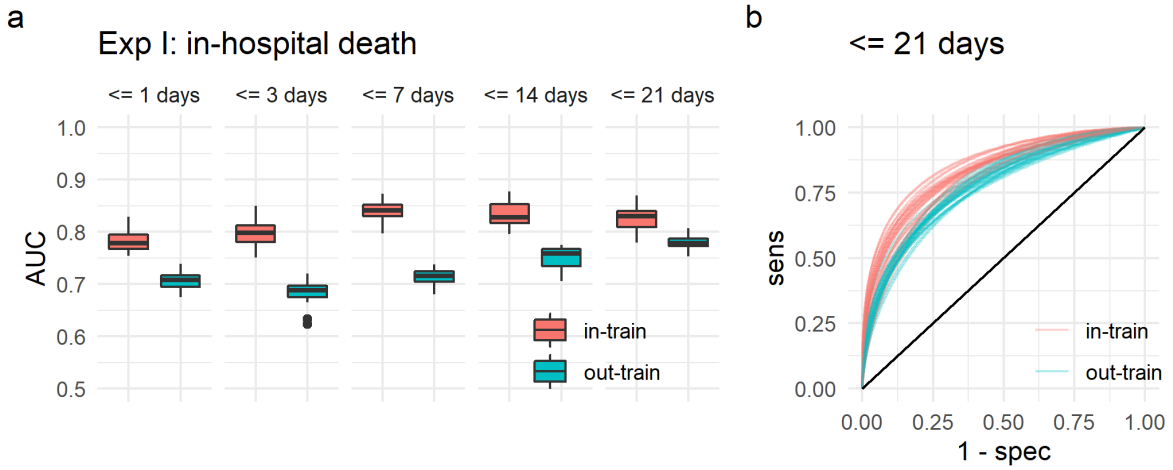


Figure 18. Results of dynamic prediction modeling Experiment I. The target of this prediction task was whether patient died in-hospital. (**a**) boxplots of AUC for each one of the cutoffs of the dynamic modeling simulation. (**b**) ROC curve for the experiment considering all available data up to 21 days from hospital arrival. Each prediction scenario was repeated 25 times with a different random seed. The random seed was maintained to be the same across the cutoff.

5.4.2. Experiment II. Predicting SCI

Experiment II results can be seen in figure 19. The performance of all models to predict whether a patient had an SCI measured by AUC was high for the in-train sample for each one of the cutoff dynamic models. As early as day 1, we can observe a good prediction using the same data to train the ElasticNet models. The prediction performance for the out-train sample

reduces considerable, including for the 21 days models, with an AUC ~ 0.6. The drop in performance clearly illustrates the difficulty to generalize the ElasticNet models in this scenario for the detection of whether a patient with a spine trauma presents with SCI. Nonetheless, it is encouraging that it is possible to train models to detect the presence of SCI better than random with only blood trajectory probabilities.

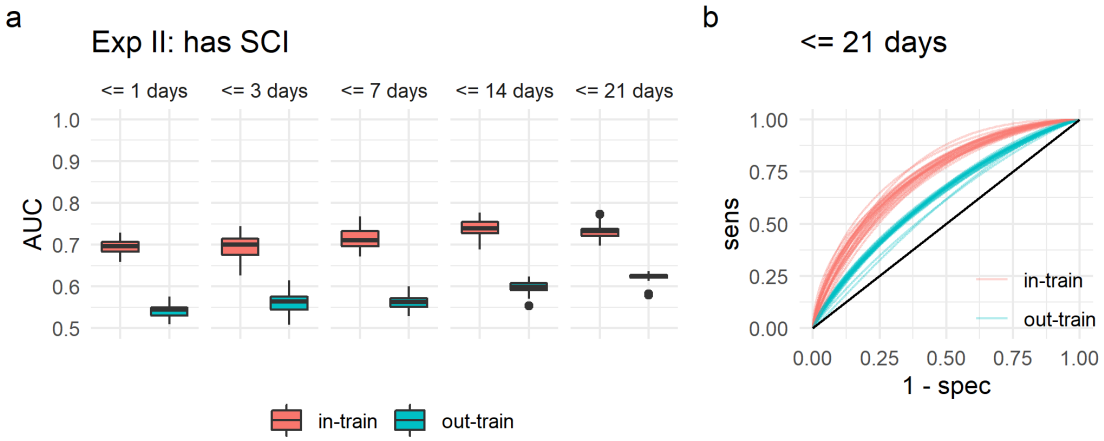


Figure 19. Results of dynamic prediction modeling Experiment II. The target of this prediction task was whether patients with spine trauma had an SCI. (**a**) boxplots of AUC for each one of the cutoffs of the dynamic modeling simulation. (**b**) ROC curve for the experiment considering all available data up to 21 days from hospital arrival. Each prediction scenario was repeated 25 times with a different random seed. The random seed was maintained to be the same across the cutoff.

5.4.3. Experiment III. Predicting SCI severity in an external cohort

Experiment III was set to study whether the prediction of trajectory classes from patients external to the trajectory modeling can have predictive power for SCI severity. Figure 20 shows the spaghetti plots for the extracted lab analytes from the TRACK-SCI cohort after processing the data in the same way we did for the MIMIC values. Those where then used in our dynamic prediction experiment workflow (Fig. 17). Given the sample size was small (n = 137), instead of having an out-of-train sample for estimating prediction generalizability, we used LOOCV. The dynamic prediction task for this experiment was set to classify whether SCI patients in the TRACK-SCI cohort presented an AIS grade of A or B vs. C, D or E in its latest available in-hospital timepoint. Effectively, this means we trained a model to distinguish on whether patients presented motor complete (AIS A or B) vs. motor incomplete (AIS C, D or E) neurological deficits. Figure 21 show the results of Experiment III. Similar to Experiments I and III, AUC for in-train sample was considerably higher than LOOCV, reaching values as high as 0.87. With one day of data, AUC for LOOCV was 0.66, reaching 0.72 when 21 days of data were used. This indicates that the prediction models were overfitted.

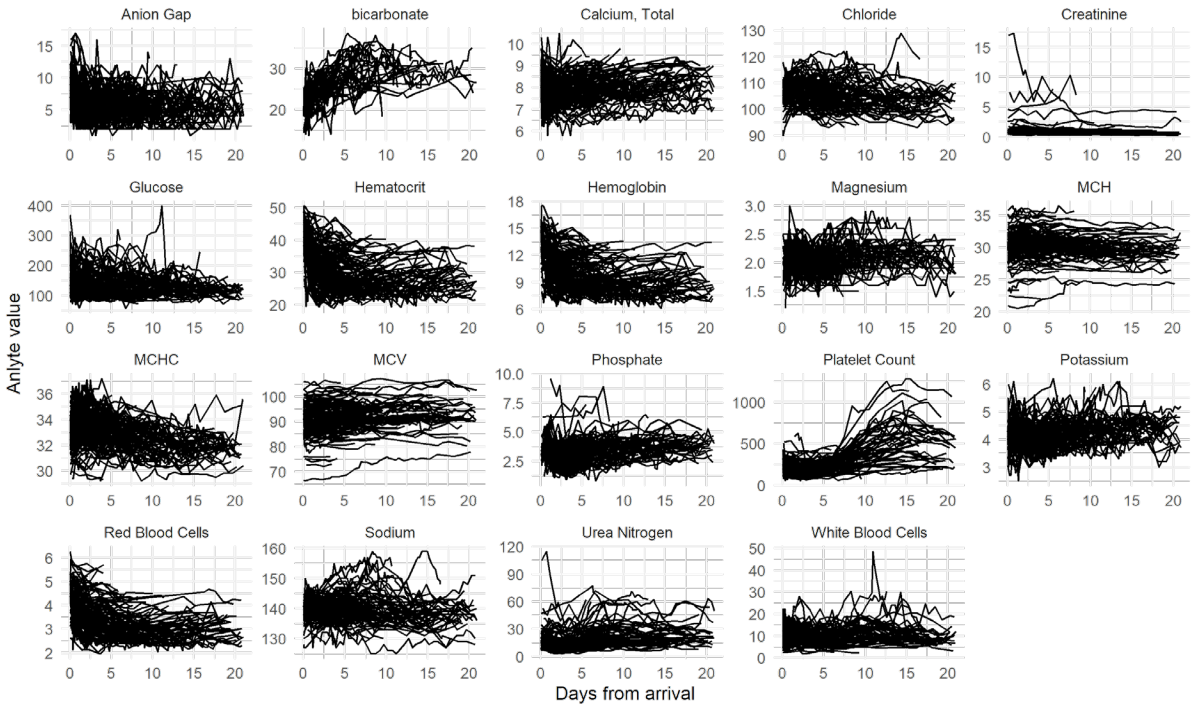


Figure 20. Spaghetti plots for the modeling set of laboratory analytes in the TRACK-SCI cohort

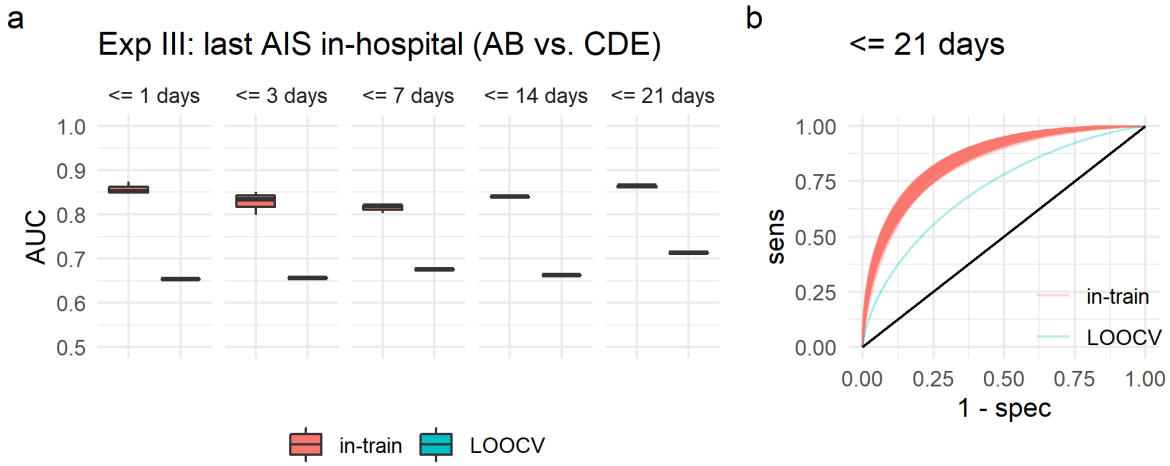


Figure 21. Results of dynamic prediction modeling Experiment III. The target of this prediction task was whether latest AIS grade in-hospital was AB vs. CDE (i.e., motor complete vs. motor incomplete). (**a**) boxplots of AUC for each one of the cutoffs of the dynamic modeling simulation. (**b**) ROC curve for the experiment considering all available data up to 21 days from hospital arrival. Each prediction scenario was repeated 25 times with a different random seed. The random seed was maintained to be the same across the cutoff.

6 Discussion

This work has studied the temporal dynamics of blood analytes early on after spinal cord injury through trajectory modeling. We have demonstrated distinct temporal patterns of blood analytes after SCI, that those follow non-linear trends for the most part, and that there is heterogeneity in spine trauma patients that can be observed from a data-driven analytical framework. Importantly, we demonstrate that routinely collected in-hospital data have enough

signal for patient stratification and subtyping, increasing interest in using blood analytes as predictor biomarkers for neurological conditions, including SCI. This also demonstrates the potential of using data collected in real-world scenarios to research SCI and presumably other neurotraumatic conditions. Finally, we provide evidence for the need to model these metrics through a more complex set of tools that can account for heterogeneity and non-linearity and accommodate the non-Gaussian distribution nature of some of these metrics.

6.1. Modeling heterogeneous blood trajectories after acute injury

Previous work has shown that blood analyte changes can be associated with SCI, and this association can be used in predictive models as biomarkers (Gurcay et al., 2009; Kyritsis et al., 2021; Leister et al., 2021). The premise is that the pathophysiological events triggered by the injury to the spinal cord can be proxied in the blood through complex signal integration pathways (Jogia, Kopp, et al., 2021; Kyritsis et al., 2021). We observed distinctive non-linear trajectories of blood analytes early after SCI or spine trauma. After a thorough examination of modeling parameters, we show that non-Gaussian link functions are better for modeling the continuous random deviations of some analytes. In contrast, other analytes are well captured under Gaussian assumptions. Linear methods such as linear mixed models assume Gaussian distributions for the random effects and error, which is not always true when modeling longitudinal data. Although simulation studies show that LMMs are robust to deviations from the Gaussian assumption, it can bias the estimation in the case of missing temporal data (Lu et al., 2009). Since the asynchrony nature of real-world blood laboratory work, it is granted to observe several patterns of temporal missingness. Therefore, to better model potential non-Gaussian deviations, we used latent processes modeled through a link function between the observed data and the linear model (Proust-Lima et al., 2017).

The obtained trajectories represent distinctive demographics and clinical characteristics. The chemical blood analytes are major electrolytes that regulate normal physiological function, and their dysregulation can signal several pathophysiological events (Balcı et al., 2013). Two distinctive trajectories emerged for each electrolyte, with a low membership class presenting a higher in-hospital mortality rate and characterized by non-linear changes with high dynamic range and values above normal. Although these trajectory classes are unrelated univariate, these were all also associated with a higher average number of clinical diagnostics, a surrogate measure for comorbidities and medical events. It could indicate distinctive trajectories in severe patients with a higher likelihood of dying at the hospital than patients with less physiological distress or a capacity to regulate homeostasis. This is perhaps not surprising as blood analytes are considered to calculate measures of overall patient severity scores (Bouch & Thompson, 2008). Thus, the finite models mainly capture the heterogeneity of patient severity.

Regarding the hematology values, hematocrit, hemoglobin, and red blood cells are highly correlated. As expected, the three trajectories found for each analyte are very similar in their temporal trend. Class trajectory 1, characterized by high initial values and a slow steady reduction, presented the lower proportion of patients with SCI and spine fracture, and the lower mortality rate. Class 2 evolved from high levels to rapid decay by day 7. Class 3 sustained low values along the 21 days. The class 2 and 3 trajectories in hematocrit, hemoglobin, and red blood cells presented a higher mortality rate and proportion of SCI patients with spine fracture than class 1. The major difference between class 2 and 3 in these three analytes was in demographics, where class 2 was younger and with more representation of males than class 3. All levels of hematocrit and red blood cell counts were below considered

normal levels in the general population, an effect previously described in acute SCI (Brown et al., 2020).

In general, white blood cell values were above normal, which has been described (Brown et al., 2020). It is expected given the inflammatory process and immune dysregulation occurring after SC (Jogia, Lübstorf, et al., 2021). Class 2 had sustained levels at the high limit of a normal range. This class was constituted by the majority of the patients and presented a lower mortality rate than class 1 and class 3, and with a lower proportion of SCI Fracture patients. Class 1 was older, with a high mortality rate and higher number of diagnostics than class 3. After SCI, distinct populations of white blood cells dynamically change their blood presence over time following pathophysiological processes triggered by the injury (Jogia, Lübstorf, et al., 2021). It is possible then that a global count of white blood cells lack the resolution to describe the temporal patterns and that individual levels of different cell types should be adequate to describe heterogeneity. Future work should address this oversight.

6.2. Blood analyte trajectories as biomarkers for SCI

Predicting patient outcomes after SCI using routinely collected information during early hospitalization can have high utility for patient management, prognostication, and clinical research. Previous work has shown the association of blood markers to distinct patient characteristics and outcomes after SCI (Brown et al., 2020; Gurcay et al., 2009; Jogia, Kopp, et al., 2021; Kyritsis et al., 2021; Leister et al., 2021). Here we extended this line of work by incorporating the class trajectory of each biomarker as predictors in a dynamic process. Using the posterior probability of the modeled class trajectories, we trained classification models with decent performance to detect whether patients will die during the hospital stay as early as one day after hospital arrival. Prediction performance increased in the out-of-train sample as more temporal data was considered for predicting trajectories. This suggests that predicting which trajectory of blood analytes patients will follow can serve as predictors in dynamic prognostication models of whether a patient will die.

In the case of predicting whether a patient has an SCI or not after spine trauma, the performance was moderate at best and worse than in the case of predicting patient death. This could be explained by the fact that spine trauma is a significant traumatic event, usually companioned by polytraumatic processes (e.g., other broken bones) and damage to parts of the body, which probably triggers several pathophysiological changes. The addition of injury to the cord may be minor in the overall early pathology after trauma. Another complication is that SCI is a rare event compared with all patients with traumatic spine damage. In our work, we used spine trauma patients with no finding of SCI as the closest trauma “control” for patients with traumatic SCI. Our search resulted in ~15% of the total spine trauma patients having a diagnosis for SCI in the MIMIC dataset. This imbalance also increases the challenge of building classification models. Nonetheless, our approach shows the potential to detect SCI in patients with spine trauma. The addition of other predictors such as demographics would potentially increase classification performance. As in the case of detecting mortality, the prediction performance increased as more temporal data was included. This effect is somewhat expected as the inclusion of more data should reduce the uncertainty of trajectory membership classification. Overall, the results suggest that predicting the blood analyte trajectory in spine trauma patients has biomarker utility for early SCI diagnostic.

We also investigated whether predicting analyte trajectory in a cohort of SCI patients that was not used for modeling (TRACK-SCI patients) could be used to prognosticate the

severity of the neurological deficit. The results are promising in two folds. On one side, we demonstrate that early prediction of analyte trajectory in an external cohort has prediction utility. This is important as it signifies that we can derive trajectory models from real-world data generalizable to data collected in a different context (i.e., another hospital). On the other side, we show that prognostication of the degree of neurological deficits is possible by just using laboratory data. In addition, the prediction improves as more data is included in determining patient trajectory.

The dynamic prediction simulation is encouraging. By modeling the trajectory of analytes, we can determine the probability of following a given trajectory as early as one day after hospital arrival, which provides information into the predicted "future" of the pathophysiological events by forecasting a patient's changes in blood values. Using the trajectory membership as a latent feature, we can then build predictive models for patient diagnostic and prognostication. This indicates that real-time prediction models using available in-hospital data in clinical support systems that help clinicians guide SCI patient management are possible. Future work should be oriented to expand on the possible implementation of these models.

6.3. The use of real-world data to perform research in SCI

Performing research in clinical SCI is challenging due to the relatively low incidence with respect to other medical issues and the complex heterogeneity of the population. Studies usually use high constraints in the inclusion and exclusion criteria to deal with this heterogeneity, focusing on a very narrow segment of the population, resulting in a reduction in sample size. This is illustrated by the median number of subjects in clinical SCI studies is 32 (calculated from clinicaltrials.gov). An alternative framework is to embrace heterogeneity and use it to the research advantage through data-driven methodologies that can discover unseen or unexpected patterns of associations, hoping that those are important for the question at hand. In this work, we used this research framework through a discovery and modeling process of routinely collected data present in electronic health records to then apply those models to a cohort of interest. Data in real-world scenarios are messy (Chan et al., 2010; Cowie et al., 2017; Suvarna, 2018), often with many missing data, uncontrolled confounding mechanisms, and challenging to explain associations. On the other hand, they offer volume and variety, two common characteristics of big-data (Ohmann et al., 2017; Peek et al., 2014). Since low incidence and heterogeneity are distinctive features of SCI, real-world data offers a different venue for research and discovery, with the ultimate goal of producing knowledge that can drive clinical decisions and patient improvement.

One of the strengths of the methodology used in this work is that it allows for flexible modeling of the complex temporal patterns after injury. Using longitudinal finite mixture models, we provide evidence that there are different temporal pattern trajectories after spine trauma and that those are, for the most part, non-linear. In addition, we demonstrate how to accommodate non-Gaussian response distributions through non-linear link functions of a latent process mixed model. Previous work on SCI has ignored these factors, which may affect the utility of the models. The drawback is that model estimation becomes computationally expensive, as the number of parameters to estimate grows considerably when accommodating all the characteristic features of blood analyte trajectories. This model complexity also requires a big sample size, which, as discussed above, can be an issue in SCI research. Nonetheless, we demonstrated that model development using electronic health records is feasible and that these models can be generalized to external cohorts with utility for

predicting tasks. This should open new exciting venues for SCI research since the knowledge gained from electronic medical records is transferable.

7 Conclusions

7.1 Conclusions

We sought to study dynamic trends of blood analytes in SCI as potential biomarkers in this work. We demonstrated the utility of modeling heterogeneous temporal trends of blood analytes collected in real-world scenarios to predict diverse diagnostic and prognostic tasks in spine trauma and SCI patients. Longitudinal finite mixture models are powerful tools to describe multi-class trajectories of blood analytes, potentially capturing latent pathophysiological events. Given that SCI and other neurological pathologies evolve, studying the non-linear dynamic changes of any biomarker level is more valuable than considering predictions with a static cut-off level. We also simulated a dynamic prediction process to perform prediction tasks as data get available in the hospital settings, demonstrating the potential utility of this approach to build clinical support systems assisting in patient management. Furthermore, the knowledge gained from real-world data is transferable to cohorts of interest, which suggest that this data can be used for SCI research and model development. The models and methodology used in this work can be extended to other types of data and pathologies. Future research should be oriented to improve upon the models and the limitations of the present work. We hope that by offering accessibility to the code and trained models, the field of SCI research can expand on this work toward developing multi-analyte trajectory biomarker models.

7.2 Limitations and future work

The present work is a proof-of-concept for multi-trajectory blood analyte biomarker discovery. As such, there are several limitations, and much more can be done and expanded. For convenience, we performed a shallow cohort selection through ICD diagnostic codes of EHR data, which might have introduced extreme heterogeneity in the cohort and potentially affected the models' capacity to split heterogeneities into smaller cohort subtypes. Future work should address this by a more thorough examination of inclusion and exclusion criteria and by using better characterized SCI patient cohorts for both modeling and prediction experiments. Also, for convenience, the search in the parameter space for the finite mixed models was limited, and for some analytes, we might not have resolved the best model possible. Future work should consider other parametrizations of the link functions and higher orders for the non-linear transformations. Moreover, we focused on the most common blood analytes, which may or may not be the most useful predictors. For instance, differential blood white cell counts that include different types of leucocytes show distinctive associations in SCI, and their trajectories are potentially helpful for prediction in this population (Jogia, Lübstorf, et al., 2021). Given the drop in sample size, we did not include those in this work; however, the current approach can be expanded to include these and other analytes. Furthermore, here we focused on the prediction task. However, we hypothesize that the derived trajectories are proxies for pathophysiological events. Therefore, a deeper examination of the clinical characteristics of each trajectory subpopulation for better patient phenotype understanding is granted. This better understanding can also derive future development of prediction modeling by pointing to

stratifying factors serving as other early predictors of SCI, potentially improving prediction performance. In addition, we limited the analysis to 21 days as the sample size quickly dropped after that. This is sufficient for our selected cohort; however, a subset of patients stays in the hospital for longer. Given that length of hospital stay in SCI is related to patient severity, other trajectories may emerge when more extended periods are considered. Finally, we limited the prediction task to a subset, but many other tasks can be explored.

As a proof-of-concept, we hope that his work sets the bases for further development of dynamic biomarkers in neurotrauma and other neurological conditions.

7.3 Plan following

The initial plan was ambitious and was set under some uncertainty since the data quality was not fully known in advance. The work progressed following the created plan, although some changes were made over time. One of the significant challenges during modeling was the computational cost of some of the models, taking hours to days to finalize the process, even after parallelizing the computation. Those periods were used to advance the writing of this manuscript and deepen into the mathematics behind the employed models. Overall, the progress was good and on time.

8 Glossary

Spinal cord injury (SCI). Injury affecting the spinal cord

Real-world data. Data is derived from real-world scenarios during everyday activities such as regular medical practice in contraposition to data collected as part of a design study. Electronic Health Records for research are an example of real-world data.

Electronic Health Records (EHR). Records are generated during routine medical practice in electronic format (i.e., digital) to contain individuals' medical and health information.

Big-data. Data is big-data when it presents in high volume, velocity, and variety. Working with datasets with these characteristics has specific challenges that require tailored approaches. This has made big-data a phrase to refer to datasets and the methodologies dedicated to mining big datasets.

Longitudinal data. Data collected over time is said to be longitudinal. It is characterized by repeated measures of the same variables over different periods per subject. Although this definition could also include data collected in a continuous high sampling process, usually those are distinguished from longitudinal data by referring to them as time-series and functional data.

Finite Mixture Model (FMM). Statistical probabilistic models in which a finite number of distributions are defined in combination (i.e., mixture). These models can approximate complex distributions through a mixture of well-known distributions such as Gaussian and determine latent groups or classes by estimating unseen homogeneous groups of observations.

Trajectory. Temporal trend describes the dynamic changes or evolution of a variable or multiple variables in a given subject or group of subjects.

Blood analyte. Substances measured in the blood, such as electrolytes and blood cell times, usually using laboratory procedures.

Pathology. Clinical characteristics defining the typical behavior of a disease or medical condition.

Pathophysiology. The physiological processes and events that are associated with a disease or injury.

Biomarker. A measurable substance whose presence indicates some phenomena such as disease or pathophysiological event. Biomarkers are typically used as biological indicators either in isolation or in combination.

9 References

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets: Multiple factor analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(2), 149–179. https://doi.org/10.1002/wics.1246

Ahuja, C. S., Wilson, J. R., Nori, S., Kotter, M. R. N., Druschel, C., Curt, A., & Fehlings, M. G. (2017). Traumatic spinal cord injury. *Nature Reviews Disease Primers*, *3*(1), 1–21. https://doi.org/10.1038/nrdp.2017.18

Albayar, A. A., Roche, A., Swiatkowski, P., Antar, S., Ouda, N., Emara, E., Smith, D. H., Ozturk, A. K., & Awad, B. I. (2019). Biomarkers in Spinal Cord Injury: Prognostic Insights and Future Potentials. *Frontiers in Neurology*, *10*, 27. https://doi.org/10.3389/fneur.2019.00027

Alizadeh, A., Dyck, S. M., & Karimi-Abdolrezaee, S. (2019). Traumatic Spinal Cord Injury: An Overview of Pathophysiology, Models and Acute Injury Mechanisms. *Frontiers in Neurology*, *10*, 282. https://doi.org/10.3389/fneur.2019.00282

Balcı, A. K., Koksal, O., Kose, A., Armagan, E., Ozdemir, F., Inal, T., & Oner, N. (2013). General characteristics of patients with electrolyte imbalance admitted to emergency department. *World Journal of Emergency Medicine*, *4*(2), 113–116. https://doi.org/10.5847/wjem.j.issn.1920-8642.2013.02.005

Betz, R., Biering-Sørensen, F., Burns, S. P., Donovan, W., Graves, D. E., Guest, J., Jones, L., Kirshblum, S., Krassioukov, A., Mulcahey, M. J., Schmidt Read, M., Rodriguez, G. M., Rupp, R., Schuld, C., Tansey, K., Walden, K., & ASIA and ISCoS International Standards Committee. (2019). The 2019 revision of the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI)—What's new? *Spinal Cord*, *57*(10), 815–817. https://doi.org/10.1038/s41393-019-0350-9

Bock, R. D. (1979). Univariate and multivariate analysis of variance of time-structured data. *Longitudinal Research in the Study of Behavior and Development*, 199–231.

Bouch, D. C., & Thompson, J. P. (2008). Severity scoring systems in the critically ill. *Continuing Education in Anaesthesia Critical Care & Pain*, *8*(5), 181–185. https://doi.org/10.1093/bjaceaccp/mkn033

Bourguignon, L., Vo, A. K., Tong, B., Geisler, F., Mach, O., Maier, D., Kramer, J. L. K., Grassner, L., & Jutzeler, C. R. (2021). Natural Progression of Routine Laboratory Markers after Spinal Trauma: A Longitudinal, Multi-Cohort Study. *Journal of Neurotrauma*, *38*(15), 2151–2161. https://doi.org/10.1089/neu.2021.0012

Brown, S. J., Harrington, G. M. B., Hulme, C. H., Morris, R., Bennett, A., Tsang, W.-H., Osman, A., Chowdhury, J., Kumar, N., & Wright, K. T. (2020). A Preliminary Cohort Study Assessing Routine Blood Analyte Levels and Neurological Outcome after Spinal Cord Injury. *Journal of Neurotrauma*, *37*(3), 466–480. https://doi.org/10.1089/neu.2019.6495

Chan, K. S., Fowles, J. B., & Weiner, J. P. (2010). Review: Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature. *Medical Care Research and Review*, *67*(5), 503–527. https://doi.org/10.1177/1077558709359007

Chen, S.-H., Bu, X.-L., Jin, W.-S., Shen, L.-L., Wang, J., Zhuang, Z.-Q., Zhang, T., Zeng, F., Yao, X.-Q., Zhou, H.-D., & Wang, Y.-J. (2017). Altered peripheral profile of blood cells in Alzheimer disease. *Medicine*, *96*(21), e6843. https://doi.org/10.1097/MD.0000000000006843

Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J. P., Southworth, M. R., Stough, W. G., Thoenes, M., Zannad, F., & Zalewski, A. (2017). Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, *106*(1), 1–9. https://doi.org/10.1007/s00392-016-1025-6

Dong, X., Nao, J., Shi, J., & Zheng, D. (2019). Predictive Value of Routine Peripheral Blood Biomarkers in Alzheimer's Disease. *Frontiers in Aging Neuroscience*, *11*. https://www.frontiersin.org/article/10.3389/fnagi.2019.00332

Failli, V., Kopp, M. A., Gericke, C., Martus, P., Klingbeil, S., Brommer, B., Laginha, I., Chen, Y., DeVivo, M. J., Dirnagl, U., & Schwab, J. M. (2012). Functional neurological recovery after spinal cord injury is impaired in patients with infections. *Brain: A Journal of Neurology*, *135*(Pt 11), 3238–3250. https://doi.org/10.1093/brain/aws267

Fitzmaurice, G. M., & Ravichandran, C. (2008). A Primer in Longitudinal Data Analysis. *Circulation*, *118*(19), 2005–2010. https://doi.org/10.1161/CIRCULATIONAHA.107.714618

Fouad, K., Popovich, P. G., Kopp, M. A., & Schwab, J. M. (2021). The neuroanatomical-functional paradox in spinal cord injury. *Nature Reviews. Neurology*, *17*(1), 53–62. https://doi.org/10.1038/s41582-020-00436-x

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Furlan, J. C., Krassioukov, A. V., & Fehlings, M. G. (2006). Hematologic abnormalities within the first week after acute isolated traumatic cervical spinal cord injury: A case-control cohort study. *Spine*, *31*(23), 2674–2683. https://doi.org/10.1097/01.brs.0000244569.91204.01

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, *65*, 1–34. https://doi.org/10.18637/jss.v065.i04

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, *101*(23). https://doi.org/10.1161/01.CIR.101.23.e215

Gueorguieva, R., & Krystal, J. H. (2004). Move Over ANOVA: Progress in Analyzing Repeated-Measures Data andIts Reflection in Papers Published in the Archives of General Psychiatry. *Archives of General Psychiatry*, *61*(3), 310–317. https://doi.org/10.1001/archpsyc.61.3.310

Gurcay, E., Bal, A., Gurcay, A. G., & Cakci, A. (2009). Evaluation of blood and serum markers in spinal cord injured patients with pressure sores. *Saudi Medical Journal*, *30*(3), 413–417.

Haefeli, J., Mabray, M. C., Whetstone, W. D., Dhall, S. S., Pan, J. Z., Upadhyayula, P., Manley, G. T., Bresnahan, J. C., Beattie, M. S., Ferguson, A. R., & Talbott, J. F. (2017). Multivariate Analysis of MRI Biomarkers for Predicting Neurologic Impairment in Cervical Spinal Cord Injury. *AJNR. American Journal of Neuroradiology*, *38*(3), 648–655. https://doi.org/10.3174/ajnr.A5021

Harrington, G. M. B., Cool, P., Hulme, C., Osman, A., Chowdhury, J. R., Kumar, N., Budithi, S., & Wright, K. (2021). Routinely Measured Hematological Markers Can Help to Predict American Spinal Injury Association Impairment Scale Scores after Spinal Cord Injury. *Journal of Neurotrauma*, *38*(3), 301–308. https://doi.org/10.1089/neu.2020.7144

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hawryluk, G., Whetstone, W., Saigal, R., Ferguson, A., Talbott, J., Bresnahan, J., Dhall, S., Pan, J., Beattie, M., & Manley, G. (2015). Mean Arterial Blood Pressure Correlates with Neurological Recovery after Human Spinal Cord Injury: Analysis of High Frequency Physiologic Data. *Journal of Neurotrauma*, *32*(24), 1958–1967. https://doi.org/10.1089/neu.2014.3778

Huie, J. R., Diaz-Arrastia, R., Yue, J. K., Sorani, M. D., Puccio, A. M., Okonkwo, D. O., Manley, G. T., Ferguson, A. R., & TRACK-TBI Investigators. (2019). Testing a Multivariate Proteomic Panel for Traumatic Brain Injury Biomarker Discovery: A TRACK-TBI Pilot Study. *Journal of Neurotrauma*, *36*(1), 100–110. https://doi.org/10.1089/neu.2017.5449

James, S. L., Theadom, A., Ellenbogen, R. G., Bannick, M. S., Montjoy-Venning, W., Lucchesi, L. R., Abbasi, N., Abdulkader, R., Abraha, H. N., Adsuar, J. C., Afarideh, M., Agrawal, S., Ahmadi, A., Ahmed, M. B., Aichour, A. N., Aichour, I., Aichour, M. T. E., Akinyemi, R. O., Akseer, N., … Murray, C. J. L. (2019). Global, regional, and national burden of traumatic brain injury and spinal cord injury, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(1), 56–87. https://doi.org/10.1016/S1474-4422(18)30415-0

Jennrich, R. I., & Sampson, P. F. (1976). Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation. *Technometrics*, *18*(1), 11–17. https://doi.org/10.1080/00401706.1976.10489395

Jiang, H., & Eskridge, K. M. (2000). BIAS IN PRINCIPAL COMPONENTS ANALYSIS DUE TO CORRELATED OBSERVATIONS. *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1247

Jogia, T., Kopp, M. A., Schwab, J. M., & Ruitenberg, M. J. (2021). Peripheral white blood cell responses as emerging biomarkers for patient stratification and prognosis in acute spinal cord injury. *Current Opinion in Neurology*, *34*(6), 796–803. https://doi.org/10.1097/WCO.0000000000000995

Jogia, T., Lübstorf, T., Jacobson, E., Scriven, E., Atresh, S., Nguyen, Q. H., Liebscher, T., Schwab, J. M., Kopp, M. A., Walsham, J., Campbell, K. E., & Ruitenberg, M. J. (2021).

Prognostic value of early leukocyte fluctuations for recovery from traumatic spinal cord injury. *Clinical and Translational Medicine*, *11*(1), e272. https://doi.org/10.1002/ctm2.272

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.35

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jung, T., & Wickrama, K. a. S. (2008). An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Khorasanizadeh, M., Yousefifard, M., Eskian, M., Lu, Y., Chalangari, M., Harrop, J. S., Jazayeri, S. B., Seyedpour, S., Khodaei, B., Hosseini, M., & Rahimi-Movaghar, V. (2019). Neurological recovery following traumatic spinal cord injury: A systematic review and meta-analysis. *Journal of Neurosurgery. Spine*, 1–17. https://doi.org/10.3171/2018.10.SPINE18802

Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing*, *6*(2), 151–157. https://doi.org/10.1177/1099800404267682

Krzysko, M., Smiałowski, T., & Wołynski, W. (2014). Analysis of multivariate repeated measures data using a MANOVA model and principal components. *Biometrical Letters*, *51*(2), 103–114. https://doi.org/10.2478/bile-2014-0008

Kuhn, M. (2021). *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret

Kumar, R., Lim, J., Mekary, R. A., Rattani, A., Dewan, M. C., Sharif, S. Y., Osorio-Fonseca, E., & Park, K. B. (2018). Traumatic Spinal Injury: Global Epidemiology and Worldwide Volume. *World Neurosurgery*, *113*, e345–e363. https://doi.org/10.1016/j.wneu.2018.02.033

Kwon, B. K., Bloom, O., Wanner, I.-B., Curt, A., Schwab, J. M., Fawcett, J., & Wang, K. K. (2019). Neurochemical biomarkers in spinal cord injury. *Spinal Cord*, *57*(10), 819–831. https://doi.org/10.1038/s41393-019-0319-8

Kwon, B. K., Stammers, A. M. T., Belanger, L. M., Bernardo, A., Chan, D., Bishop, C. M., Slobogean, G. P., Zhang, H., Umedaly, H., Giffin, M., Street, J., Boyd, M. C., Paquette, S. J., Fisher, C. G., & Dvorak, M. F. (2010). Cerebrospinal fluid inflammatory cytokines and biomarkers of injury severity in acute human spinal cord injury. *Journal of Neurotrauma*, *27*(4), 669–682. https://doi.org/10.1089/neu.2009.1080

Kyritsis, N., Torres-Espín, A., Schupp, P. G., Huie, J. R., Chou, A., Duong-Fernandez, X., Thomas, L. H., Tsolinas, R. E., Hemmerle, D. D., Pascual, L. U., Singh, V., Pan, J. Z., Talbott, J. F., Whetstone, W. D., Burke, J. F., DiGiorgio, A. M., Weinstein, P. R., Manley, G. T., Dhall, S. S., … Beattie, M. S. (2021). Diagnostic blood RNA profiles for human acute spinal cord injury. *The Journal of Experimental Medicine*, *218*(3). https://doi.org/10.1084/jem.20201795

Lai, D., Xu, H., Koller, D., Foroud, T., & Gao, S. (2016). A multivariate finite mixture latent trajectory model with application to dementia studies. *Journal of Applied Statistics*, *43*(14), 2503–2523. https://doi.org/10.1080/02664763.2016.1141181

Laursen, B. P., & Hoff, E. (2006). Person-Centered and Variable-Centered Approaches to Longitudinal Data. *Merrill-Palmer Quarterly*, *52*(3), 377–389. https://doi.org/10.1353/mpq.2006.0029

Leister, I., Linde, L. D., Vo, A. K., Haider, T., Mattiassich, G., Grassner, L., Schaden, W., Resch, H., Jutzeler, C. R., Geisler, F. H., Kramer, J. L. K., & Aigner, L. (2021). Routine Blood Chemistry Predicts Functional Recovery After Traumatic Spinal Cord Injury: A Post Hoc Analysis. *Neurorehabilitation and Neural Repair*, *35*(4), 321–333. https://doi.org/10.1177/1545968321992328

Liebscher, T., Ludwig, J., Lübstorf, T., Kreutzträger, M., Auhuber, T., Grittner, U., Schäfer, B., Wüstner, G., Ekkernkamp, A., & Kopp, M. A. (2022). Cervical Spine Injuries with Acute Traumatic Spinal Cord Injury: Spinal Surgery Adverse Events and Their Association with Neurological and Functional Outcome. *Spine*, *47*(1), E16–E26. https://doi.org/10.1097/BRS.0000000000004124

Lu, N., Tang, W., He, H., Yu, Q., Crits-Christoph, P., Zhang, H., & Tu, X. (2009). On the Impact of Parametric Assumptions and Robust Alternatives for Longitudinal Data Analysis. *Biometrical Journal. Biometrische Zeitschrift*, *51*(4), 627–643. https://doi.org/10.1002/bimj.200800186

Magrini, A. (2022). *Gbmt* [R]. https://github.com/alessandromagrini/gbmt (Original work published 2021)

Magrini, A. (2022). Assessment of agricultural sustainability in European Union countries: A group-based multivariate trajectory approach. *AStA Advances in Statistical Analysis*. https://doi.org/10.1007/s10182-022-00437-9

Merritt, C. H., Taylor, M. A., Yelton, C. J., & Ray, S. K. (2019). Economic impact of traumatic spinal cord injuries in the United States. *Neuroimmunology and Neuroinflammation*, *6*. https://doi.org/10.20517/2347-8659.2019.15

Microsoft and Hong Ooi. (2019). glmnetUtils: Utilities for "Glmnet." *R Package Version 1.1.2*.

Muthén, B. (2004). Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data. In D. Kaplan, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 346–369). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n19

Muthen, B., & Asparouhov, T. (2008). *Growth mixture modeling: Analysis with non-Gaussian random effects*. 24.

Nagin, D. S. (2014). Group-Based Trajectory Modeling: An Overview. *Annals of Nutrition and Metabolism*, *65*(2–3), 205–210. https://doi.org/10.1159/000360229

Nagin, D. S., Jones, B. L., Passos, V. L., & Tremblay, R. E. (2018). Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*, *27*(7), 2015–2023. https://doi.org/10.1177/0962280216673085

Ohmann, C., Banzi, R., Canham, S., Battaglia, S., Matei, M., Ariyo, C., Becnel, L., Bierer, B., Bowers, S., Clivio, L., Dias, M., Druml, C., Faure, H., Fenner, M., Galvez, J., Ghersi, D., Gluud, C., Groves, T., Houston, P., … Demotes-Mainard, J. (2017). Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open*, *7*(12), e018647. https://doi.org/10.1136/bmjopen-2017-018647

Peek, N., Holmes, J. H., & Sun, J. (2014). Technical challenges for big data in biomedicine and health: Data sources, infrastructure, and analytics. *Yearbook of Medical Informatics*, *9*, 42–47. https://doi.org/10.15265/IY-2014-0018

Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software*, *78*, 1–56. https://doi.org/10.18637/jss.v078.i02

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Ram, N., & Grimm, K. J. (2009). Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups. *International Journal of Behavioral Development*, *33*(6), 565–576. https://doi.org/10.1177/0165025409343765

Redner, R. A., & Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, *26*(2), 195–239. https://doi.org/10.1137/1026034

Roberts, T. T., Leonard, G. R., & Cepela, D. J. (2017). Classifications In Brief: American Spinal Injury Association (ASIA) Impairment Scale. *Clinical Orthopaedics and Related Research*, *475*(5), 1499–1504. https://doi.org/10.1007/s11999-016-5133-4

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rovine, M. J., & McDermott, P. A. (2018). Latent Growth Curve and Repeated Measures ANOVA Contrasts: What the Models are Telling You. *Multivariate Behavioral Research*, *53*(1), 90–101. https://doi.org/10.1080/00273171.2017.1387511

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. JSTOR. https://doi.org/10.2307/2335739

Ruiz, I. A., Squair, J. W., Phillips, A. A., Lukac, C. D., Huang, D., Oxciano, P., Yan, D., & Krassioukov, A. V. (2017). Incidence and Natural Progression of Neurogenic Shock after Traumatic Spinal Cord Injury. *Journal of Neurotrauma*, *35*(3), 461–466. https://doi.org/10.1089/neu.2016.4947

Schober, P., & Vetter, T. R. (2018). Repeated Measures Designs and Analysis of Longitudinal Data: If at First You Do Not Succeed—Try, Try Again. *Anesthesia and Analgesia*, *127*(2), 569–575. https://doi.org/10.1213/ANE.0000000000003511

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Seo, S. (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. *Undefined*. https://www.semanticscholar.org/paper/A-Review-and-Comparison-of-Methods-for-Detecting-in-Seo/cb868f0b242b9623b7544a58b6a21647dfa138a5

Singh, V., Rana, R. K., & Singhal, R. (2013). Analysis of repeated measurement data in the clinical trials. *Journal of Ayurveda and Integrative Medicine*, *4*(2), 77–81. https://doi.org/10.4103/0975-9476.113872

Squair, J. W., Bélanger, L. M., Tsang, A., Ritchie, L., Mac-Thiong, J.-M., Parent, S., Christie, S., Bailey, C., Dhall, S., Street, J., Ailon, T., Paquette, S., Dea, N., Fisher, C. G., Dvorak, M. F., West, C. R., & Kwon, B. K. (2017). Spinal cord perfusion pressure predicts neurologic recovery in acute spinal cord injury. *Neurology*, *89*(16), 1660–1667. https://doi.org/10.1212/WNL.0000000000004519

Suvarna, V. R. (2018). Real world evidence (RWE)—Are we (RWE) ready? *Perspectives in Clinical Research*, *9*(2), 61–63. https://doi.org/10.4103/picr.PICR_36_18

Talbott, J. F., Whetstone, W. D., Readdy, W. J., Ferguson, A. R., Bresnahan, J. C., Saigal, R., Hawryluk, G. W. J., Beattie, M. S., Mabray, M. C., Pan, J. Z., Manley, G. T., & Dhall, S. S. (2015). The Brain and Spinal Injury Center score: A novel, simple, and

reproducible method for assessing the severity of acute cervical spinal cord injury with axial T2-weighted MRI findings. *Journal of Neurosurgery. Spine*, *23*(4), 495–504. https://doi.org/10.3171/2015.1.SPINE141033

Teuling, N. D., Pauws, S., Heuvel, E. van den, & N.V, C. © 2021 K. P. (2022). *latrend: A Framework for Clustering Longitudinal Data* (1.2.1) [Computer software]. https://CRAN.R-project.org/package=latrend

Torres-Espín, A., Haefeli, J., Ehsanian, R., Torres, D., Almeida, C. A., Huie, J. R., Chou, A., Morozov, D., Sanderson, N., Dirlikov, B., Suen, C. G., Nielson, J. L., Kyritsis, N., Hemmerle, D. D., Talbott, J. F., Manley, G. T., Dhall, S. S., Whetstone, W. D., Bresnahan, J. C., … The TRACK-SCI Investigators. (2021). Topological network analysis of patient similarity for precision management of acute blood pressure in spinal cord injury. *ELife*, *10*, e68015. https://doi.org/10.7554/eLife.68015

Tsolinas, R. E., Burke, J. F., DiGiorgio, A. M., Thomas, L. H., Duong-Fernandez, X., Harris, M. H., Yue, J. K., Winkler, E. A., Suen, C. G., Pascual, L. U., Ferguson, A. R., Huie, J. R., Pan, J. Z., Hemmerle, D. D., Singh, V., Torres-Espin, A., Omondi, C., Kyritsis, N., Haefeli, J., … Dhall, S. S. (2020). Transforming Research and Clinical Knowledge in Spinal Cord Injury (TRACK-SCI): An overview of initial enrollment and demographics. *Neurosurgical Focus*, *48*(5), E6. https://doi.org/10.3171/2020.2.FOCUS191030

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co.

Van Der Leeden, R. (1998). Multilevel Analysis of Repeated Measures Data. *Quality and Quantity*, *32*(1), 15–29. https://doi.org/10.1023/A:1004233225855

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, *43*, 100323. https://doi.org/10.1016/j.alcr.2019.100323

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, *23*(1), 42–59. https://doi.org/10.1177/0962280212445834

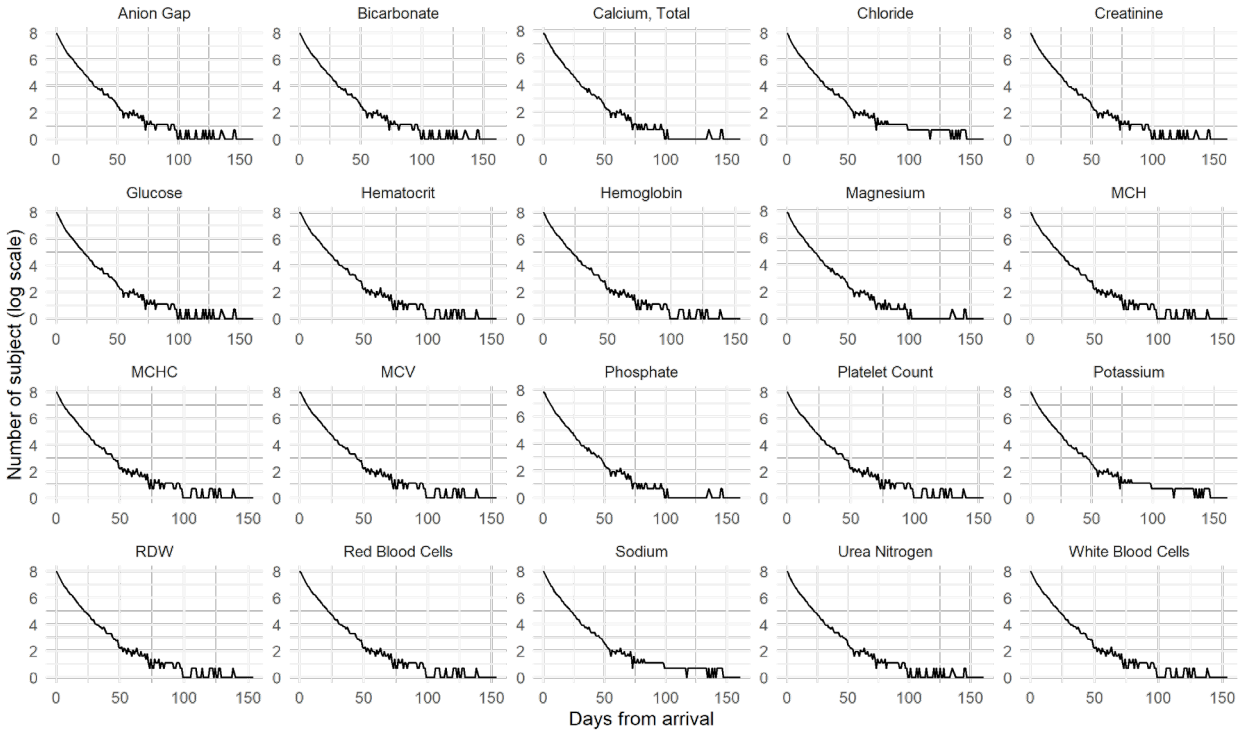
10 Annex

Annexed Table 1. Variables from MIMIC-III and IV used in this work

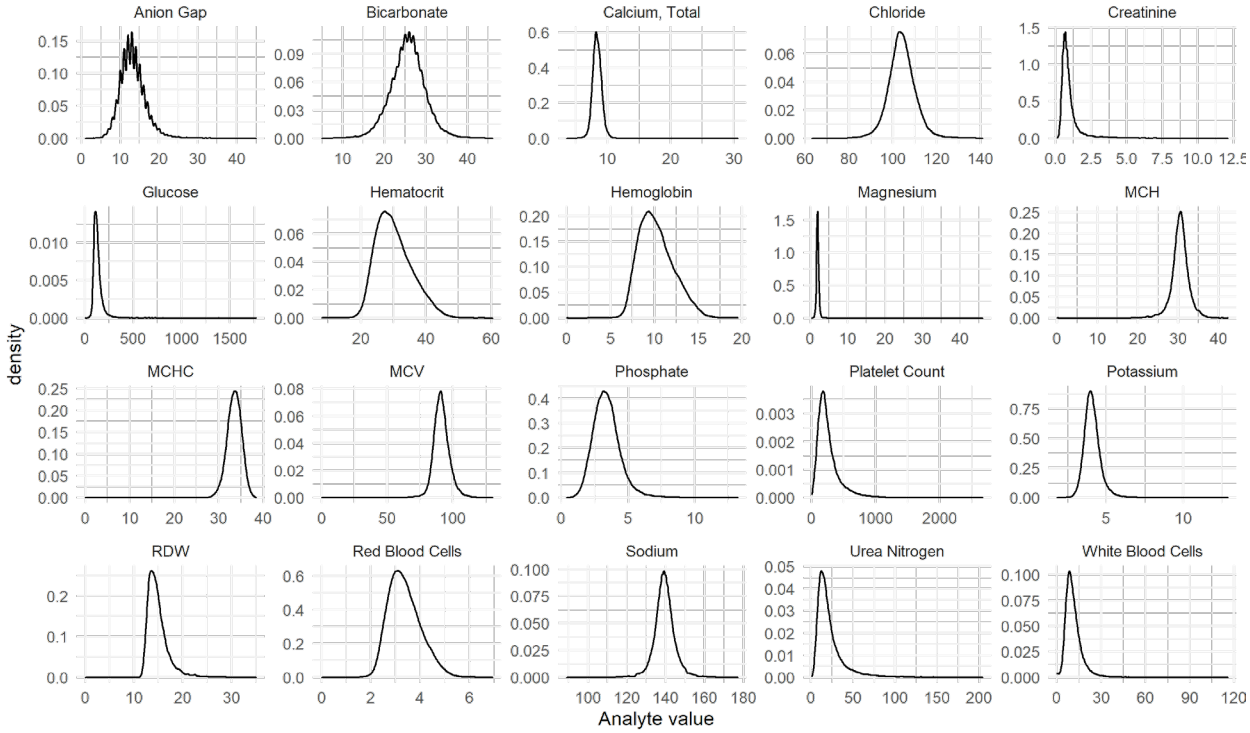
Variable	DataBase/Table	Variable type	Description
MIMIC-III			
SUBJECT_ID	MIMIC-III/PATIENTS MIMIC-III/DIAGNOSES_ICD	numeric	Patient unique identifier. Common key across tables
HADM_ID	MIMIC-III/ADMISSIONS MIMIC-III/DIAGNOSES_ICD MIMIC-III/LABEVENTS	numeric	Unique identifier for a hospital admission. Same SUBJECT_ID may have one or more HADM_ID. Common key across tables
ICD9_CODE	MIMIC-III/DIAGNOSES_ICD	String	ICD9 code
SEQ_NUM	MIMIC-III/DIAGNOSES_ICD	numeric	The numeric sequence of ICD9 code appearance
GENDER	MIMIC-III/PATIENTS	String	Patient's gender
EDREGTIME	MIMIC-III/ADMISSIONS	Date and time	Date and time of patent registration to ED. Surrogate for patient hospital arrival. The date is shifted to preserve privacy.
DOB	MIMIC-III/PATIENTS	Date and time	Date of patient's birth. The date is shifted to preserve privacy and used to calculate the patient's age at arrival.
ITEMID	MIMIC-III/LABEVENTS MIMIC-III/D_LABITEMS	numeric	Unique identifier for the laboratory analyte type
CHARTIME	MIMIC-III/LABEVENTS	Date and time	Date and time of sample extraction for specific laboratory analyte. The date and time are shifted to preserve privacy
VALUENUM	MIMIC-III/LABEVENTS	numeric	Numeric results of the laboratory analyte
LABEL	MIMIC-III/D_LABITEMS	String	Name of the laboratory analyte
FLUID	MIMIC-III/D_LABITEMS	String	Name of the biological sample or fluid
CATEGORY	MIMIC-III/D_LABITEMS	String	Type of laboratory analyte
LOINC_CODE	MIMIC-III/D_LABITEMS	String	LOINC code for each analyte
MIMIC-IV			
subject_id	MIMIC-IV/core/patients	numeric	Patient unique identifier. Common key across tables
hadm_id	MIMIC-IV/core/admissions MIMIC-IV/hosp/diagnoses_icd MIMIC-IV/hosp/labevents	numeric	Unique identifier for a hospital admission. Same SUBJECT_ID may have one or more HADM_ID. Common key across tables
icd_code	MIMIC-IV/hosp/diagnoses_icd	string	ICD code
seq_num	MIMIC-IV/hosp/diagnoses_icd	numeric	Numeric sequence of ICD9 code appearence
edregtime	MIMIC-IV/core/admissions	Date and time	Date and time of patent registration to ED. Surrogate for patient hospital arrival. The date is shifted to preserve privacy.
gender	MIMIC-IV/core/patients	String	Patient's gender
anchore_age	MIMIC-IV/core/patients	numeric	Age of the patient in a block of 3 years at the time of hospital arrival.
itemid	MIMIC-IV/hospt/labevents MIMIC-IV/hospt/d_labitems	numeric	Unique identifier for the laboratory analyte type
chartime	MIMIC-IV/hospt/labevents	Date and time	Date and time of sample extraction for specific laboratory analyte. The date and time are shifted to preserve privacy
valuenum	MIMIC-IV/hospt/labevents	numeric	Numeric results of the laboratory analyte
label	MIMIC-IV/hospt/d_labitems	String	Name of the laboratory analyte
fluid	MIMIC-IV/hospt/d_labitems	String	Name of the biological sample or fluid
category	MIMIC-IV/hospt/d_labitems	String	Type of laboratory analyte
loinc_code	MIMIC-IV/hospt/d_labitems	String	LOINC code for each analyte

Annexed Table 2. Blood analytes collected in the majority of patients

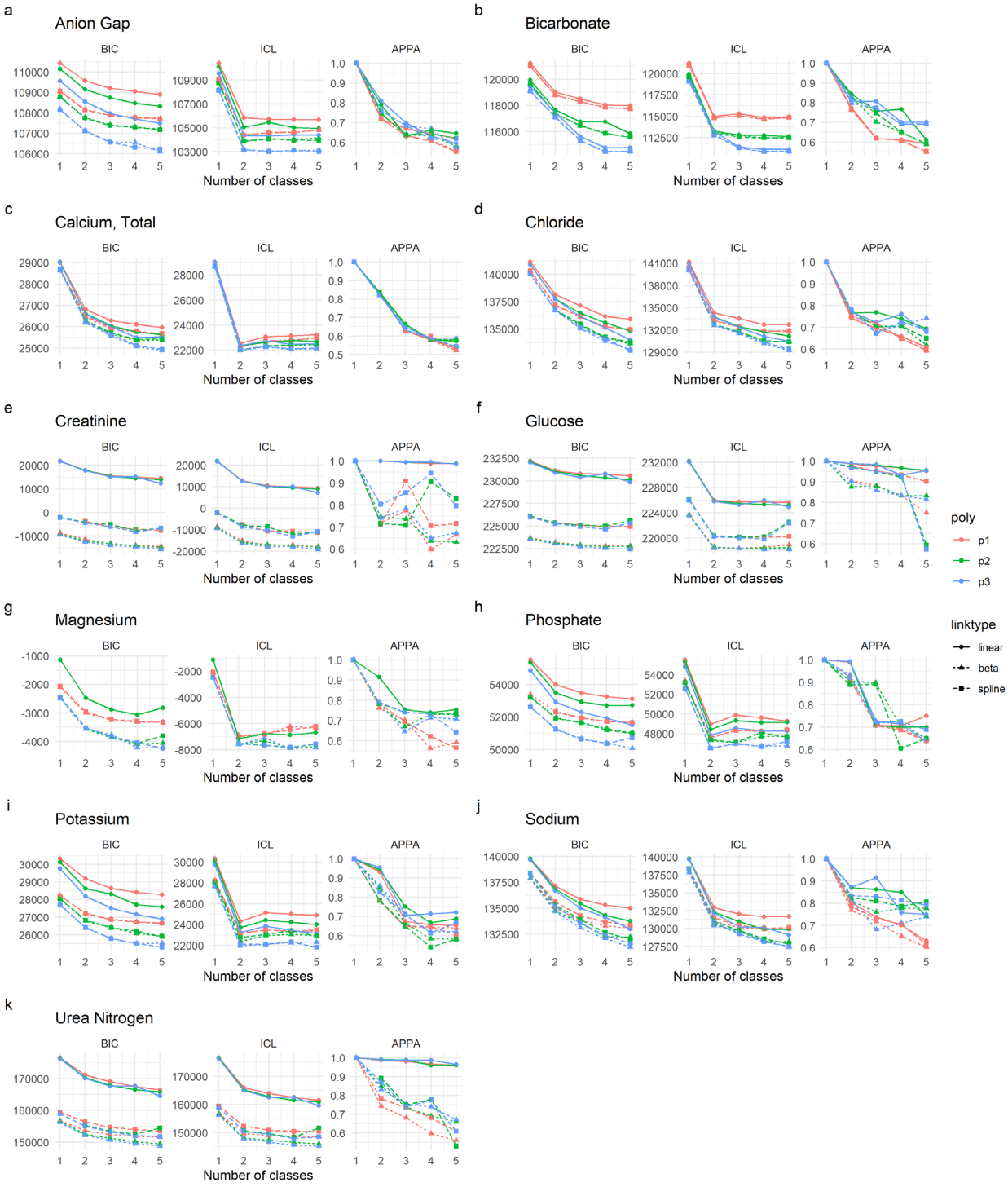
itemid	LOINC	label	fluid	category	proportion of patients
51221	4544-3	Hematocrit	Blood	Hematology	0.98
51301	804-5	White Blood	Blood	Hematology	0.97
51222	718-7	Hemoglobin	Blood	Hematology	0.97
51248	785-6	MCH (mean corpuscular hemoglobin)	Blood	Hematology	0.97
51249	786-4	MCHC (mean corpuscular hemoglobin concentration)	Blood	Hematology	0.97
51250	787-2	MCV (mean corpuscular volume)	Blood	Hematology	0.97
51277	788-0	RDW (red cell distribution width)	Blood	Hematology	0.97
51279	789-8	Red Blood	Blood	Hematology	0.97
51265	777-3	Platelet Count	Blood	Hematology	0.97
50912	2160-0	Creatinine	Blood	Chemistry	0.97
51006	3094-0	Urea Nitrogen	Blood	Chemistry	0.97
50971	2823-3	Potassium	Blood	Chemistry	0.96
50902	2075-0	Chloride	Blood	Chemistry	0.96
50983	2951-2	Sodium	Blood	Chemistry	0.96
50868	1863-0	Anion Gap	Blood	Chemistry	0.96
50882	1963-8	Bicarbonate	Blood	Chemistry	0.96
50931	2345-7	Glucose	Blood	Chemistry	0.96
50960	2601-3	Magnesium	Blood	Chemistry	0.91
50893	2000-8	Calcium, Total	Blood	Chemistry	0.90
50970	2777-1	Phosphate	Blood	Chemistry	0.90



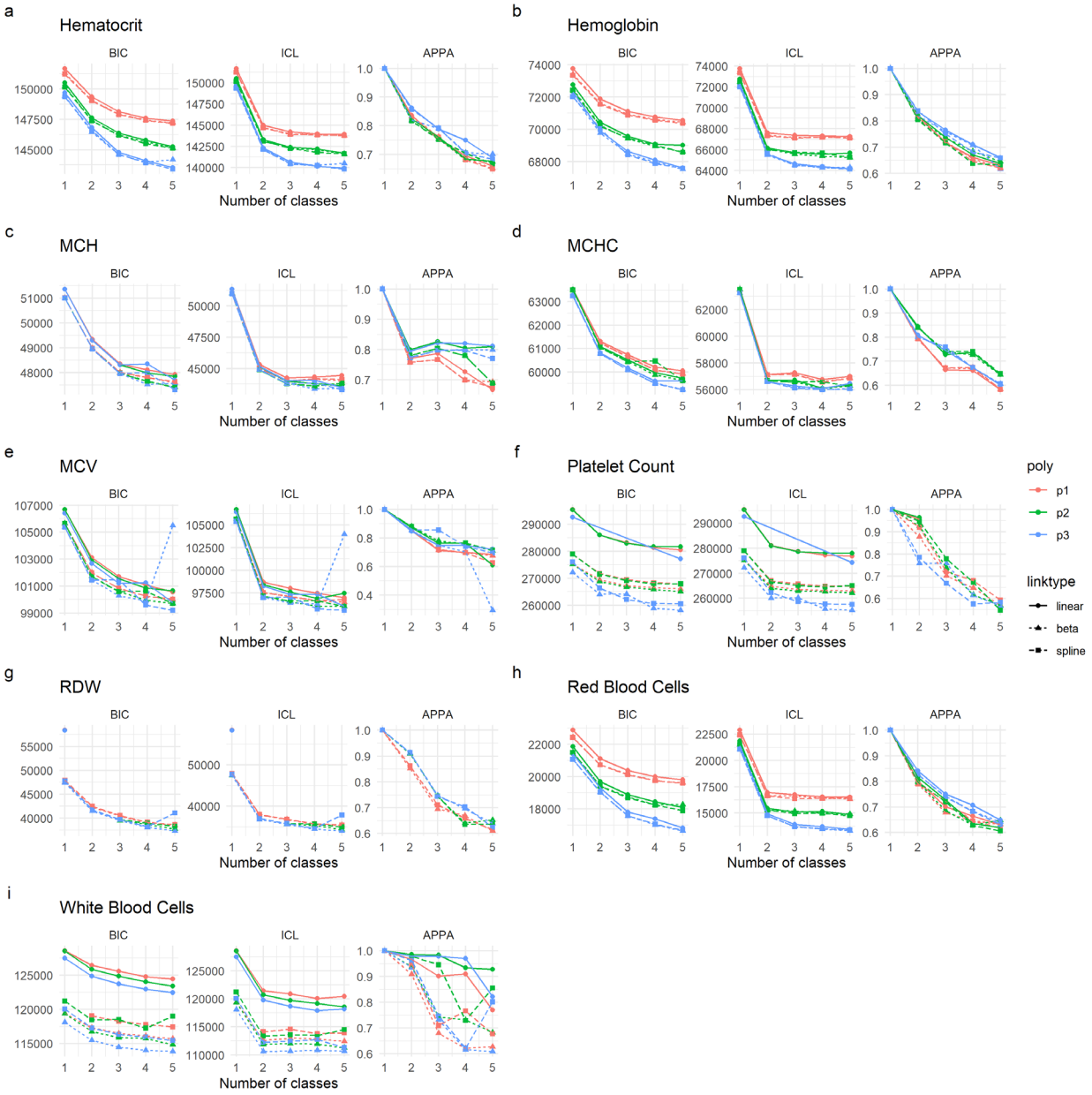
Annexed Figure 1. Number of subjects (in log scale) with data up to a given day from hospital arrival for each analyte in the modeling set



Annexed Figure 2. Marginal distribution of analytes in the modeling set before extreme value detection and processing.



Annexed Figure 3. LCGA model selection metrics (BIC, ICL and APPA) for each one of the chemical analytes in the modeling set.



Annexed Figure 4. LCGA model selection metrics (BIC, ICL and APPA) for each one of the hematology analytes in the modeling set.

Annexed Table 3. Anion Gap. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	110436.6	110436.6	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	110154.4	110154.4	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	109566.8	109566.8	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	109076.7	109076.7	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	108797.5	108797.5	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	108179.9	108179.9	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	109049.4	109049.4	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	108771.3	108771.3	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	108153.3	108153.3	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	109572.9	105845.9	0.72	39.08	60.92	NA	NA	NA
2	linear	9	p2	109153.5	105049.3	0.79	24.31	75.69	NA	NA	NA
2	linear	11	p3	108550.0	104335.8	0.81	20.69	79.31	NA	NA	NA
2	beta	9	p1	108168.6	104428.1	0.72	47.23	52.77	NA	NA	NA
2	beta	11	p2	107772.4	103888.5	0.75	38.31	61.69	NA	NA	NA
2	**beta**	**13**	**p3**	**107131.6**	**103178.4**	**0.76**	**28.85**	**71.15**	**NA**	**NA**	**NA**
2	spline	10	p1	108139.7	104398.2	0.72	46.85	53.15	NA	NA	NA
2	spline	12	p2	107746.8	103867.5	0.75	38.65	61.35	NA	NA	NA
2	**spline**	**14**	**p3**	**107105.8**	**103161.2**	**0.76**	**28.88**	**71.12**	**NA**	**NA**	**NA**
3	linear	10	p1	109211.7	105729.7	0.67	7.85	77.46	14.69	NA	NA
3	linear	13	p2	108742.6	105455.2	0.63	9.58	44.15	46.27	NA	NA
3	linear	16	p3	107964.5	104330.3	0.70	16.77	14.00	69.23	NA	NA
3	beta	12	p1	107899.1	104603.2	0.63	7.04	62.88	30.08	NA	NA
3	beta	15	p2	107407.2	104089.9	0.64	25.42	16.85	57.73	NA	NA
3	**beta**	**18**	**p3**	**106584.1**	**103029.9**	**0.68**	**20.12**	**17.15**	**62.73**	**NA**	**NA**
3	spline	13	p1	107874.9	104575.3	0.63	6.69	62.65	30.65	NA	NA
3	spline	16	p2	107383.3	104061.8	0.64	16.85	25.88	57.27	NA	NA
3	**spline**	**19**	**p3**	**106563.0**	**103000.2**	**0.69**	**22.23**	**16.19**	**61.58**	**NA**	**NA**
4	linear	13	p1	109054.7	105702.1	0.64	5.35	23.15	69.73	1.77	NA
4	linear	17	p2	108478.1	105031.2	0.66	8.58	1.69	68.92	20.81	NA
4	linear	21	p3	107701.7	104414.6	0.63	3.04	17.12	27.12	52.73	NA
4	beta	15	p1	107780.1	104607.7	0.61	5.54	73.00	18.15	3.31	NA
4	beta	19	p2	107305.4	103994.7	0.64	33.35	12.23	1.88	52.54	NA
4	beta	23	p3	106552.5	103068.5	0.67	2.00	17.15	18.54	62.31	NA
4	spline	16	p1	107757.8	104599.6	0.61	5.42	18.38	72.65	3.54	NA
4	spline	20	p2	107290.2	103944.9	0.64	11.58	1.81	31.46	55.15	NA
4	spline	24	p3	106313.5	103109.4	0.62	5.08	18.19	33.69	43.04	NA
5	linear	16	p1	108903.4	105668.6	0.62	5.35	2.12	41.77	49.27	1.50
5	linear	21	p2	108316.5	104959.2	0.65	18.85	8.08	1.62	69.15	2.31
5	linear	26	p3	107491.5	104397.6	0.59	22.65	42.50	4.19	28.92	1.73
5	beta	18	p1	107724.0	104809.1	0.56	4.69	40.19	4.65	48.92	1.54
5	beta	23	p2	107190.4	104160.1	0.58	1.73	5.00	46.00	10.96	36.31
5	beta	28	p3	106102.6	103126.1	0.57	44.15	5.38	30.65	15.00	4.81
5	spline	19	p1	107707.1	104825.9	0.55	4.58	4.92	39.46	49.27	1.77
5	spline	24	p2	107173.3	103942.5	0.62	5.31	2.50	72.38	11.77	8.04
5	spline	29	p3	106219.1	103000.7	0.62	5.19	2.31	15.23	64.27	13.00

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability

Annexed Table 4. Bicarbonate. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	121222.8	121222.8	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	119924.4	119924.4	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	119408.6	119408.6	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	120972.5	120972.5	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	119630.2	119630.2	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	119087.5	119087.5	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	120972.1	120972.1	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	119633.4	119633.4	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	119093.2	119093.2	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	119015.2	114989.3	0.77	74.23	25.77	NA	NA	NA
2	linear	9	p2	117656.3	113266.7	0.84	80.42	19.58	NA	NA	NA
2	linear	11	p3	117439.1	113236.0	0.81	79.15	20.85	NA	NA	NA
2	beta	9	p1	118790.2	114816.7	0.76	71.73	28.27	NA	NA	NA
2	beta	11	p2	117382.4	113070.3	0.83	23.73	76.27	NA	NA	NA
2	beta	13	p3	117134.4	112988.0	0.80	73.73	26.27	NA	NA	NA
2	spline	10	p1	118787.4	114810.3	0.76	71.88	28.12	NA	NA	NA
2	spline	12	p2	117383.0	113064.9	0.83	23.35	76.65	NA	NA	NA
2	spline	14	p3	117089.2	112787.0	0.83	30.15	69.85	NA	NA	NA
3	linear	10	p1	118501.5	115282.3	0.62	22.96	13.85	63.19	NA	NA
3	linear	13	p2	116747.4	112813.7	0.76	20.38	72.35	7.27	NA	NA
3	**linear**	**16**	**p3**	**115615.9**	**111420.1**	**0.81**	**20.46**	**73.27**	**6.27**	**NA**	**NA**
3	beta	12	p1	118279.2	115059.5	0.62	14.50	60.85	24.65	NA	NA
3	beta	15	p2	116443.4	112790.2	0.70	33.81	7.81	58.38	NA	NA
3	beta	18	p3	115304.7	111292.6	0.77	22.38	10.00	67.62	NA	NA
3	spline	13	p1	118275.6	115055.3	0.62	14.46	61.04	24.50	NA	NA
3	spline	16	p2	116447.7	112565.1	0.75	22.69	7.85	69.46	NA	NA
3	spline	19	p3	115307.9	111289.7	0.77	22.15	9.92	67.92	NA	NA
4	linear	13	p1	118030.0	114852.4	0.61	6.96	10.12	70.54	12.38	NA
4	linear	17	p2	116743.3	112755.9	0.77	72.04	6.54	0.54	20.88	NA
4	**linear**	**21**	**p3**	**114754.8**	**111117.6**	**0.70**	**11.15**	**27.38**	**56.46**	**5.00**	**NA**
4	beta	15	p1	117836.7	114674.8	0.61	13.27	9.96	68.96	7.81	NA
4	beta	19	p2	115852.2	112468.3	0.65	46.50	7.77	40.27	5.46	NA
4	beta	23	p3	114461.2	110875.0	0.69	12.58	53.42	5.69	28.31	NA
4	spline	16	p1	117831.7	114668.0	0.61	9.88	13.31	7.73	69.08	NA
4	spline	20	p2	115854.1	112468.0	0.65	46.77	40.19	7.58	5.46	NA
4	spline	24	p3	114461.2	110866.2	0.69	12.27	53.77	5.73	28.23	NA
5	linear	16	p1	117963.6	114885.6	0.59	6.42	7.81	7.92	73.92	3.92
5	linear	21	p2	115825.8	112647.4	0.61	10.04	59.15	22.77	4.54	3.50
5	linear	26	p3	114759.6	111111.4	0.70	10.92	0.50	5.15	55.69	27.73
5	beta	18	p1	117737.4	114864.4	0.55	55.58	1.23	13.85	8.38	20.96
5	beta	23	p2	115550.1	112484.5	0.59	31.27	11.96	49.19	4.42	3.15
5	beta	28	p3	114477.8	110891.3	0.69	0.54	12.27	52.38	5.65	29.15
5	spline	19	p1	117731.8	114859.2	0.55	1.23	55.31	13.96	8.38	21.12
5	spline	24	p2	115549.1	112479.1	0.59	11.88	49.62	30.92	4.46	3.12
5	spline	29	p3	114477.3	110884.7	0.69	12.27	0.54	28.88	5.65	52.65

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability

Annexed Table 5. Calcium. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	29028.37	29028.37	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	29022.51	29022.51	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	29005.72	29005.72	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	28664.13	28664.13	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	28656.65	28656.65	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	28636.82	28636.82	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	28701.97	28701.97	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	28695.90	28695.90	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	28677.71	28677.71	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	26816.92	22551.19	0.84	84.38	15.62	NA	NA	NA
2	**linear**	**9**	**p2**	**26617.48**	**22351.83**	**0.84**	**80.74**	**19.26**	**NA**	**NA**	**NA**
2	linear	11	p3	26583.16	22372.66	0.82	79.05	20.95	NA	NA	NA
2	beta	9	p1	26452.49	22232.63	0.83	82.89	17.11	NA	NA	NA
2	beta	11	p2	26234.21	21986.69	0.83	79.64	20.36	NA	NA	NA
2	beta	13	p3	26191.71	22008.41	0.82	77.41	22.59	NA	NA	NA
2	spline	10	p1	26499.94	22258.15	0.83	83.24	16.76	NA	NA	NA
2	spline	12	p2	26286.12	22024.65	0.83	80.03	19.97	NA	NA	NA
2	spline	14	p3	26247.47	22047.38	0.82	78.03	21.97	NA	NA	NA
3	linear	10	p1	26291.93	23090.22	0.63	38.25	49.49	12.26	NA	NA
3	**linear**	**13**	**p2**	**26050.07**	**22653.88**	**0.66**	**44.68**	**45.42**	**9.91**	**NA**	**NA**
3	linear	16	p3	26000.43	22773.13	0.63	29.13	50.98	19.89	NA	NA
3	beta	12	p1	25923.08	22714.95	0.63	40.41	45.73	13.86	NA	NA
3	beta	15	p2	25674.00	22303.11	0.66	42.76	45.38	11.86	NA	NA
3	beta	18	p3	25571.03	22251.52	0.65	31.32	46.28	22.40	NA	NA
3	spline	13	p1	25970.93	22764.26	0.63	39.98	46.63	13.39	NA	NA
3	spline	16	p2	25723.81	22343.70	0.66	42.68	46.16	11.16	NA	NA
3	spline	19	p3	25638.89	22327.33	0.65	47.22	31.36	21.42	NA	NA
4	linear	13	p1	26114.90	23147.84	0.58	24.59	7.60	60.61	7.20	NA
4	linear	17	p2	25745.35	22781.67	0.58	33.56	47.34	8.26	10.85	NA
4	linear	21	p3	25493.69	22487.25	0.59	17.19	15.07	56.46	11.28	NA
4	beta	15	p1	25747.26	22800.68	0.58	56.97	7.91	27.53	7.60	NA
4	beta	19	p2	25350.92	22393.82	0.58	7.79	33.95	45.11	13.16	NA
4	beta	23	p3	25063.62	22071.65	0.59	18.60	20.60	49.22	11.59	NA
4	spline	16	p1	25798.13	22742.26	0.60	28.47	57.91	8.89	4.74	NA
4	spline	20	p2	25406.33	22450.54	0.58	33.71	8.22	45.50	12.57	NA
4	spline	24	p3	25129.83	22131.93	0.59	19.66	51.64	17.38	11.32	NA
5	linear	16	p1	25971.94	23248.85	0.53	7.28	59.59	7.71	22.00	3.41
5	linear	21	p2	25636.78	22718.48	0.57	45.38	0.90	33.20	11.98	8.54
5	linear	26	p3	25532.91	22526.48	0.59	17.19	56.46	0.00	15.07	11.28
5	beta	18	p1	25645.85	22981.45	0.52	10.02	54.93	22.63	8.85	3.56
5	beta	23	p2	25382.31	22425.20	0.58	7.79	33.95	0.00	45.11	13.16
5	beta	28	p3	24891.48	22123.56	0.54	7.60	20.75	44.68	17.03	9.95
5	spline	19	p1	25683.56	23005.04	0.52	8.93	56.38	8.42	22.75	3.52
5	spline	24	p2	25437.71	22481.92	0.58	33.71	0.00	45.50	8.22	12.57
5	spline	29	p3	24948.25	22172.34	0.54	20.28	16.37	46.91	7.20	9.24

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability

Annexed Table 6. Cloride. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	141138.9	141138.9	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	140914.5	140914.5	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	140909.5	140909.5	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	140303.8	140303.8	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	140047.3	140047.3	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	140046.6	140046.6	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	140327.5	140327.5	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	140081.8	140081.8	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	140079.8	140079.8	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	138176.8	134326.2	0.74	61.54	38.46	NA	NA	NA
2	linear	9	p2	137766.3	133781.1	0.77	53.92	46.08	NA	NA	NA
2	linear	11	p3	137753.7	133749.7	0.77	57.62	42.38	NA	NA	NA
2	beta	9	p1	137145.2	133237.0	0.75	63.19	36.81	NA	NA	NA
2	beta	11	p2	136701.6	132665.9	0.78	55.92	44.08	NA	NA	NA
2	beta	13	p3	136696.3	132641.9	0.78	58.35	41.65	NA	NA	NA
2	spline	10	p1	137210.5	133309.7	0.75	64.15	35.85	NA	NA	NA
2	spline	12	p2	136780.1	132753.0	0.77	57.31	42.69	NA	NA	NA
2	spline	14	p3	136771.9	132722.2	0.78	60.08	39.92	NA	NA	NA
3	linear	10	p1	137145.6	133530.0	0.70	6.12	19.04	74.85	NA	NA
3	linear	13	p2	136444.0	132444.0	0.77	17.15	75.85	7.00	NA	NA
3	linear	16	p3	136138.7	132380.0	0.72	17.69	13.54	68.77	NA	NA
3	beta	12	p1	136126.9	132408.4	0.72	21.15	72.50	6.35	NA	NA
3	beta	15	p2	135361.8	131703.2	0.70	27.54	61.12	11.35	NA	NA
3	beta	18	p3	135047.8	131525.1	0.68	32.23	32.19	35.58	NA	NA
3	spline	13	p1	136149.0	132415.0	0.72	21.23	72.58	6.19	NA	NA
3	spline	16	p2	135463.1	131816.8	0.70	26.65	11.23	62.12	NA	NA
3	spline	19	p3	135147.8	131655.0	0.67	33.12	34.35	32.54	NA	NA
4	linear	13	p1	136158.5	132736.2	0.66	6.27	1.88	67.58	24.27	NA
4	linear	17	p2	135593.5	131754.7	0.74	6.88	5.77	80.00	7.35	NA
4	**linear**	**21**	**p3**	**135100.9**	**131143.0**	**0.76**	**6.58**	**75.15**	**10.69**	**7.58**	**NA**
4	beta	15	p1	135201.6	131835.1	0.65	7.73	62.62	27.77	1.88	NA
4	beta	19	p2	134212.0	130547.9	0.70	14.27	68.00	3.08	14.65	NA
4	**beta**	**23**	**p3**	**133885.1**	**130193.5**	**0.71**	**13.42**	**22.38**	**61.31**	**2.88**	**NA**
4	spline	16	p1	135206.9	131836.0	0.65	7.73	62.77	27.54	1.96	NA
4	spline	20	p2	134249.1	130573.3	0.71	13.23	14.85	68.92	3.00	NA
4	spline	24	p3	134141.3	130372.6	0.72	6.42	9.77	70.00	13.81	NA
5	linear	16	p1	135901.4	132740.1	0.61	6.12	3.08	61.38	28.00	1.42
5	linear	21	p2	134807.8	131200.1	0.69	7.00	11.81	6.62	72.12	2.46
5	linear	26	p3	134003.8	130477.3	0.68	1.62	21.65	23.73	2.04	50.96
5	beta	18	p1	134943.6	131860.0	0.59	7.85	1.54	57.15	3.54	29.92
5	beta	23	p2	133633.4	130419.6	0.62	9.31	15.35	19.42	1.08	54.85
5	beta	28	p3	133114.3	129252.0	0.74	8.69	70.31	7.50	10.62	2.88
5	spline	19	p1	134941.4	131854.0	0.59	7.58	1.62	57.81	29.46	3.54
5	spline	24	p2	133784.9	130409.8	0.65	10.54	6.77	64.46	15.23	3.00
5	spline	29	p3	133027.3	129459.1	0.69	3.96	59.85	16.65	18.42	1.12

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability

Annexed Table 7. Creatinine. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	21923.05	21923.05	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	21816.48	21816.48	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	21770.73	21770.73	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	-8558.82	-8558.82	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	-9065.80	-9065.80	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	-9467.56	-9467.56	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	*	*	*	*	*	*	*	*
1	spline	8	p2	-1986.83	-1986.83	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	-2189.22	-2189.22	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	18008.87	12811.01	1.00	2.27	97.73	NA	NA	NA
2	linear	9	p2	17933.98	12735.78	1.00	2.19	97.81	NA	NA	NA
2	linear	11	p3	17842.81	12643.90	1.00	2.15	97.85	NA	NA	NA
2	**beta**	**9**	**p1**	**-11304.72**	**-15045.83**	**0.72**	**31.49**	**68.51**	**NA**	**NA**	**NA**
2	beta	11	p2	-11999.38	-15883.42	0.75	35.99	64.01	NA	NA	NA
2	beta	13	p3	-12425.80	-16286.27	0.74	50.52	49.48	NA	NA	NA
2	spline	10	p1	-3785.07	-7488.57	0.71	73.51	26.49	NA	NA	NA
2	spline	12	p2	-4159.43	-7883.33	0.72	33.91	66.09	NA	NA	NA
2	spline	14	p3	-4379.73	-8557.86	0.80	14.88	85.12	NA	NA	NA
3	linear	10	p1	15640.41	10469.10	0.99	1.92	95.50	2.58	NA	NA
3	linear	13	p2	15360.67	10186.99	0.99	1.92	95.46	2.61	NA	NA
3	linear	16	p3	15173.67	9998.35	0.99	2.04	95.35	2.61	NA	NA
3	**beta**	**12**	**p1**	**-13264.62**	**-17290.95**	**0.77**	**2.27**	**80.43**	**17.30**	**NA**	**NA**
3	beta	15	p2	-13140.42	-16968.73	0.74	10.19	70.67	19.15	NA	NA
3	beta	18	p3	-13960.55	-18043.91	0.78	8.07	78.82	13.11	NA	NA
3	spline	13	p1	-5941.25	-10669.54	0.91	3.65	90.73	5.61	NA	NA
3	spline	16	p2	-4920.64	-8604.18	0.71	39.98	3.11	56.90	NA	NA
3	spline	19	p3	-5895.30	-10346.31	0.86	8.80	86.62	4.58	NA	NA
4	linear	13	p1	15181.73	10038.30	0.99	1.85	0.69	94.96	2.50	NA
4	linear	17	p2	14581.36	9412.75	0.99	0.81	1.69	94.96	2.54	NA
4	linear	21	p3	15212.99	10037.67	0.99	2.04	95.35	0.00	2.61	NA
4	beta	15	p1	-13966.05	-17080.39	0.60	1.88	26.37	58.86	12.88	NA
4	beta	19	p2	-14087.84	-17403.82	0.64	3.81	21.18	64.24	10.77	NA
4	beta	23	p3	-14542.07	-17924.60	0.65	6.42	23.49	59.05	11.03	NA
4	spline	16	p1	-6995.12	-10671.50	0.71	2.38	74.05	20.84	2.73	NA
4	spline	20	p2	-7461.05	-12169.69	0.91	2.54	88.74	6.54	2.19	NA
4	spline	24	p3	-8124.12	-13037.75	0.94	2.42	4.08	91.27	2.23	NA
5	linear	16	p1	14520.56	9380.11	0.99	0.88	2.50	0.81	93.54	2.27
5	linear	21	p2	14102.31	8961.80	0.99	0.38	1.69	94.73	1.54	1.65
5	linear	26	p3	12332.60	7199.29	0.99	0.62	2.15	93.77	0.81	2.65
5	beta	18	p1	-14353.40	-17813.36	0.67	1.81	6.69	59.75	27.10	4.65
5	beta	23	p2	-14744.44	-18030.64	0.63	3.00	6.31	41.83	38.06	10.80
5	beta	28	p3	-15608.22	-19115.01	0.67	2.58	9.23	65.86	14.49	7.84
5	spline	19	p1	-7566.47	-11289.71	0.72	2.08	1.88	75.20	18.49	2.35
5	spline	24	p2	-6637.76	-10957.86	0.83	6.07	0.00	3.92	86.39	3.61
5	spline	29	p3	-6611.82	-10755.93	0.80	3.08	7.69	86.04	0.00	3.19

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 8. Glucose. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	232188.0	232188.0	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	232129.3	232129.3	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	232026.3	232026.3	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	223687.9	223687.9	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	223635.1	223635.1	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	223537.1	223537.1	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	*	*	*	*	*	*	*	*
1	spline	8	p2	226076.4	226076.4	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	225978.5	225978.5	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	231131.9	226001.3	0.99	5.08	94.92	NA	NA	NA
2	linear	9	p2	231033.7	225896.2	0.99	5.12	94.88	NA	NA	NA
2	linear	11	p3	230920.2	225782.4	0.99	5.08	94.92	NA	NA	NA
2	**beta**	**9**	**p1**	**223192.8**	**218480.4**	**0.91**	**6.85**	**93.15**	**NA**	**NA**	**NA**
2	beta	11	p2	223131.9	218584.8	0.87	10.38	89.62	NA	NA	NA
2	beta	13	p3	223044.7	218341.7	0.90	8.23	91.77	NA	NA	NA
2	spline	10	p1	225395.2	220318.3	0.98	94.85	5.15	NA	NA	NA
2	spline	12	p2	225355.6	220312.4	0.97	6.15	93.85	NA	NA	NA
2	spline	14	p3	225260.8	220226.8	0.97	6.42	93.58	NA	NA	NA
3	linear	10	p1	230781.2	225712.0	0.97	4.38	93.85	1.77	NA	NA
3	linear	13	p2	230589.5	225482.5	0.98	4.65	93.96	1.38	NA	NA
3	linear	16	p3	230377.3	225262.8	0.98	2.77	2.65	94.58	NA	NA
3	**beta**	**12**	**p1**	**222905.1**	**218326.8**	**0.88**	**3.85**	**1.88**	**94.27**	**NA**	**NA**
3	beta	15	p2	222885.0	218310.8	0.88	2.54	6.54	90.92	NA	NA
3	beta	18	p3	222745.3	218286.7	0.86	5.42	5.96	88.62	NA	NA
3	spline	13	p1	225097.0	220142.7	0.95	93.31	1.69	5.00	NA	NA
3	spline	16	p2	225096.8	220162.1	0.95	2.46	6.31	91.23	NA	NA
3	spline	19	p3	224958.9	220013.5	0.95	2.46	6.42	91.12	NA	NA
4	linear	13	p1	230719.0	225689.6	0.97	5.12	0.62	92.77	1.50	NA
4	linear	17	p2	230323.0	225301.0	0.97	4.38	92.23	2.04	1.35	NA
4	linear	21	p3	230778.7	225935.7	0.93	92.65	5.69	0.00	1.65	NA
4	beta	15	p1	222835.1	218501.4	0.83	3.04	1.73	6.08	89.15	NA
4	beta	19	p2	222705.1	218377.6	0.83	3.23	6.62	87.38	2.77	NA
4	beta	23	p3	222525.1	218191.2	0.83	4.96	2.69	87.46	4.88	NA
4	spline	16	p1	224974.4	220125.8	0.93	1.54	3.92	91.50	3.04	NA
4	spline	20	p2	225000.9	220204.4	0.92	2.46	6.46	1.69	89.38	NA
4	spline	24	p3	224652.3	219824.5	0.93	2.19	4.96	91.27	1.58	NA
5	linear	16	p1	230556.6	225611.0	0.95	6.69	1.65	89.85	1.46	0.35
5	linear	21	p2	230117.1	225147.1	0.96	4.42	0.65	1.69	91.35	1.88
5	linear	26	p3	229837.9	224877.2	0.95	1.96	91.19	1.58	0.58	4.69
5	beta	18	p1	222840.0	218939.0	0.75	3.50	5.81	2.08	86.88	1.73
5	beta	23	p2	222736.5	218409.0	0.83	3.23	6.62	87.38	0.00	2.77
5	beta	28	p3	222423.6	218188.1	0.81	2.85	4.12	2.62	86.35	4.08
5	spline	19	p1	224944.8	220256.9	0.90	0.81	3.65	91.27	1.08	3.19
5	spline	24	p2	225623.6	222537.7	0.59	5.19	26.15	0.04	0.00	68.62
5	spline	29	p3	225351.0	222372.6	0.57	3.27	12.69	0.00	77.96	6.08

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 9. Hematocrit. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	151697.9	151697.9	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	150524.6	150524.6	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	149697.8	149697.8	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	151313.5	151313.5	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	150213.0	150213.0	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	149421.9	149421.9	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	151242.0	151242.0	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	150154.8	150154.8	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	149366.0	149366.0	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	149362.7	144996.8	0.84	51.76	48.24	NA	NA	NA
2	linear	9	p2	147617.8	143310.0	0.82	47.82	52.18	NA	NA	NA
2	linear	11	p3	146824.2	142319.1	0.86	71.32	28.68	NA	NA	NA
2	beta	9	p1	149079.8	144724.3	0.83	55.09	44.91	NA	NA	NA
2	beta	11	p2	147421.4	143146.6	0.82	51.49	48.51	NA	NA	NA
2	beta	13	p3	146465.8	142183.1	0.82	40.89	59.11	NA	NA	NA
2	spline	10	p1	149031.3	144694.8	0.83	55.05	44.95	NA	NA	NA
2	spline	12	p2	147381.9	143116.4	0.82	51.49	48.51	NA	NA	NA
2	spline	14	p3	146605.1	142128.2	0.86	71.71	28.29	NA	NA	NA
3	linear	10	p1	148139.1	144212.5	0.75	11.91	59.80	28.29	NA	NA
3	linear	13	p2	146377.6	142405.1	0.76	14.17	49.50	36.33	NA	NA
3	**linear**	**16**	**p3**	**144785.7**	**140663.9**	**0.79**	**19.26**	**44.56**	**36.18**	**NA**	**NA**
3	beta	12	p1	147936.5	143958.4	0.76	10.22	62.67	27.11	NA	NA
3	beta	15	p2	146223.4	142280.4	0.75	48.55	17.69	33.77	NA	NA
3	beta	18	p3	144615.6	140483.0	0.79	44.03	34.65	21.32	NA	NA
3	spline	13	p1	147894.3	143925.1	0.76	63.25	10.11	26.65	NA	NA
3	spline	16	p2	146194.3	142268.7	0.75	17.27	49.23	33.50	NA	NA
3	spline	19	p3	144604.7	140470.6	0.79	19.64	47.78	32.58	NA	NA
4	linear	13	p1	147604.2	143971.7	0.70	10.72	27.79	47.17	14.32	NA
4	linear	17	p2	145785.8	142211.8	0.68	14.59	18.91	45.25	21.25	NA
4	**linear**	**21**	**p3**	**144087.3**	**140172.6**	**0.75**	**14.82**	**31.85**	**17.73**	**35.60**	**NA**
4	beta	15	p1	147445.3	143855.7	0.69	9.88	33.04	43.57	13.51	NA
4	beta	19	p2	145664.9	142105.7	0.68	51.76	14.13	12.63	21.48	NA
4	beta	23	p3	143953.7	140260.3	0.71	44.14	18.99	15.24	21.63	NA
4	spline	16	p1	147410.8	143852.3	0.68	44.41	32.31	9.80	13.48	NA
4	spline	20	p2	145503.3	141822.5	0.70	12.71	44.49	22.21	20.60	NA
4	spline	24	p3	143929.6	140252.4	0.70	21.98	46.59	17.15	14.28	NA
5	linear	16	p1	147367.0	143922.2	0.66	10.87	25.34	43.80	18.38	1.61
5	linear	21	p2	145249.4	141716.8	0.68	13.97	19.33	29.06	30.47	7.16
5	linear	26	p3	143526.0	139931.8	0.69	12.98	9.61	14.62	44.56	18.22
5	beta	18	p1	147211.2	143809.9	0.65	9.72	23.09	45.87	19.72	1.61
5	beta	23	p2	145137.5	141613.4	0.67	13.21	31.47	28.64	18.87	7.81
5	beta	28	p3	144178.0	140515.7	0.70	22.51	30.17	6.55	17.88	22.89
5	spline	19	p1	147175.5	143786.8	0.65	9.57	22.17	46.90	19.75	1.61
5	spline	24	p2	145124.2	141637.4	0.67	13.06	30.86	29.21	18.80	8.08
5	spline	29	p3	143398.0	139825.9	0.68	31.24	18.95	13.17	24.73	11.91

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 10. Hemoglobin. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	73769.28	73769.28	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	72758.97	72758.97	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	72320.83	72320.83	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	73413.67	73413.67	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	72468.49	72468.49	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	72070.44	72070.44	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	73344.64	73344.64	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	72409.56	72409.56	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	72014.11	72014.11	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	71887.85	67614.69	0.82	49.52	50.48	NA	NA	NA
2	linear	9	p2	70441.03	66165.70	0.82	39.72	60.28	NA	NA	NA
2	linear	11	p3	69899.50	65552.89	0.83	31.08	68.92	NA	NA	NA
2	beta	9	p1	71604.38	67367.92	0.81	55.01	44.99	NA	NA	NA
2	beta	11	p2	70246.38	66042.10	0.81	41.87	58.13	NA	NA	NA
2	beta	13	p3	69765.25	65496.83	0.82	32.81	67.19	NA	NA	NA
2	spline	10	p1	71546.27	67333.11	0.81	55.70	44.30	NA	NA	NA
2	spline	12	p2	70193.62	66005.76	0.80	41.64	58.36	NA	NA	NA
2	spline	14	p3	70029.64	65660.94	0.84	74.07	25.93	NA	NA	NA
3	linear	10	p1	71110.02	67374.35	0.72	10.87	50.48	38.65	NA	NA
3	linear	13	p2	69575.04	65733.69	0.74	12.99	40.11	46.91	NA	NA
3	**linear**	**16**	**p3**	**68635.46**	**64646.77**	**0.77**	**16.60**	**35.96**	**47.45**	**NA**	**NA**
3	beta	12	p1	70927.78	67167.08	0.72	9.22	56.55	34.23	NA	NA
3	beta	15	p2	69424.55	65594.09	0.74	42.76	12.41	44.83	NA	NA
3	beta	18	p3	68475.53	64521.36	0.76	45.52	37.11	17.36	NA	NA
3	spline	13	p1	70878.57	67127.97	0.72	57.74	8.95	33.31	NA	NA
3	spline	16	p2	69450.62	65718.41	0.72	18.86	38.11	43.03	NA	NA
3	spline	19	p3	68421.78	64508.12	0.75	36.80	45.64	17.56	NA	NA
4	linear	13	p1	70771.33	67328.74	0.66	9.26	30.89	50.40	9.45	NA
4	linear	17	p2	69076.28	65582.73	0.67	8.84	35.77	34.69	20.71	NA
4	**linear**	**21**	**p3**	**68095.38**	**64399.43**	**0.71**	**13.52**	**27.85**	**39.11**	**19.52**	**NA**
4	beta	15	p1	70605.06	67188.54	0.66	7.53	34.46	48.87	9.14	NA
4	beta	19	p2	68940.93	65373.73	0.69	41.95	29.58	9.30	19.17	NA
4	beta	23	p3	67964.81	64279.52	0.71	30.66	13.14	37.19	19.02	NA
4	spline	16	p1	70562.28	67195.24	0.65	7.57	33.96	49.48	8.99	NA
4	spline	20	p2	69008.77	65685.06	0.64	20.71	53.98	9.57	15.75	NA
4	spline	24	p3	67874.84	64327.82	0.68	23.93	49.02	18.79	8.26	NA
5	linear	16	p1	70549.33	67262.38	0.63	5.49	21.28	39.45	1.50	32.27
5	linear	21	p2	69009.40	65685.72	0.64	13.48	19.36	25.93	22.78	18.44
5	linear	26	p3	67613.18	64173.11	0.66	10.56	24.01	36.07	22.13	7.22
5	beta	18	p1	70432.34	67196.62	0.62	4.34	35.84	5.42	47.37	7.03
5	beta	23	p2	68611.38	65252.63	0.65	8.95	28.81	19.86	38.65	3.73
5	beta	28	p3	67527.44	64316.75	0.62	11.06	24.51	32.73	21.28	10.41
5	spline	19	p1	70354.94	67125.08	0.62	4.80	17.63	44.72	31.39	1.46
5	spline	24	p2	68570.87	65281.07	0.63	9.26	26.55	19.63	40.45	4.11
5	spline	29	p3	67576.20	64142.86	0.66	9.22	38.69	17.98	8.76	25.36

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 11. Magnesium. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	*	*	*	*	*	*	*	*
1	linear	5	p2	-1141.81	-1141.81	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	*	*	*	*	*	*	*	*
1	beta	6	p1	-2107.71	-2107.71	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	-2489.68	-2489.68	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	-2498.37	-2498.37	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	-2071.61	-2071.61	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	-2458.76	-2458.76	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	-2470.24	-2470.24	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	*	*	*	*	*	*	*	*
2	linear	9	p2	-2476.35	-7164.50	0.91	94.73	5.27	NA	NA	NA
2	linear	11	p3	*	*	*	*	*	*	*	*
2	beta	9	p1	-3003.55	-6905.01	0.76	12.60	87.40	NA	NA	NA
2	beta	11	p2	-3563.49	-7561.17	0.78	80.26	19.74	NA	NA	NA
2	beta	13	p3	-3559.42	-7586.43	0.79	19.51	80.49	NA	NA	NA
2	spline	10	p1	-2962.96	-6989.53	0.79	10.38	89.62	NA	NA	NA
2	spline	12	p2	-3514.31	-7532.35	0.78	19.24	80.76	NA	NA	NA
2	**spline**	**14**	**p3**	**-3512.70**	**-7557.73**	**0.79**	**19.08**	**80.92**	**NA**	**NA**	**NA**
3	linear	10	p1	*	*	*	*	*	*	*	*
3	linear	13	p2	-2880.04	-6741.85	0.75	13.62	83.22	3.16	NA	NA
3	linear	16	p3	*	*	*	*	*	*	*	*
3	beta	12	p1	-3252.18	-6848.79	0.70	6.83	6.52	86.66	NA	NA
3	beta	15	p2	-3863.03	-7296.79	0.67	50.64	46.08	3.28	NA	NA
3	beta	18	p3	-3760.60	-7074.22	0.65	10.81	75.38	13.81	NA	NA
3	spline	13	p1	-3219.60	-6775.16	0.69	6.28	6.87	86.85	NA	NA
3	spline	16	p2	-3840.56	-7640.23	0.74	83.61	11.90	4.49	NA	NA
3	**spline**	**19**	**p3**	**-3832.12**	**-7629.19**	**0.74**	**83.11**	**13.03**	**3.86**	**NA**	**NA**
4	linear	13	p1	*	*	*	*	*	*	*	*
4	linear	17	p2	-3058.40	-6851.57	0.74	2.26	80.14	15.92	1.68	NA
4	linear	21	p3	*	*	*	*	*	*	*	*
4	beta	15	p1	-3322.22	-6204.70	0.56	70.43	2.42	21.54	5.62	NA
4	beta	19	p2	-4085.14	-7822.96	0.73	2.58	10.92	84.35	2.15	NA
4	beta	23	p3	-4217.66	-7876.75	0.71	6.20	10.77	2.97	80.06	NA
4	spline	16	p1	-3280.49	-6468.64	0.62	3.98	4.56	84.00	7.45	NA
4	spline	20	p2	-4054.20	-7810.71	0.73	1.87	11.43	2.38	84.32	NA
4	**spline**	**24**	**p3**	**-4037.02**	**-7790.03**	**0.73**	**2.34**	**11.94**	**83.89**	**1.83**	**NA**
5	linear	16	p1	*	*	*	*	*	*	*	*
5	linear	21	p2	-2817.25	-6679.05	0.75	13.62	83.22	0.00	0.00	3.16
5	linear	26	p3	*	*	*	*	*	*	*	*
5	beta	18	p1	-3318.07	-6351.89	0.59	0.86	4.21	79.79	4.49	10.65
5	beta	23	p2	-4053.75	-7791.56	0.73	2.58	10.92	0.00	84.35	2.15
5	beta	28	p3	-4207.18	-7831.36	0.71	0.70	10.77	6.05	79.52	2.97
5	spline	19	p1	-3329.88	-6226.31	0.57	76.20	2.11	16.11	4.21	1.37
5	spline	24	p2	-3797.80	-7563.27	0.73	0.00	12.10	82.79	0.70	4.41
5	spline	29	p3	-4236.94	-7525.39	0.64	3.75	14.87	2.77	76.24	2.38

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 12. MCH. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	51361.75	51361.75	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	51355.96	51355.96	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	51352.73	51352.73	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	51002.91	51002.91	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	51001.58	51001.58	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	50996.44	50996.44	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	50999.59	50999.59	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	50998.31	50998.31	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	50993.14	50993.14	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	49329.15	45317.40	0.77	25.16	74.84	NA	NA	NA
2	linear	9	p2	49293.74	45131.36	0.80	20.40	79.60	NA	NA	NA
2	linear	11	p3	49295.92	45160.24	0.79	21.55	78.45	NA	NA	NA
2	beta	9	p1	48973.90	45025.41	0.76	29.54	70.46	NA	NA	NA
2	beta	11	p2	48952.28	44888.91	0.78	25.51	74.49	NA	NA	NA
2	beta	13	p3	48951.83	44925.26	0.77	27.12	72.88	NA	NA	NA
2	spline	10	p1	48973.24	45026.80	0.76	29.12	70.88	NA	NA	NA
2	spline	12	p2	48951.24	44886.92	0.78	25.09	74.91	NA	NA	NA
2	spline	14	p3	48950.86	44923.71	0.77	26.82	73.18	NA	NA	NA
3	linear	10	p1	48353.42	44256.69	0.79	15.44	80.79	3.76	NA	NA
3	**linear**	**13**	**p2**	**48307.71**	**44005.48**	**0.83**	**12.26**	**83.71**	**4.03**	**NA**	**NA**
3	linear	16	p3	48304.14	44023.35	0.82	12.56	83.44	4.00	NA	NA
3	beta	12	p1	48008.25	44015.61	0.77	16.17	78.76	5.07	NA	NA
3	beta	15	p2	47973.40	43790.99	0.80	12.68	82.14	5.19	NA	NA
3	beta	18	p3	47966.09	43814.56	0.80	13.37	81.18	5.46	NA	NA
3	spline	13	p1	48000.71	44006.85	0.77	15.98	78.95	5.07	NA	NA
3	spline	16	p2	47966.35	43783.28	0.80	12.79	82.02	5.19	NA	NA
3	spline	19	p3	47959.18	43806.84	0.80	13.37	81.25	5.38	NA	NA
4	linear	13	p1	48109.87	44324.35	0.73	3.23	13.56	79.52	3.69	NA
4	**linear**	**17**	**p2**	**47981.31**	**43790.23**	**0.81**	**13.56**	**2.38**	**82.52**	**1.54**	**NA**
4	linear	21	p3	48345.06	44073.54	0.82	12.41	0.00	83.40	4.19	NA
4	beta	15	p1	47809.98	44168.86	0.70	15.90	76.45	2.88	4.76	NA
4	beta	19	p2	47656.01	43592.59	0.78	13.22	82.02	1.73	3.03	NA
4	beta	23	p3	47523.06	43363.44	0.80	13.83	80.68	4.96	0.54	NA
4	spline	16	p1	47802.86	44161.41	0.70	2.92	15.87	76.45	4.76	NA
4	spline	20	p2	47649.53	43588.51	0.78	13.22	82.02	1.73	3.03	NA
4	spline	24	p3	47998.50	43846.16	0.80	13.37	81.25	0.00	5.38	NA
5	linear	16	p1	47923.25	44448.74	0.67	2.96	13.10	3.19	78.22	2.54
5	**linear**	**21**	**p2**	**47849.50**	**43631.79**	**0.81**	**12.41**	**0.61**	**82.06**	**1.19**	**3.73**
5	linear	26	p3	47692.61	43466.71	0.81	12.75	0.35	81.64	4.15	1.11
5	beta	18	p1	47598.48	43988.55	0.69	2.80	76.14	5.07	14.29	1.69
5	beta	23	p2	47409.45	43822.98	0.69	3.92	76.76	14.71	1.46	3.15
5	beta	28	p3	47562.38	43402.76	0.80	13.83	0.00	80.68	4.96	0.54
5	spline	19	p1	47638.52	44114.94	0.68	14.83	2.77	77.68	1.96	2.77
5	spline	24	p2	47402.69	43817.25	0.69	3.92	14.75	76.72	1.46	3.15
5	spline	29	p3	47310.40	43299.33	0.77	13.60	80.33	3.42	2.27	0.38

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 13. MCHC. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	63530.88	63530.88	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	63538.51	63538.51	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	63287.39	63287.39	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	63492.14	63492.14	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	63502.96	63502.96	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	63245.29	63245.29	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	63493.72	63493.72	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	63501.59	63501.59	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	63246.55	63246.55	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	61304.89	57146.83	0.80	20.78	79.22	NA	NA	NA
2	**linear**	**9**	**p2**	**61088.50**	**56692.96**	**0.84**	**17.10**	**82.90**	**NA**	**NA**	**NA**
2	**linear**	**11**	**p3**	**60822.30**	**56612.98**	**0.81**	**22.28**	**77.72**	**NA**	**NA**	**NA**
2	beta	9	p1	61245.91	57110.91	0.79	19.63	80.37	NA	NA	NA
2	beta	11	p2	61049.74	56676.74	0.84	83.10	16.90	NA	NA	NA
2	beta	13	p3	60775.59	56603.52	0.80	22.13	77.87	NA	NA	NA
2	spline	10	p1	61243.30	57109.92	0.79	19.48	80.52	NA	NA	NA
2	spline	12	p2	61048.34	56674.96	0.84	16.71	83.29	NA	NA	NA
2	spline	14	p3	60774.02	56603.67	0.80	21.86	78.14	NA	NA	NA
3	linear	10	p1	60742.76	57280.93	0.66	6.49	34.11	59.39	NA	NA
3	**linear**	**13**	**p2**	**60510.13**	**56721.26**	**0.73**	**8.61**	**17.94**	**73.45**	**NA**	**NA**
3	**linear**	**16**	**p3**	**60169.67**	**56288.05**	**0.75**	**10.18**	**17.33**	**72.49**	**NA**	**NA**
3	beta	12	p1	60648.41	57155.07	0.67	6.45	32.35	61.20	NA	NA
3	beta	15	p2	60443.08	56590.78	0.74	17.90	73.84	8.26	NA	NA
3	beta	18	p3	60093.77	56157.81	0.76	9.10	17.06	73.84	NA	NA
3	spline	13	p1	60642.18	57141.79	0.67	6.26	61.24	32.50	NA	NA
3	spline	16	p2	60436.35	56584.35	0.74	8.03	18.56	73.42	NA	NA
3	spline	19	p3	60083.87	56126.16	0.76	9.14	73.68	17.17	NA	NA
4	linear	13	p1	60212.96	56766.31	0.66	5.26	60.12	34.00	0.61	NA
4	**linear**	**17**	**p2**	**59964.33**	**56126.27**	**0.74**	**4.42**	**23.01**	**71.72**	**0.85**	**NA**
4	linear	21	p3	59615.71	56101.16	0.68	6.92	44.14	0.77	48.18	NA
4	beta	15	p1	60083.65	56596.62	0.67	4.96	32.04	61.81	1.19	NA
4	beta	19	p2	59871.36	56087.41	0.73	0.88	70.27	24.36	4.49	NA
4	beta	23	p3	59513.58	56005.93	0.67	6.30	45.33	0.85	47.52	NA
4	spline	16	p1	60075.44	56562.40	0.67	4.69	29.93	64.16	1.23	NA
4	spline	20	p2	60467.81	56615.81	0.74	8.03	0.00	18.56	73.42	NA
4	spline	24	p3	59507.48	56002.90	0.67	6.15	45.52	0.88	47.45	NA
5	linear	16	p1	60042.45	57012.94	0.58	2.65	42.22	46.75	7.72	0.65
5	linear	21	p2	59736.13	56378.38	0.64	3.92	16.98	55.67	0.42	23.01
5	linear	26	p3	59609.50	56480.04	0.60	5.53	20.09	45.56	27.31	1.50
5	beta	18	p1	59917.73	56881.53	0.58	2.77	7.26	44.41	44.41	1.15
5	beta	23	p2	59638.55	56284.52	0.64	0.92	22.24	55.74	17.17	3.92
5	beta	28	p3	59247.42	56091.90	0.61	4.65	39.45	21.59	0.92	33.38
5	spline	19	p1	59910.05	56872.01	0.58	2.80	7.07	44.45	44.45	1.23
5	spline	24	p2	59630.56	56250.23	0.65	3.80	56.70	16.87	1.00	21.63
5	spline	29	p3	59239.72	56084.95	0.61	4.73	21.24	40.22	0.92	32.89

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 14. MVC. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	106676.49	106676.49	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	106679.60	106679.60	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	106418.23	106418.23	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	105643.68	105643.68	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	105651.39	105651.39	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	105387.58	105387.58	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	105636.15	105636.15	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	105643.87	105643.87	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	105380.82	105380.82	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	103127.91	98693.40	0.85	87.25	12.75	NA	NA	NA
2	linear	9	p2	102964.86	98419.05	0.87	87.01	12.99	NA	NA	NA
2	linear	11	p3	102673.76	98251.39	0.85	86.13	13.87	NA	NA	NA
2	beta	9	p1	101999.73	97548.51	0.86	84.56	15.44	NA	NA	NA
2	beta	11	p2	101737.37	97150.75	0.88	85.02	14.98	NA	NA	NA
2	beta	13	p3	101459.54	97002.30	0.86	82.02	17.98	NA	NA	NA
2	spline	10	p1	101989.27	97531.38	0.86	84.71	15.29	NA	NA	NA
2	spline	12	p2	101726.24	97138.08	0.88	84.90	15.10	NA	NA	NA
2	spline	14	p3	101449.15	96990.85	0.86	81.98	18.02	NA	NA	NA
3	linear	10	p1	101716.75	98006.32	0.71	59.32	37.38	3.30	NA	NA
3	linear	13	p2	101564.34	97585.26	0.76	70.30	27.20	2.50	NA	NA
3	linear	16	p3	101204.62	97314.06	0.75	63.62	34.19	2.19	NA	NA
3	beta	12	p1	100843.83	97110.61	0.72	64.35	30.96	4.69	NA	NA
3	**beta**	**15**	**p2**	**100599.89**	**96535.38**	**0.78**	**73.18**	**23.51**	**3.30**	**NA**	**NA**
3	beta	18	p3	100288.74	96385.86	0.75	65.16	31.19	3.65	NA	NA
3	spline	13	p1	100838.92	97095.78	0.72	64.50	30.81	4.69	NA	NA
3	spline	16	p2	100593.65	96608.36	0.77	70.15	25.82	4.03	NA	NA
3	spline	19	p3	101488.47	97030.18	0.86	81.98	0.00	18.02	NA	NA
4	linear	13	p1	101085.80	97442.80	0.70	5.57	14.75	76.83	2.84	NA
4	linear	17	p2	100864.79	96892.04	0.76	3.84	16.17	77.99	2.00	NA
4	linear	21	p3	101243.94	97353.38	0.75	63.62	0.00	34.19	2.19	NA
4	beta	15	p1	100266.60	96640.96	0.70	4.11	16.63	75.49	3.76	NA
4	**beta**	**19**	**p2**	**99974.50**	**96023.76**	**0.76**	**3.15**	**17.63**	**76.26**	**2.96**	**NA**
4	beta	23	p3	99921.26	96259.93	0.70	57.86	4.19	35.61	2.34	NA
4	spline	16	p1	100262.15	96630.58	0.70	4.07	75.30	16.90	3.73	NA
4	spline	20	p2	100625.10	96639.81	0.77	70.15	25.82	0.00	4.03	NA
4	spline	24	p3	99589.90	95731.79	0.74	3.84	73.53	19.59	3.03	NA
5	linear	16	p1	100553.71	96986.80	0.69	5.49	75.57	15.02	3.34	0.58
5	linear	21	p2	100644.82	97458.73	0.61	3.42	9.45	65.46	20.02	1.65
5	linear	26	p3	99787.85	96028.89	0.72	4.00	20.05	72.49	2.92	0.54
5	beta	18	p1	99995.70	96469.92	0.68	4.15	73.95	16.67	3.92	1.31
5	beta	23	p2	99718.65	96014.24	0.71	2.88	74.64	16.52	4.11	1.84
5	beta	28	p3	105489.91	103962.99	0.29	63.20	0.00	0.00	0.00	36.80
5	spline	19	p1	100027.20	96770.04	0.63	3.38	65.73	22.97	6.19	1.73
5	spline	24	p2	99705.82	95976.41	0.72	3.19	75.45	15.33	4.00	2.04
5	spline	29	p3	99207.07	95578.60	0.70	1.61	5.80	70.15	19.44	3.00

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 15. Phosphate. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	55519.66	55519.66	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	55357.75	55357.75	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	54859.94	54859.94	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	53382.92	53382.92	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	53249.56	53249.56	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	52655.75	52655.75	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	*	*	*	*	*	*	*	*
1	spline	8	p2	53211.38	53211.38	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	52616.65	52616.65	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	54001.17	48945.92	0.99	3.56	96.44	NA	NA	NA
2	linear	9	p2	53507.78	48440.48	0.99	3.09	96.91	NA	NA	NA
2	linear	11	p3	52918.69	47850.82	0.99	2.98	97.02	NA	NA	NA
2	beta	9	p1	52332.67	47583.35	0.93	6.46	93.54	NA	NA	NA
2	**beta**	**11**	**p2**	**51944.67**	**47323.01**	**0.91**	**9.44**	**90.56**	**NA**	**NA**	**NA**
2	**beta**	**13**	**p3**	**51266.20**	**46502.44**	**0.93**	**8.30**	**91.70**	**NA**	**NA**	**NA**
2	spline	10	p1	52305.50	47642.49	0.91	7.56	92.44	NA	NA	NA
2	**spline**	**12**	**p2**	**51917.90**	**47378.52**	**0.89**	**10.30**	**89.70**	**NA**	**NA**	**NA**
2	**spline**	**14**	**p3**	**51236.95**	**46525.70**	**0.92**	**8.85**	**91.15**	**NA**	**NA**	**NA**
3	linear	10	p1	53499.36	49884.77	0.71	2.74	19.62	77.63	NA	NA
3	linear	13	p2	52942.33	49335.96	0.71	2.19	56.72	41.09	NA	NA
3	linear	16	p3	52297.81	48580.26	0.73	2.27	33.10	64.63	NA	NA
3	beta	12	p1	51953.53	48305.55	0.71	4.00	78.81	17.20	NA	NA
3	**beta**	**15**	**p2**	**51661.81**	**47076.68**	**0.90**	**3.02**	**6.78**	**90.21**	**NA**	**NA**
3	**beta**	**18**	**p3**	**50668.06**	**47009.61**	**0.72**	**49.98**	**43.56**	**6.46**	**NA**	**NA**
3	spline	13	p1	51943.82	48329.33	0.71	4.19	77.87	17.94	NA	NA
3	**spline**	**16**	**p2**	**51642.01**	**47099.31**	**0.89**	**7.21**	**89.74**	**3.06**	**NA**	**NA**
3	**spline**	**19**	**p3**	**50639.30**	**46952.64**	**0.72**	**6.62**	**54.95**	**38.43**	**NA**	**NA**
4	linear	13	p1	53253.23	49642.22	0.71	2.12	79.55	1.37	16.96	NA
4	linear	17	p2	52696.57	49119.32	0.70	0.51	2.19	49.90	47.40	NA
4	linear	21	p3	51919.44	48280.68	0.71	0.82	3.68	25.73	69.76	NA
4	beta	15	p1	51724.40	48181.73	0.69	2.55	77.67	3.02	16.76	NA
4	beta	19	p2	51288.54	47689.05	0.70	2.55	4.39	76.26	16.80	NA
4	beta	23	p3	50385.88	46701.39	0.72	1.72	6.31	60.95	31.02	NA
4	spline	16	p1	51718.97	48212.87	0.69	2.66	3.09	76.15	18.10	NA
4	spline	20	p2	51200.73	48107.63	0.61	3.02	17.39	50.14	29.46	NA
4	spline	24	p3	50359.33	46656.69	0.73	61.77	6.35	30.16	1.72	NA
5	linear	16	p1	53106.28	49277.75	0.75	1.41	0.55	3.56	87.39	7.09
5	linear	21	p2	52727.95	49150.70	0.70	0.51	2.19	0.00	49.90	47.40
5	linear	26	p3	51502.69	48252.82	0.64	1.68	0.67	11.16	44.61	41.87
5	beta	18	p1	51673.57	48392.01	0.64	2.98	3.21	2.31	83.90	7.60
5	beta	23	p2	51011.09	47687.49	0.65	2.70	0.90	66.04	20.17	10.18
5	beta	28	p3	50070.45	46763.49	0.65	0.71	2.78	39.87	16.45	40.19
5	spline	19	p1	51663.40	48400.27	0.64	2.86	2.31	2.82	8.97	83.04
5	spline	24	p2	51001.44	47679.52	0.65	2.70	0.90	65.18	20.56	10.65
5	spline	29	p3	50710.83	47194.17	0.69	30.79	8.97	59.54	0.27	0.43

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 16. Platelete count. Results of the LCGA exploratory analysis. Bold marks the selected models for

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	295446.9	295446.9	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	295408.0	295408.0	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	292618.9	292618.9	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	275230.2	275230.2	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	275208.7	275208.7	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	272207.7	272207.7	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	278994.3	278994.3	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	279001.9	279001.9	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	276091.6	276091.6	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	286125.9	281151.7	0.96	8.14	91.86	NA	NA	NA
2	linear	9	p2	286057.1	281036.7	0.96	92.16	7.84	NA	NA	NA
2	linear	11	p3	*	*	*	*	*	*	*	*
2	beta	9	p1	269465.3	264897.4	0.88	84.17	15.83	NA	NA	NA
2	beta	11	p2	268879.5	264054.7	0.93	85.09	14.91	NA	NA	NA
2	**beta**	**13**	**p3**	**264072.2**	**260128.1**	**0.76**	**61.51**	**38.49**	**NA**	**NA**	**NA**
2	spline	10	p1	271812.1	267025.5	0.92	87.98	12.02	NA	NA	NA
2	spline	12	p2	271541.5	266611.2	0.95	88.40	11.60	NA	NA	NA
2	**spline**	**14**	**p3**	**266377.7**	**262288.7**	**0.79**	**80.56**	**19.44**	**NA**	**NA**	**NA**
3	linear	10	p1	283165.3	278743.8	NaN	4.07	84.02	11.87	NA	NA
3	linear	13	p2	283008.7	278663.6	NaN	82.67	13.33	3.96	NA	NA
3	linear	16	p3	*	*	*	*	*	*	*	*
3	beta	12	p1	267189.9	263538.3	0.70	27.43	62.58	9.99	NA	NA
3	beta	15	p2	266704.1	262862.5	0.74	62.35	29.27	8.37	NA	NA
3	**beta**	**18**	**p3**	**264111.5**	**260167.4**	**0.76**	**61.51**	**0.00**	**38.49**	**NA**	**NA**
3	spline	13	p1	269475.5	265722.8	0.72	68.31	23.47	8.22	NA	NA
3	spline	16	p2	269141.7	265088.0	0.78	72.45	21.86	5.69	NA	NA
3	**spline**	**19**	**p3**	**262189.2**	**258710.9**	**0.67**	**17.52**	**70.42**	**12.06**	**NA**	**NA**
4	linear	13	p1	281367.9	277281.6	NaN	0.92	80.29	5.76	12.99	NA
4	linear	17	p2	281744.3	278088.4	NaN	77.18	7.53	3.03	12.22	NA
4	linear	21	p3	*	*	*	*	*	*	*	*
4	beta	15	p1	266488.6	263120.8	0.65	9.87	44.26	42.26	3.61	NA
4	beta	19	p2	265887.2	262680.2	0.62	49.87	11.60	7.26	31.27	NA
4	beta	23	p3	258909.6	255715.9	0.61	7.95	20.09	62.27	9.68	NA
4	spline	16	p1	268387.5	264842.1	0.68	7.99	63.89	3.11	25.01	NA
4	spline	20	p2	268028.8	264528.9	0.67	57.28	30.46	8.87	3.38	NA
4	spline	24	p3	260700.3	257707.4	0.57	9.87	57.86	21.28	10.99	NA
5	linear	16	p1	280582.5	276929.8	NaN	0.92	75.68	5.69	14.25	3.42
5	linear	21	p2	281687.7	277969.7	NaN	76.68	0.00	13.75	2.46	7.07
5	linear	26	p3	277173.2	274242.5	NaN	38.03	0.00	48.37	11.37	2.19
5	beta	18	p1	266028.1	263067.1	0.57	9.76	17.40	8.03	61.77	3.03
5	beta	23	p2	265069.0	262111.4	0.57	26.78	6.49	18.02	4.03	44.68
5	beta	28	p3	258144.3	255200.5	0.57	5.26	16.90	58.16	8.68	10.99
5	spline	19	p1	267938.5	264853.5	0.59	7.68	7.18	12.87	2.96	69.30
5	spline	24	p2	267954.1	265109.2	0.55	55.17	4.53	29.04	5.30	5.95
5	spline	29	p3	260582.5	257547.4	0.58	10.07	0.00	46.95	33.73	9.26

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 17. Potassium. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	30312.29	30312.29	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	30120.97	30120.97	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	29742.75	29742.75	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	28231.70	28231.70	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	28063.18	28063.18	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	27717.38	27717.38	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	28201.18	28201.18	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	28034.74	28034.74	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	27692.15	27692.15	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	29175.55	24325.06	0.93	8.07	91.93	NA	NA	NA
2	linear	9	p2	28621.30	23713.95	0.94	8.30	91.70	NA	NA	NA
2	linear	11	p3	28186.86	23223.47	0.95	7.65	92.35	NA	NA	NA
2	beta	9	p1	27234.58	23178.49	0.78	24.98	75.02	NA	NA	NA
2	beta	11	p2	26789.86	22373.30	0.85	17.45	82.55	NA	NA	NA
2	**beta**	**13**	**p3**	**26409.39**	**21940.37**	**0.86**	**16.10**	**83.90**	**NA**	**NA**	**NA**
2	spline	10	p1	27210.63	23140.19	0.78	23.52	76.48	NA	NA	NA
2	spline	12	p2	26836.34	22762.64	0.78	34.17	65.83	NA	NA	NA
2	**spline**	**14**	**p3**	**26411.13**	**22114.33**	**0.83**	**79.86**	**20.14**	**NA**	**NA**	**NA**
3	linear	10	p1	28634.65	25127.93	0.67	5.07	27.63	67.29	NA	NA
3	linear	13	p2	28319.95	24410.54	0.75	4.11	23.21	72.67	NA	NA
3	linear	16	p3	27520.04	23849.11	0.71	36.63	5.84	57.53	NA	NA
3	beta	12	p1	26881.05	23517.56	0.65	6.61	49.92	43.47	NA	NA
3	beta	15	p2	26401.06	22937.43	0.67	63.07	12.41	24.52	NA	NA
3	**beta**	**18**	**p3**	**25793.24**	**22112.75**	**0.71**	**27.59**	**62.76**	**9.65**	**NA**	**NA**
3	spline	13	p1	26860.11	23490.30	0.65	6.15	51.46	42.39	NA	NA
3	spline	16	p2	26411.28	23034.93	0.65	13.45	22.83	63.72	NA	NA
3	**spline**	**19**	**p3**	**25790.83**	**22099.41**	**0.71**	**9.07**	**26.29**	**64.64**	**NA**	**NA**
4	linear	13	p1	28415.99	25003.81	0.66	5.07	0.92	38.93	55.07	NA
4	linear	17	p2	27710.15	24238.53	0.67	13.72	5.11	75.98	5.19	NA
4	linear	21	p3	27164.90	23457.35	0.71	2.50	14.68	77.21	5.61	NA
4	beta	15	p1	26738.72	23391.98	0.64	6.76	58.84	33.74	0.65	NA
4	beta	19	p2	26073.96	23041.74	0.58	14.22	10.53	64.37	10.88	NA
4	beta	23	p3	25494.95	22263.94	0.62	46.73	6.07	10.45	36.74	NA
4	spline	16	p1	26713.97	23357.39	0.65	6.34	0.65	59.68	33.32	NA
4	spline	20	p2	26212.11	23406.02	0.54	15.14	6.92	58.99	18.95	NA
4	spline	24	p3	25509.03	22319.65	0.61	5.92	48.00	34.86	11.22	NA
5	linear	16	p1	28289.70	24888.08	0.65	1.58	0.92	3.07	40.70	53.73
5	linear	21	p2	27573.59	23997.99	0.69	1.50	4.80	77.67	10.91	5.11
5	linear	26	p3	26891.09	23140.15	0.72	11.45	3.00	79.36	3.38	2.81
5	beta	18	p1	26665.81	23339.78	0.64	1.61	37.55	5.11	55.07	0.65
5	beta	23	p2	25923.09	22895.60	0.58	15.53	6.34	18.22	1.73	58.19
5	beta	28	p3	25534.27	22303.26	0.62	46.73	6.07	0.00	10.45	36.74
5	spline	19	p1	26651.86	23508.59	0.60	4.46	0.65	43.20	4.15	47.54
5	spline	24	p2	25910.65	22889.10	0.58	6.23	15.03	19.37	57.69	1.69
5	spline	29	p3	25298.05	21826.35	0.67	5.73	8.72	76.10	5.00	4.46

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 18. RDW. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	58409.39	58409.39	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	58380.52	58380.52	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	58385.69	58385.69	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	47569.39	47569.39	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	47442.19	47442.19	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	47449.00	47449.00	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	47887.62	47887.62	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	47772.98	47772.98	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	47780.49	47780.49	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	*	*	*	*	*	*	*	*
2	linear	9	p2	*	*	*	*	*	*	*	*
2	linear	11	p3	*	*	*	*	*	*	*	*
2	**beta**	**9**	**p1**	**42329.51**	**37893.27**	**0.85**	**84.59**	**15.41**	**NA**	**NA**	**NA**
2	**beta**	**11**	**p2**	**41494.01**	**36759.63**	**0.91**	**85.90**	**14.10**	**NA**	**NA**	**NA**
2	beta	13	p3	41506.64	36760.54	0.91	85.94	14.06	NA	NA	NA
2	**spline**	**10**	**p1**	**42494.89**	**38004.28**	**0.86**	**85.75**	**14.25**	**NA**	**NA**	**NA**
2	**spline**	**12**	**p2**	**41739.74**	**36988.80**	**0.91**	**86.52**	**13.48**	**NA**	**NA**	**NA**
2	spline	14	p3	41753.81	36999.72	0.91	86.52	13.48	NA	NA	NA
3	linear	10	p1	*	*	*	*	*	*	*	*
3	linear	13	p2	*	*	*	*	*	*	*	*
3	linear	16	p3	*	*	*	*	*	*	*	*
3	**beta**	**12**	**p1**	**40427.29**	**36810.18**	**0.69**	**68.04**	**22.63**	**9.34**	**NA**	**NA**
3	**beta**	**15**	**p2**	**39529.38**	**35646.28**	**0.75**	**70.11**	**21.94**	**7.95**	**NA**	**NA**
3	beta	18	p3	39533.69	35640.57	0.75	70.96	21.21	7.84	NA	NA
3	**spline**	**13**	**p1**	**40563.04**	**36860.20**	**0.71**	**70.53**	**8.22**	**21.24**	**NA**	**NA**
3	**spline**	**16**	**p2**	**39712.54**	**35837.21**	**0.74**	**70.11**	**22.47**	**7.41**	**NA**	**NA**
3	spline	19	p3	39713.91	35843.97	0.74	70.38	21.97	7.65	NA	NA
4	linear	13	p1	*	*	*	*	*	*	*	*
4	linear	17	p2	*	*	*	*	*	*	*	*
4	linear	21	p3	*	*	*	*	*	*	*	*
4	beta	15	p1	39053.07	35570.99	0.67	7.53	30.46	58.51	3.50	NA
4	beta	19	p2	38723.55	35387.13	0.64	36.38	8.22	49.17	6.22	NA
4	beta	23	p3	38082.95	34457.40	0.70	67.08	22.55	5.42	4.96	NA
4	spline	16	p1	39175.83	35740.42	0.66	8.11	54.40	34.08	3.42	NA
4	spline	20	p2	38944.17	35634.41	0.64	52.63	8.34	32.23	6.80	NA
4	spline	24	p3	38254.97	34592.20	0.70	68.65	5.30	21.55	4.49	NA
5	linear	16	p1	*	*	*	*	*	*	*	*
5	linear	21	p2	*	*	*	*	*	*	*	*
5	linear	26	p3	*	*	*	*	*	*	*	*
5	beta	18	p1	38526.61	35349.00	0.61	11.06	6.42	8.57	70.38	3.57
5	beta	23	p2	37728.12	34333.89	0.65	27.97	3.42	58.36	7.34	2.92
5	beta	28	p3	37348.78	34019.31	0.64	12.33	5.46	8.26	70.23	3.73
5	spline	19	p1	38622.15	35418.98	0.62	71.23	6.45	8.03	11.06	3.23
5	spline	24	p2	38286.20	34976.12	0.64	9.99	4.30	7.65	72.42	5.65
5	spline	29	p3	41077.26	37825.40	0.62	69.73	5.92	0.00	9.22	15.14

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 19. Red Blood Cells. Results of the LCGA exploratory analysis. Bold marks the selected models for

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	22886.53	22886.53	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	21875.89	21875.89	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	21407.21	21407.21	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	22449.18	22449.18	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	21511.41	21511.41	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	21085.81	21085.81	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	22440.65	22440.65	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	21507.13	21507.13	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	21083.04	21083.04	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	21132.46	16960.87	0.80	46.41	53.59	NA	NA	NA
2	linear	9	p2	19694.14	15442.45	0.82	36.34	63.66	NA	NA	NA
2	**linear**	**11**	**p3**	**19317.61**	**14933.21**	**0.84**	**72.69**	**27.31**	**NA**	**NA**	**NA**
2	beta	9	p1	20744.74	16624.90	0.79	51.98	48.02	NA	NA	NA
2	beta	11	p2	19420.34	15262.45	0.80	39.80	60.20	NA	NA	NA
2	beta	13	p3	19050.45	14743.74	0.83	71.26	28.74	NA	NA	NA
2	spline	10	p1	20741.52	16628.92	0.79	51.90	48.10	NA	NA	NA
2	spline	12	p2	19419.92	15265.86	0.80	39.49	60.51	NA	NA	NA
2	spline	14	p3	19049.58	14744.33	0.83	71.11	28.89	NA	NA	NA
3	linear	10	p1	20400.29	16747.88	0.70	11.41	41.68	46.91	NA	NA
3	linear	13	p2	18894.07	15108.78	0.73	37.11	10.91	51.98	NA	NA
3	**linear**	**16**	**p3**	**17811.97**	**13905.77**	**0.75**	**52.55**	**26.78**	**20.67**	**NA**	**NA**
3	beta	12	p1	20165.93	16637.25	0.68	13.37	27.78	58.86	NA	NA
3	beta	15	p2	18754.45	15181.46	0.69	37.30	22.70	39.99	NA	NA
3	beta	18	p3	17573.48	13710.83	0.74	47.64	31.73	20.63	NA	NA
3	spline	13	p1	20130.18	16358.32	0.72	7.41	52.55	40.03	NA	NA
3	spline	16	p2	18693.05	14945.04	0.72	11.18	38.92	49.90	NA	NA
3	spline	19	p3	17575.73	13721.10	0.74	31.54	47.83	20.63	NA	NA
4	linear	13	p1	20012.17	16554.95	0.66	8.53	23.97	57.74	9.76	NA
4	linear	17	p2	18454.90	15145.01	0.64	37.84	10.60	19.94	31.62	NA
4	**linear**	**21**	**p3**	**17396.20**	**13716.27**	**0.71**	**50.83**	**13.98**	**13.95**	**21.24**	**NA**
4	beta	15	p1	19760.85	16392.32	0.65	7.57	30.00	54.05	8.37	NA
4	beta	19	p2	18247.25	14970.22	0.63	20.09	13.02	48.91	17.98	NA
4	beta	23	p3	17020.86	13460.07	0.68	37.46	27.51	12.02	23.01	NA
4	spline	16	p1	19760.45	16400.68	0.65	7.49	30.00	54.25	8.26	NA
4	spline	20	p2	18248.96	14980.00	0.63	13.18	48.75	20.25	17.83	NA
4	spline	24	p3	17054.40	13499.92	0.68	26.51	46.25	7.84	19.40	NA
5	linear	16	p1	19809.50	16523.62	0.63	5.15	30.23	49.71	12.45	2.46
5	linear	21	p2	18098.62	14870.52	0.62	43.14	9.49	21.48	23.01	2.88
5	linear	26	p3	16825.73	13446.11	0.65	37.57	15.41	10.95	26.66	9.41
5	beta	18	p1	19624.61	16333.24	0.63	4.34	32.42	4.96	53.86	4.42
5	beta	23	p2	18283.02	14907.09	0.65	63.43	11.22	0.00	12.64	12.72
5	beta	28	p3	16643.59	13301.86	0.64	5.95	23.47	10.45	40.03	20.09
5	spline	19	p1	19606.68	16392.88	0.62	4.84	8.95	49.75	33.38	3.07
5	spline	24	p2	17902.02	14743.25	0.61	15.10	21.44	41.95	17.94	3.57
5	spline	29	p3	16685.29	13394.19	0.63	16.71	35.57	12.14	25.89	9.68

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 20. Sodium. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	139759.4	139759.4	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	139759.3	139759.3	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	139697.8	139697.8	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	137902.3	137902.3	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	137907.7	137907.7	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	137858.5	137858.5	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	138354.1	138354.1	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	138358.6	138358.6	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	138303.0	138303.0	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	137139.1	132960.9	0.80	87.46	12.54	NA	NA	NA
2	linear	9	p2	136836.4	132311.0	0.87	88.81	11.19	NA	NA	NA
2	linear	11	p3	136713.2	132172.3	0.87	89.19	10.81	NA	NA	NA
2	beta	9	p1	135078.2	131082.8	0.77	80.92	19.08	NA	NA	NA
2	beta	11	p2	134799.7	130582.1	0.81	80.27	19.73	NA	NA	NA
2	beta	13	p3	134685.2	130445.3	0.82	82.31	17.69	NA	NA	NA
2	spline	10	p1	135613.6	131523.2	0.79	83.73	16.27	NA	NA	NA
2	spline	12	p2	135322.6	131028.0	0.83	82.85	17.15	NA	NA	NA
2	spline	14	p3	135203.3	130863.8	0.83	84.38	15.62	NA	NA	NA
3	linear	10	p1	135886.4	132044.8	0.74	13.73	3.19	83.08	NA	NA
3	linear	13	p2	135472.5	130984.6	0.86	88.12	7.69	4.19	NA	NA
3	linear	16	p3	135030.6	130271.4	0.92	4.46	89.04	6.50	NA	NA
3	beta	12	p1	133855.6	130122.4	0.72	17.85	78.42	3.73	NA	NA
3	**beta**	**15**	**p2**	**133582.1**	**129625.5**	**0.76**	**10.77**	**80.81**	**8.42**	**NA**	**NA**
3	beta	18	p3	133150.6	129602.4	0.68	60.85	26.08	13.08	NA	NA
3	spline	13	p1	134286.2	130475.6	0.73	16.00	79.54	4.46	NA	NA
3	**spline**	**16**	**p2**	**133937.5**	**129724.9**	**0.81**	**10.42**	**84.77**	**4.81**	**NA**	**NA**
3	spline	19	p3	133583.8	129266.2	0.83	8.58	83.62	7.81	NA	NA
4	linear	13	p1	135319.1	131666.5	0.70	5.77	2.19	84.38	7.65	NA
4	linear	17	p2	134314.9	129895.2	0.85	6.15	6.23	85.08	2.54	NA
4	linear	21	p3	134132.8	130192.2	0.76	5.27	4.27	5.77	84.69	NA
4	beta	15	p1	133314.9	129923.5	0.65	6.81	10.35	79.38	3.46	NA
4	**beta**	**19**	**p2**	**132243.5**	**128201.7**	**0.78**	**7.04**	**78.19**	**4.00**	**10.77**	**NA**
4	beta	23	p3	132088.1	128393.3	0.71	6.15	6.35	8.46	79.04	NA
4	spline	16	p1	133682.4	130011.7	0.71	8.04	82.73	7.15	2.08	NA
4	**spline**	**20**	**p2**	**132668.3**	**128574.3**	**0.79**	**6.62**	**10.38**	**79.62**	**3.38**	**NA**
4	spline	24	p3	132393.5	128163.3	0.81	6.69	6.69	4.81	81.81	NA
5	linear	16	p1	135008.8	131725.1	0.63	1.50	1.88	14.96	77.65	4.00
5	linear	21	p2	133778.4	129919.5	0.74	4.46	81.35	5.27	5.92	3.00
5	linear	26	p3	133012.6	129111.7	0.75	10.27	6.19	78.92	2.85	1.77
5	beta	18	p1	133074.5	129938.6	0.60	2.62	13.96	6.50	74.65	2.27
5	beta	23	p2	132234.5	128187.5	0.78	10.00	7.04	78.38	3.38	1.19
5	beta	28	p3	131255.5	127404.7	0.74	3.12	6.23	77.73	6.35	6.58
5	spline	19	p1	133360.0	130157.8	0.62	3.23	2.88	62.12	30.12	1.65
5	spline	24	p2	132062.8	127868.6	0.81	5.08	6.88	81.54	3.58	2.92
5	spline	29	p3	131610.4	127490.2	0.79	5.15	7.88	81.54	1.96	3.46

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 21. Urea Nitrogen. Results of the LCGA exploratory analysis. Bold marks the selected models for

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	176644.1	176644.1	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	176567.7	176567.7	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	176350.6	176350.6	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	156828.2	156828.2	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	156695.3	156695.3	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	156261.1	156261.1	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	159486.5	159486.5	1.00	100.00	NA	NA	NA	NA
1	spline	8	p2	*	*	*	*	*	*	*	*
1	spline	9	p3	158963.5	158963.5	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	171182.4	166051.9	0.99	95.19	4.81	NA	NA	NA
2	linear	9	p2	170417.0	165263.7	0.99	96.62	3.38	NA	NA	NA
2	linear	11	p3	170158.9	165011.9	0.99	96.31	3.69	NA	NA	NA
2	beta	9	p1	153608.1	149746.3	0.74	60.96	39.04	NA	NA	NA
2	**beta**	**11**	**p2**	**152547.9**	**148141.5**	**0.85**	**77.69**	**22.31**	**NA**	**NA**	**NA**
2	beta	13	p3	152206.1	147882.4	0.83	75.08	24.92	NA	NA	NA
2	spline	10	p1	156382.3	152304.3	0.78	81.08	18.92	NA	NA	NA
2	spline	12	p2	155361.1	150726.2	0.89	85.42	14.58	NA	NA	NA
2	spline	14	p3	155038.5	150509.7	0.87	84.15	15.85	NA	NA	NA
3	linear	10	p1	169094.8	163999.2	0.98	2.88	92.73	4.38	NA	NA
3	linear	13	p2	167968.9	162835.5	0.99	3.08	93.65	3.27	NA	NA
3	linear	16	p3	167697.2	162572.5	0.99	2.88	93.69	3.42	NA	NA
3	beta	12	p1	152401.4	148854.9	0.68	24.65	66.12	9.23	NA	NA
3	**beta**	**15**	**p2**	**151218.8**	**147380.0**	**0.74**	**61.96**	**29.92**	**8.12**	**NA**	**NA**
3	beta	18	p3	150631.0	146706.3	0.75	55.42	25.27	19.31	NA	NA
3	spline	13	p1	154689.4	150868.9	0.73	21.54	72.77	5.69	NA	NA
3	spline	16	p2	153523.3	149623.0	0.75	35.65	59.77	4.58	NA	NA
3	spline	19	p3	153274.9	149424.2	0.74	30.27	63.77	5.96	NA	NA
4	linear	13	p1	167575.1	162549.4	0.97	3.12	1.62	91.65	3.62	NA
4	linear	17	p2	166519.6	161526.6	0.96	2.42	4.62	90.73	2.23	NA
4	linear	21	p3	167736.5	162611.8	0.99	2.88	0.00	93.69	3.42	NA
4	beta	15	p1	151846.7	148740.1	0.60	10.77	41.00	42.42	5.81	NA
4	**beta**	**19**	**p2**	**150193.7**	**146605.6**	**0.69**	**10.19**	**23.81**	**59.73**	**6.27**	**NA**
4	beta	23	p3	149472.2	145626.9	0.74	17.73	16.65	59.54	6.08	NA
4	spline	16	p1	153998.0	150453.4	0.68	5.54	4.50	62.81	27.15	NA
4	spline	20	p2	152503.5	148451.5	0.78	8.19	72.73	15.54	3.54	NA
4	spline	24	p3	152022.1	147992.5	0.77	13.92	13.23	69.46	3.38	NA
5	linear	16	p1	166471.3	161480.3	0.96	89.92	1.50	0.73	4.19	3.65
5	linear	21	p2	165877.5	160885.1	0.96	1.81	2.69	4.46	89.46	1.58
5	linear	26	p3	164551.8	159535.1	0.96	5.19	89.54	1.73	2.35	1.19
5	beta	18	p1	151531.5	148602.1	0.56	4.27	59.04	17.23	14.31	5.15
5	beta	23	p2	149454.0	146021.0	0.66	13.54	21.38	15.35	46.54	3.19
5	beta	28	p3	148745.1	145247.4	0.67	20.85	15.92	47.00	8.96	7.27
5	spline	19	p1	153545.4	150371.8	0.61	2.12	5.08	65.12	14.46	13.23
5	spline	24	p2	154464.6	151701.2	0.53	65.38	1.54	14.31	10.31	8.46
5	spline	29	p3	151737.3	148564.4	0.61	14.08	16.46	23.62	40.69	5.15

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 22. White blood cells. Results of the LCGA exploratory analysis. Bold marks the selected models for GMM

k	link	npm	poly	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
1	linear	4	p1	128591.8	128591.8	1.00	100.00	NA	NA	NA	NA
1	linear	5	p2	128551.2	128551.2	1.00	100.00	NA	NA	NA	NA
1	linear	6	p3	127511.4	127511.4	1.00	100.00	NA	NA	NA	NA
1	beta	6	p1	119411.9	119411.9	1.00	100.00	NA	NA	NA	NA
1	beta	7	p2	119378.4	119378.4	1.00	100.00	NA	NA	NA	NA
1	beta	8	p3	118115.0	118115.0	1.00	100.00	NA	NA	NA	NA
1	spline	7	p1	*	*	*	*	*	*	*	*
1	spline	8	p2	121214.8	121214.8	1.00	100.00	NA	NA	NA	NA
1	spline	9	p3	120052.1	120052.1	1.00	100.00	NA	NA	NA	NA
2	linear	7	p1	126485.4	121456.2	0.97	96.50	3.50	NA	NA	NA
2	linear	9	p2	125869.8	120740.4	0.99	96.97	3.03	NA	NA	NA
2	linear	11	p3	124861.6	119769.8	0.98	96.62	3.38	NA	NA	NA
2	beta	9	p1	117299.8	112569.2	0.91	93.39	6.61	NA	NA	NA
2	beta	11	p2	116764.7	111879.8	0.94	92.59	7.41	NA	NA	NA
2	**beta**	**13**	**p3**	**115471.3**	**110605.4**	**0.93**	**93.01**	**6.99**	**NA**	**NA**	**NA**
2	spline	10	p1	119052.7	114125.7	0.95	95.58	4.42	NA	NA	NA
2	spline	12	p2	118452.4	113360.9	0.98	96.27	3.73	NA	NA	NA
2	spline	14	p3	117273.4	112262.7	0.96	95.47	4.53	NA	NA	NA
3	linear	10	p1	125596.3	120900.9	0.90	93.01	1.11	5.88	NA	NA
3	linear	13	p2	124870.8	119753.9	0.98	1.92	95.01	3.07	NA	NA
3	linear	16	p3	123754.2	118660.5	0.98	1.34	95.39	3.27	NA	NA
3	beta	12	p1	116480.6	112942.9	0.68	28.04	68.69	3.27	NA	NA
3	beta	15	p2	115882.6	112012.4	0.74	55.01	41.99	3.00	NA	NA
3	**beta**	**18**	**p3**	**114474.2**	**110666.1**	**0.73**	**64.12**	**33.04**	**2.84**	**NA**	**NA**
3	spline	13	p1	118268.2	114575.9	0.71	19.02	78.03	2.96	NA	NA
3	spline	16	p2	118508.9	113586.6	0.95	93.05	0.42	6.53	NA	NA
3	spline	19	p3	116316.9	112434.8	0.75	24.66	72.84	2.50	NA	NA
4	linear	13	p1	124798.0	120064.2	0.91	1.11	0.77	92.89	5.22	NA
4	linear	17	p2	124043.3	119182.7	0.93	1.11	6.68	91.47	0.73	NA
4	linear	21	p3	122977.1	117926.4	0.97	2.92	0.15	93.55	3.38	NA
4	beta	15	p1	116113.1	112881.3	0.62	4.88	47.56	44.79	2.77	NA
4	beta	19	p2	115723.8	111917.9	0.73	0.77	41.14	55.40	2.69	NA
4	beta	23	p3	114007.9	110804.7	0.62	23.43	32.04	42.45	2.07	NA
4	spline	16	p1	117776.4	113786.5	0.77	8.03	4.99	86.09	0.88	NA
4	spline	20	p2	117247.8	113450.4	0.73	57.20	0.81	39.84	2.15	NA
4	spline	24	p3	115969.8	112733.2	0.62	10.72	21.97	64.58	2.73	NA
5	linear	16	p1	124472.4	120463.7	0.77	0.54	1.00	86.25	9.07	3.15
5	linear	21	p2	123411.4	118582.7	0.93	0.15	2.57	6.49	89.93	0.85
5	linear	26	p3	122479.3	118200.7	0.82	0.92	82.48	12.64	3.23	0.73
5	beta	18	p1	115704.6	112437.7	0.63	5.84	0.61	71.61	18.25	3.69
5	beta	23	p2	114822.8	111267.2	0.68	11.41	5.84	9.14	72.15	1.46
5	beta	28	p3	113833.5	110663.5	0.61	2.15	28.31	33.35	34.00	2.19
5	spline	19	p1	117439.1	113913.8	0.68	2.73	14.56	78.33	3.57	0.81
5	spline	24	p2	118995.4	114540.2	0.86	88.40	0.00	0.00	0.00	11.60
5	spline	29	p3	115449.5	111274.6	0.80	2.38	3.30	83.52	9.30	1.50

k: number of classes, link: the link function, npm: number of parameters, poly: polynomial order, %class: % of the subjects classified as such class by the maximal posterior probability. *Model did not converge.

Annexed Table 23. Results of the GMM model selection process.

Analyte	k	link	np	d	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
Anion Gap	1	spline	18	3	104817.97	104817.97	1.00	100.00	NA	NA	NA	NA
Anion Gap	1	beta	17	3	104836.07	104836.07	1.00	100.00	NA	NA	NA	NA
Anion Gap	2	spline	23	3	104770.61	99661.40	0.98	1.69	98.31	NA	NA	NA
Anion Gap	2	beta	22	3	104781.70	99681.87	0.98	2.00	98.00	NA	NA	NA
Anion Gap	3	spline	28	3	104809.92	101991.79	0.54	2.46	97.54	0.00	NA	NA
Anion Gap	3	beta	27	3	104731.24	99744.00	0.96	2.12	96.19	1.69	NA	NA
Bicarbonate	1	linear	15	3	111559.02	111559.02	1.00	100.00	NA	NA	NA	NA
Bicarbonate	2	linear	20	3	111598.34	108876.42	0.52	74.23	25.77	NA	NA	NA
Bicarbonate	3	linear	25	3	111549.39	108924.95	0.50	42.96	53.81	3.23	NA	NA
Bicarbonate	4	linear	30	3	111494.58	108694.66	0.54	0.00	95.31	1.38	3.31	NA
Calcium, Total	1	linear	10	2	24868.36	24868.36	1.00	100.00	NA	NA	NA	NA
Calcium, Total	2	linear	14	2	24899.74	22238.62	0.52	49.49	50.51	NA	NA	NA
Calcium, Total	3	linear	18	2	24899.08	21743.94	0.62	9.59	90.41	0.00	NA	NA
Chloride	1	spline	18	3	127978.08	127978.08	1.00	100.00	NA	NA	NA	NA
Chloride	1	beta	17	3	127986.17	127986.17	1.00	100.00	NA	NA	NA	NA
Chloride	2	spline	23	3	128017.39	125348.34	0.51	68.77	31.23	NA	NA	NA
Chloride	2	beta	22	3	127905.22	122725.31	1.00	0.85	99.15	NA	NA	NA
Chloride	3	spline	28	3	127751.87	122649.54	0.98	96.65	0.88	2.46	NA	NA
Chloride	3	beta	27	3	127783.88	122676.67	0.98	1.00	96.58	2.42	NA	NA
Chloride	4	spline	33	3	127791.19	124103.64	0.71	0.00	96.23	1.00	2.77	NA
Chloride	4	beta	32	3	127752.49	122659.69	0.98	1.00	0.46	96.54	2.00	NA
Creatinine	1	beta	8	1	*	*	*	*	*	*	*	*
Creatinine	2	beta	11	1	*	*	*	*	*	*	*	*
Creatinine	3	beta	14	1	*	*	*	*	*	*	*	*
Glucose	1	beta	8	1	222910.79	222910.79	1.00	100.00	NA	NA	NA	NA
Glucose	2	beta	11	1	222867.98	218145.30	0.91	10.46	89.54	NA	NA	NA
Glucose	3	beta	14	1	222891.57	219572.77	0.64	12.35	87.65	0.00	NA	NA
Hematocrit	1	linear	15	3	138816.51	138816.51	1.00	100.00	NA	NA	NA	NA
Hematocrit	2	linear	20	3	138642.39	133945.07	0.90	29.98	70.02	NA	NA	NA
Hematocrit	3	linear	25	3	138572.13	134284.57	0.82	31.93	13.51	54.56	NA	NA
Hematocrit	4	linear	30	3	138568.02	134292.21	0.82	13.25	30.82	55.40	0.54	NA
Hemoglobin	1	linear	15	3	64883.30	64883.30	1.00	100.00	NA	NA	NA	NA
Hemoglobin	2	linear	20	3	64756.12	60125.12	0.89	27.39	72.61	NA	NA	NA
Hemoglobin	3	linear	25	3	64672.17	60404.33	0.82	26.74	11.37	61.89	NA	NA
Hemoglobin	4	linear	30	3	64685.26	60438.74	0.82	26.24	12.52	60.43	0.81	NA
Magnesium	1	spline	18	3	-4799.88	-4799.88	1.00	100.00	NA	NA	NA	NA
Magnesium	1	beta	17	3	-4824.88	-4824.88	1.00	100.00	NA	NA	NA	NA
Magnesium	2	spline	23	3	-4861.69	-9859.52	0.97	1.95	98.05	NA	NA	NA
Magnesium	2	beta	22	3	-4785.64	-7440.29	0.52	48.61	51.39	NA	NA	NA
Magnesium	3	spline	28	3	-4867.33	-9682.87	0.94	0.90	97.19	1.91	NA	NA
Magnesium	3	beta	27	3	-4897.79	-9695.65	0.94	2.26	1.48	96.25	NA	NA
Magnesium	4	spline	33	3	-4915.25	-9683.73	0.93	95.67	1.44	0.94	1.95	NA
Magnesium	4	beta	32	3	-4858.55	-9353.10	0.88	2.38	96.14	0.00	1.48	NA
MCH	1	linear	10	2	47087.22	47087.22	1.00	100.00	NA	NA	NA	NA

Annexed Table 23. Results of the GMM model selection process.

Analyte	k	link	np	d	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
MCH	2	linear	14	2	47118.68	44445.00	0.51	53.32	46.68	NA	NA	NA
MCH	3	linear	18	2	46805.77	41775.40	0.97	94.97	1.84	3.19	NA	NA
MCH	4	linear	22	2	46827.94	42321.11	0.87	0.00	96.85	0.61	2.54	NA
MCH	5	linear	26	2	46757.24	41813.73	0.95	0.19	92.62	0.58	3.03	3.57
MCHC	1	linear	10	2	59151.98	59151.98	1.00	100.00	NA	NA	NA	NA
MCHC	1	linear	15	3	58185.87	58185.87	1.00	100.00	NA	NA	NA	NA
MCHC	2	linear	14	2	59109.67	54145.65	0.95	96.54	3.46	NA	NA	NA
MCHC	2	linear	20	3	58225.20	55536.51	0.52	51.63	48.37	NA	NA	NA
MCHC	3	linear	18	2	59141.13	54951.47	0.80	95.85	0.00	4.15	NA	NA
MCHC	3	linear	25	3	58140.08	53821.44	0.83	56.32	41.11	2.57	NA	NA
MCHC	4	linear	22	2	59172.59	56698.34	0.48	0.00	0.00	94.05	5.95	NA
MCHC	4	linear	30	3	58131.43	53803.71	0.83	41.30	55.94	0.15	2.61	NA
MCV	1	beta	12	2	97882.11	97882.11	1.00	100.00	NA	NA	NA	NA
MCV	2	beta	16	2	97687.94	92518.59	0.99	98.92	1.08	NA	NA	NA
MCV	3	beta	20	2	97660.56	92495.81	0.99	0.15	98.81	1.04	NA	NA
MCV	4	beta	24	2	97616.21	92695.80	0.95	0.15	97.50	1.27	1.08	NA
Phosphate	1	spline	13	2	50729.21	50729.21	1.00	100.00	NA	NA	NA	NA
Phosphate	1	spline	18	3	48271.29	48271.29	1.00	100.00	NA	NA	NA	NA
Phosphate	1	beta	12	2	50745.60	50745.60	1.00	100.00	NA	NA	NA	NA
Phosphate	1	beta	17	3	48307.09	48307.09	1.00	100.00	NA	NA	NA	NA
Phosphate	2	spline	17	2	50529.81	45933.01	0.90	6.93	93.07	NA	NA	NA
Phosphate	2	spline	23	3	48032.52	42980.51	0.99	2.70	97.30	NA	NA	NA
Phosphate	2	beta	16	2	50457.54	45400.07	0.99	2.74	97.26	NA	NA	NA
Phosphate	2	beta	22	3	48047.75	42994.38	0.99	2.82	97.18	NA	NA	NA
Phosphate	3	spline	21	2	50287.94	45724.38	0.89	6.62	90.64	2.74	NA	NA
Phosphate	3	spline	28	3	48071.75	43905.32	0.82	2.86	0.00	97.14	NA	NA
Phosphate	3	beta	20	2	50488.92	46445.95	0.79	2.94	97.06	0.00	NA	NA
Phosphate	3	beta	27	3	48086.98	43632.29	0.87	2.90	97.10	0.00	NA	NA
Platelet Count	1	spline	18	3	251815.76	251815.76	1.00	100.00	NA	NA	NA	NA
Platelet Count	1	beta	17	3	248865.19	248865.19	1.00	100.00	NA	NA	NA	NA
Platelet Count	2	spline	23	3	251855.08	249155.70	0.52	9.76	90.24	NA	NA	NA
Platelet Count	2	beta	22	3	248904.51	246232.09	0.51	35.65	64.35	NA	NA	NA
Platelet Count	3	spline	28	3	251688.11	246673.63	0.96	94.93	2.84	2.23	NA	NA
Platelet Count	3	beta	27	3	248772.75	243815.02	0.95	1.96	2.50	95.54	NA	NA
Platelet Count	4	spline	33	3	251429.70	246364.85	0.97	1.08	2.34	96.27	0.31	NA
Platelet Count	4	beta	32	3	248680.38	243663.40	0.96	0.81	2.38	95.70	1.11	NA
Potassium	1	spline	18	3	24050.86	24050.86	1.00	100.00	NA	NA	NA	NA
Potassium	1	beta	17	3	24045.79	24045.79	1.00	100.00	NA	NA	NA	NA
Potassium	2	spline	23	3	24090.18	21388.18	0.52	50.38	49.62	NA	NA	NA
Potassium	2	beta	22	3	23972.63	18936.72	0.97	3.11	96.89	NA	NA	NA
Potassium	3	spline	28	3	23976.61	18999.77	0.96	2.69	96.50	0.81	NA	NA
Potassium	3	beta	27	3	23983.60	19040.70	0.95	2.57	95.77	1.65	NA	NA
RDW	1	spline	9	1	38016.26	38016.26	1.00	100.00	NA	NA	NA	NA
RDW	1	spline	13	2	34764.65	34764.65	1.00	100.00	NA	NA	NA	NA

Annexed Table 23. Results of the GMM model selection process.

Analyte	k	link	np	d	BIC	ICL	APPA	%class1	%class2	%class3	%class4	%class5
RDW	1	beta	8	1	37965.47	37965.47	1.00	100.00	NA	NA	NA	NA
RDW	1	beta	12	2	34742.24	34742.24	1.00	100.00	NA	NA	NA	NA
RDW	2	spline	12	1	38039.85	35347.36	0.52	28.54	71.46	NA	NA	NA
RDW	2	spline	17	2	34796.10	32127.97	0.51	45.14	54.86	NA	NA	NA
RDW	2	beta	11	1	37989.07	35319.56	0.51	43.57	56.43	NA	NA	NA
RDW	2	beta	16	2	34773.70	32062.45	0.52	45.10	54.90	NA	NA	NA
RDW	3	spline	15	1	37636.66	33595.66	0.78	0.00	97.04	2.96	NA	NA
RDW	3	spline	21	2	34230.04	29375.42	0.93	9.18	86.75	4.07	NA	NA
RDW	3	beta	14	1	37425.44	32534.18	0.94	6.57	90.55	2.88	NA	NA
RDW	3	beta	20	2	34238.71	29400.81	0.93	8.99	87.09	3.92	NA	NA
RDW	4	spline	18	1	37460.74	32832.29	0.89	1.31	84.13	11.72	2.84	NA
RDW	4	spline	25	2	34261.49	29873.59	0.84	9.83	0.00	85.94	4.23	NA
RDW	4	beta	17	1	37436.22	32857.33	0.88	1.27	83.71	12.14	2.88	NA
RDW	4	beta	24	2	34247.33	29617.49	0.89	1.38	13.68	80.95	4.00	NA
Red Blood Cells	1	linear	15	3	13975.47	13975.47	1.00	100.00	NA	NA	NA	NA
Red Blood Cells	2	linear	20	3	13874.45	9222.54	0.89	19.36	80.64	NA	NA	NA
Red Blood Cells	3	linear	25	3	13818.27	9586.07	0.81	18.94	10.95	70.11	NA	NA
Red Blood Cells	4	linear	30	3	13808.90	9685.69	0.79	1.65	10.76	58.74	28.85	NA
Sodium	1	spline	13	2	129674.08	129674.08	1.00	100.00	NA	NA	NA	NA
Sodium	1	beta	12	2	129344.20	129344.20	1.00	100.00	NA	NA	NA	NA
Sodium	2	spline	17	2	129373.39	124219.78	0.99	97.50	2.50	NA	NA	NA
Sodium	2	beta	16	2	129149.14	124014.29	0.99	97.35	2.65	NA	NA	NA
Sodium	3	spline	21	2	129404.84	125191.07	0.81	0.00	97.38	2.62	NA	NA
Sodium	3	beta	20	2	128871.87	123743.42	0.99	0.58	96.77	2.65	NA	NA
Sodium	4	spline	25	2	129128.67	125381.07	0.72	0.00	96.50	0.85	2.65	NA
Sodium	4	beta	24	2	128963.67	125063.68	0.75	96.04	1.04	0.00	2.92	NA
Urea Nitrogen	1	beta	12	2	146364.14	146364.14	1.00	100.00	NA	NA	NA	NA
Urea Nitrogen	2	beta	16	2	146168.70	141216.94	0.95	7.65	92.35	NA	NA	NA
Urea Nitrogen	3	beta	20	2	146200.15	141797.33	0.85	8.35	91.65	0.00	NA	NA
Urea Nitrogen	4	beta	24	2	146231.60	143608.18	0.50	10.77	0.00	89.23	0.00	NA
White Blood Cells	1	beta	17	3	110119.89	110119.89	1.00	100.00	NA	NA	NA	NA
White Blood Cells	2	beta	22	3	110159.22	107494.59	0.51	47.64	52.36	NA	NA	NA
White Blood Cells	3	beta	27	3	109955.10	104834.17	0.98	0.85	98.08	1.08	NA	NA
White Blood Cells	4	beta	32	3	109817.90	104661.20	0.99	0.85	98.12	0.88	0.15	NA

Annexed Table 24. Univatiate comparisons between clusters of Anion Gap and Chloride

Characteristic	Anion Gap				Chloride			
	1 N = 52	2 N = 2548	p-value	q-value	1 N = 22	2 N = 2578	p-value	q-value
Age	76 (58, 83)	62 (43, 79)	<0.001	**<0.001**	28 (22, 55)	63 (43, 80)	<0.001	**<0.001**
Gender			0.6	0.7			0.3	0.4
F	19 (37%)	1,036 (41%)			6 (27%)	1,049 (41%)		
M	33 (63%)	1,512 (59%)			16 (73%)	1,529 (59%)		
Ethnicity			0.7	0.7			0.4	0.4
ASIAN	2 (4.7%)	47 (2.1%)			0 (0%)	49 (2.2%)		
BLACK	1 (2.3%)	124 (5.6%)			0 (0%)	125 (5.6%)		
HISPANIC	2 (4.7%)	96 (4.3%)			1 (6.7%)	97 (4.3%)		
MULTI RACE	0 (0%)	6 (0.3%)			0 (0%)	6 (0.3%)		
OTHER	1 (2.3%)	103 (4.7%)			2 (13%)	102 (4.6%)		
WHITE	37 (86%)	1,833 (83%)			12 (80%)	1,858 (83%)		
Unknown	9	339			7	341		
Cohort			0.051	0.090			0.7	0.7
SCI Fracture	2 (3.8%)	380 (15%)			4 (18%)	378 (15%)		
SCI noFracture	3 (5.8%)	121 (4.7%)			1 (4.5%)	123 (4.8%)		
Spine Trauma	47 (90%)	2,047 (80%)			17 (77%)	2,077 (81%)		
Length of stay (days)	7 (5, 11)	7 (4, 13)	0.6	0.7	9 (4, 14)	7 (5, 12)	0.2	0.4
Unknown	0	1			0	1		
Died in hospital	20 (38%)	169 (6.6%)	<0.001	**<0.001**	14 (64%)	175 (6.8%)	<0.001	**<0.001**
Number of diagnostics	26 (17, 30)	13 (9, 19)	<0.001	**<0.001**	14 (9, 26)	13 (9, 20)	0.13	0.3

Annexed Table 25. Univatiate comparisons between clusters of Glucose and Magnesium

Characteristic	Glucose				Magnesium			
	1 N = 272	2 N = 2328	p-value	q-value	1 N = 50	2 N = 2513	p-value	q-value
Age	69 (56, 80)	62 (41, 80)	<0.001	**<0.001**	62 (49, 78)	62 (43, 79)	0.7	>0.9
Gender			0.013	**0.015**			>0.9	>0.9
F	91 (33%)	964 (41%)			20 (40%)	1,017 (40%)		
M	181 (67%)	1,364 (59%)			30 (60%)	1,496 (60%)		
Ethnicity			0.3	0.3			0.5	0.9
ASIAN	8 (3.5%)	41 (2.0%)			1 (2.7%)	48 (2.2%)		
BLACK	12 (5.2%)	113 (5.6%)			3 (8.1%)	121 (5.6%)		
HISPANIC	14 (6.1%)	84 (4.2%)			3 (8.1%)	94 (4.3%)		
MULTI RACE	0 (0%)	6 (0.3%)			0 (0%)	6 (0.3%)		
OTHER	14 (6.1%)	90 (4.4%)			2 (5.4%)	99 (4.5%)		
WHITE	181 (79%)	1,689 (83%)			28 (76%)	1,811 (83%)		
Unknown	43	305			13	334		
Cohort			0.002	**0.003**			0.2	0.6
SCI Fracture	50 (18%)	332 (14%)			10 (20%)	371 (15%)		
SCI noFracture	23 (8.5%)	101 (4.3%)			4 (8.0%)	119 (4.7%)		
Spine Trauma	199 (73%)	1,895 (81%)			36 (72%)	2,023 (81%)		
Length of stay (days)	9 (5, 18)	7 (4, 12)	<0.001	**<0.001**	8 (5, 14)	7 (5, 13)	>0.9	>0.9
Unknown	0	1			0	1		
Died in hospital	38 (14%)	151 (6.5%)	<0.001	**<0.001**	13 (26%)	176 (7.0%)	<0.001	**<0.001**
Number of diagnostics	17 (11, 26)	13 (9, 19)	<0.001	**<0.001**	20 (10, 27)	13 (9, 19)	<0.001	**<0.001**

Annexed Table 26. Univatiate comparisons between clusters of Phosphate and Potassium

Characteristic	Phosphate				Potassium			
	1 N = 72	2 N = 2481	p-value	q-value	1 N = 81	2 N = 2521	p-value	q-value
Age	70 (56, 81)	62 (43, 79)	0.006	**0.013**	73 (60, 85)	62 (43, 79)	<0.001	**<0.001**
Gender			>0.9	>0.9			0.7	0.8
F	30 (42%)	1,003 (40%)			31 (38%)	1,024 (41%)		
M	42 (58%)	1,478 (60%)			50 (62%)	1,497 (59%)		
Ethnicity			>0.9	>0.9			0.3	0.5
ASIAN	2 (3.5%)	47 (2.2%)			2 (2.8%)	47 (2.2%)		
BLACK	2 (3.5%)	122 (5.7%)			1 (1.4%)	124 (5.7%)		
HISPANIC	2 (3.5%)	94 (4.4%)			1 (1.4%)	97 (4.4%)		
MULTI RACE	0 (0%)	6 (0.3%)			0 (0%)	6 (0.3%)		
OTHER	2 (3.5%)	98 (4.6%)			5 (7.0%)	99 (4.5%)		
WHITE	49 (86%)	1,782 (83%)			62 (87%)	1,810 (83%)		
Unknown	15	332			10	338		
Cohort			0.5	0.7			0.06	0.11
SCI Fracture	7 (9.7%)	371 (15%)			5 (6.2%)	377 (15%)		
SCI noFracture	4 (5.6%)	118 (4.8%)			3 (3.7%)	122 (4.8%)		
Spine Trauma	61 (85%)	1,992 (80%)			73 (90%)	2,022 (80%)		
Length of stay (days)	8 (5, 15)	7 (5, 13)	0.5	0.7	8 (5, 13)	7 (4, 12)	0.8	0.8
Unknown	0	1			0	1		
Died in hospital	26 (36%)	161 (6.5%)	<0.001	**<0.001**	15 (19%)	174 (6.9%)	<0.001	**0.001**
Number of diagnostics	24 (17, 33)	13 (9, 19)	<0.001	**<0.001**	21 (14, 30)	13 (9, 19)	<0.001	**<0.001**

Annexed Table 27. Univatiate comparisons between clusters of Sodium and Urea Nitrogen

Characteristic	Sodium				Urea Nitrogen			
	1 N = 2531	2 N = 69	p-value	q-value	1 N = 199	2 N = 2401	p-value	q-value
Age	62 (43, 79)	78 (64, 87)	<0.001	**<0.001**	78 (69, 87)	61 (41, 78)	<0.001	**<0.001**
Gender			<0.001	**<0.001**			>0.9	>0.9
F	1,008 (40%)	47 (68%)			81 (41%)	974 (41%)		
M	1,523 (60%)	22 (32%)			118 (59%)	1,427 (59%)		
Ethnicity			0.3	0.4			0.058	0.1
ASIAN	46 (2.1%)	3 (4.5%)			8 (4.4%)	41 (2.0%)		
BLACK	122 (5.6%)	3 (4.5%)			7 (3.9%)	118 (5.7%)		
HISPANIC	98 (4.5%)	0 (0%)			5 (2.8%)	93 (4.5%)		
MULTI RACE	6 (0.3%)	0 (0%)			0 (0%)	6 (0.3%)		
OTHER	102 (4.7%)	2 (3.0%)			3 (1.7%)	101 (4.9%)		
WHITE	1,812 (83%)	58 (88%)			157 (87%)	1,713 (83%)		
Unknown	345	3			19	329		
Cohort			0.076	0.11			0.08	0.1
SCI Fracture	378 (15%)	4 (5.8%)			20 (10%)	362 (15%)		
SCI noFracture	121 (4.8%)	3 (4.3%)			13 (6.5%)	111 (4.6%)		
Spine Trauma	2,032 (80%)	62 (90%)			166 (83%)	1,928 (80%)		
Length of stay (days)	7 (5, 13)	6 (4, 11)	0.6	0.6	7 (4, 11)	7 (5, 13)	0.082	0.1
Unknown	1	0			0	1		
Died in hospital	179 (7.1%)	10 (14%)	0.03	0.053	36 (18%)	153 (6.4%)	<0.001	**<0.001**
Number of diagnostics	13 (9, 19)	17 (13, 27)	<0.001	**<0.001**	20 (14, 27)	13 (9, 19)	<0.001	**<0.001**

Annexed Table 28. Univatiate comparisons between clusters of Hematocrit and Hemoglobin

Characteristic	Hematocrit					Hemoglobin				
	1 N = 834	2 N = 353	3 N = 1425	p value	q value	1 N = 696	2 N = 296	3 N = 1,611	p value	q value
Age	64 (45, 81)	47 (29, 63)	66 (47, 81)	<0.001	**<0.001**	61 (42, 80)	44 (29, 60)	66 (47, 81)	<0.001	**<0.001**
Gender				<0.001	**<0.001**				<0.001	**<0.001**
F	295 (35%)	79 (22%)	686 (48%)			222 (32%)	59 (20%)	774 (48%)		
M	539 (65%)	274 (78%)	739 (52%)			474 (68%)	237 (80%)	837 (52%)		
Ethnicity				0.5	0.5				0.2	0.2
ASIAN	21 (2.8%)	5 (1.8%)	23 (1.9%)			17 (2.7%)	4 (1.7%)	28 (2.0%)		
BLACK	42 (5.5%)	12 (4.3%)	71 (5.8%)			32 (5.1%)	8 (3.4%)	85 (6.1%)		
HISPANIC	37 (4.9%)	17 (6.1%)	46 (3.7%)			31 (4.9%)	14 (6.0%)	53 (3.8%)		
MULTI RACE	1 (0.1%)	1 (0.4%)	4 (0.3%)			1 (0.2%)	1 (0.4%)	4 (0.3%)		
OTHER	31 (4.1%)	17 (6.1%)	58 (4.7%)			29 (4.6%)	18 (7.7%)	59 (4.2%)		
WHITE	625 (83%)	227 (81%)	1,026 (84%)			517 (82%)	189 (81%)	1,166 (84%)		
Unknown	77	74	197			69	62	216		
Cohort				<0.001	**<0.001**				<0.001	**<0.001**
SCI Fracture	93 (11%)	94 (27%)	195 (14%)			81 (12%)	70 (24%)	231 (14%)		
SCI noFracture	57 (6.8%)	5 (1.4%)	63 (4.4%)			50 (7.2%)	7 (2.4%)	68 (4.2%)		
Spine Trauma	684 (82%)	254 (72%)	1,167 (82%)			565 (81%)	219 (74%)	1,312 (81%)		
Length of stay (days)	5 (4, 9)	9 (6, 16)	8 (5, 14)	<0.001	**<0.001**	6 (4, 9)	8 (5, 16)	8 (5, 14)	<0.001	**<0.001**
Unknown	0	1	0			0	0	1		
Died in hospital	39 (4.7%)	29 (8.2%)	121 (8.5%)	0.004	**0.005**	24 (3.4%)	29 (9.8%)	136 (8.4%)	<0.001	**<0.001**
Number of diagnostics	12 (8, 18)	13 (9, 20)	14 (9, 20)	<0.001	**<0.001**	11 (8, 16)	13 (9, 20)	15 (9, 20)	<0.001	**<0.001**

Annexed Table 29. Univatiate comparisons between clusters of MCH and MCHC

Characteristic	MCH					MCHC			
	1 N = 2472	2 N = 48	3 N = 83	p value	q value	1 N = 2513	2 N = 90	p value	q value
Age	62 (43, 79)	74 (52, 83)	64 (46, 80)	0.01	**0.023**	62 (43, 79)	74 (60, 85)	<0.001	**<0.001**
Gender				0.3	0.3			0.004	**0.01**
F	1,002 (41%)	15 (31%)	38 (46%)			1,005 (40%)	50 (56%)		
M	1,470 (59%)	33 (69%)	45 (54%)			1,508 (60%)	40 (44%)		
Ethnicity				<0.001	**0.003**			0.3	0.3
ASIAN	44 (2.1%)	0 (0%)	5 (6.8%)			46 (2.1%)	3 (3.6%)		
BLACK	112 (5.2%)	1 (2.4%)	12 (16%)			116 (5.3%)	9 (11%)		
HISPANIC	89 (4.2%)	1 (2.4%)	8 (11%)			95 (4.4%)	3 (3.6%)		
MULTI RACE	6 (0.3%)	0 (0%)	0 (0%)			6 (0.3%)	0 (0%)		
OTHER	101 (4.7%)	1 (2.4%)	4 (5.4%)			102 (4.7%)	4 (4.8%)		
WHITE	1,789 (84%)	38 (93%)	45 (61%)			1,807 (83%)	65 (77%)		
Unknown	331	7	9			341	6		
Cohort				0.2	0.3			0.3	0.3
SCI Fracture	364 (15%)	9 (19%)	9 (11%)			373 (15%)	9 (10%)		
SCI noFracture	116 (4.7%)	1 (2.1%)	8 (9.6%)			119 (4.7%)	6 (6.7%)		
Spine Trauma	1,992 (81%)	38 (79%)	66 (80%)			2,021 (80%)	75 (83%)		
Length of stay (days)	7 (4, 13)	7 (5, 11)	7 (4, 12)	0.8	0.8	7 (4, 13)	7 (5, 11)	0.2	0.3
Unknown	1	0	0			1	0		
Died in hospital	180 (7.3%)	7 (15%)	2 (2.4%)	0.035	0.061	180 (7.2%)	9 (10%)	0.3	0.3
Number of diagnostics	13 (9, 19)	16 (11, 22)	17 (12, 21)	<0.001	**0.002**	13 (9, 19)	21 (17, 28)	<0.001	**<0.001**

Annexed Table 30. Univatiate comparisons between clusters of MCV and Platelet Count

Characteristic	MCV				Platelet Count				
	1 N = 2575	2 N = 28	p value	q value	1 N = 51	2 N = 65	3 N = 2487	p value	q value
Age	62 (43, 80)	73 (63, 83)	0.012	**0.028**	77 (60, 86)	50 (36, 67)	62 (43, 80)	<0.001	**<0.001**
Gender			0.8	0.8				<0.001	**<0.001**
F	1,043 (41%)	12 (43%)			36 (71%)	24 (37%)	995 (40%)		
M	1,532 (59%)	16 (57%)			15 (29%)	41 (63%)	1,492 (60%)		
Ethnicity			0.3	0.5				0.6	0.6
ASIAN	48 (2.1%)	1 (4.8%)			0 (0%)	1 (1.9%)	48 (2.2%)		
BLACK	125 (5.6%)	0 (0%)			3 (6.1%)	3 (5.8%)	119 (5.5%)		
HISPANIC	98 (4.4%)	0 (0%)			2 (4.1%)	3 (5.8%)	93 (4.3%)		
MULTI RACE	6 (0.3%)	0 (0%)			0 (0%)	0 (0%)	6 (0.3%)		
OTHER	104 (4.7%)	2 (9.5%)			0 (0%)	5 (9.6%)	101 (4.7%)		
WHITE	1,854 (83%)	18 (86%)			44 (90%)	40 (77%)	1,788 (83%)		
Unknown	340	7			2	13	332		
Cohort			0.6	0.7				0.006	**0.008**
SCI Fracture	377 (15%)	5 (18%)			7 (14%)	16 (25%)	359 (14%)		
SCI noFracture	125 (4.9%)	0 (0%)			7 (14%)	0 (0%)	118 (4.7%)		
Spine Trauma	2,073 (81%)	23 (82%)			37 (73%)	49 (75%)	2,010 (81%)		
Length of stay (days)	7 (4, 13)	7 (5, 9)	0.082	0.14	9 (4, 12)	20 (13, 27)	7 (4, 12)	<0.001	**<0.001**
Unknown	1	0			0	0	1		
Died in hospital	181 (7.0%)	8 (29%)	<0.001	**0.004**	4 (7.8%)	8 (12%)	177 (7.1%)	0.2	0.3
Number of diagnostics	13 (9, 19)	17 (10, 27)	0.009	**0.028**	16 (13, 23)	17 (9, 24)	13 (9, 19)	<0.001	**<0.001**

Annexed Table 31. Univatiate comparisons between clusters of RDW and Red Blood Cells

Characteristic	RDW					Red Blood Cells				
	1 N = 234	2 N = 2267	3 N = 102	p value	q value	1 N = 493	2 N = 285	3 N = 1825	p value	q value
Age	71 (56, 83)	62 (42, 79)	56 (37, 79)	<0.001	**<0.001**	62 (40, 80)	43 (29, 59)	65 (47, 81)	<0.001	**<0.001**
Gender				0.015	**0.026**				<0.001	**<0.001**
F	115 (49%)	897 (40%)	43 (42%)			165 (33%)	61 (21%)	829 (45%)		
M	119 (51%)	1,370 (60%)	59 (58%)			328 (67%)	224 (79%)	996 (55%)		
Ethnicity				0.12	0.14				0.042	**0.042**
ASIAN	8 (3.7%)	39 (2.0%)	2 (2.7%)			10 (2.2%)	6 (2.7%)	33 (2.1%)		
BLACK	20 (9.3%)	103 (5.2%)	2 (2.7%)			31 (6.9%)	6 (2.7%)	88 (5.5%)		
HISPANIC	8 (3.7%)	87 (4.4%)	3 (4.0%)			26 (5.8%)	15 (6.7%)	57 (3.6%)		
MULTI RACE	0 (0%)	5 (0.3%)	1 (1.3%)			1 (0.2%)	1 (0.4%)	4 (0.3%)		
OTHER	6 (2.8%)	96 (4.9%)	4 (5.3%)			22 (4.9%)	16 (7.2%)	68 (4.3%)		
WHITE	172 (80%)	1,637 (83%)	63 (84%)			357 (80%)	179 (80%)	1,336 (84%)		
Unknown	20	300	27			46	62	239		
Cohort				0.089	0.13				<0.001	**<0.001**
SCI Fracture	25 (11%)	334 (15%)	23 (23%)			45 (9.1%)	66 (23%)	271 (15%)		
SCI noFracture	10 (4.3%)	112 (4.9%)	3 (2.9%)			40 (8.1%)	9 (3.2%)	76 (4.2%)		
Spine Trauma	199 (85%)	1,821 (80%)	76 (75%)			408 (83%)	210 (74%)	1,478 (81%)		
Length of stay (days)	8 (5, 13)	7 (4, 12)	9 (6, 11)	0.2	0.2	5 (4, 9)	8 (5, 16)	8 (5, 13)	<0.001	**<0.001**
Unknown	0	1	0			0	0	1		
Died in hospital	31 (13%)	140 (6.2%)	18 (18%)	<0.001	**0.001**	20 (4.1%)	25 (8.8%)	144 (7.9%)	0.005	**0.006**
Number of diagnostics	20 (14, 25)	13 (9, 19)	15 (10, 23)	<0.001	**<0.001**	12 (8, 17)	13 (9, 20)	14 (9, 20)	<0.001	**<0.001**

Annexed Table 32. Univatiate comparisons between clusters of White blood cells

Characteristic		White Blood Cells				
		1 N = 22	2 N = 2553	3 N = 28	p value	q value
Age		82 (65, 88)	62 (43, 79)	69 (58, 82)	0.007	**0.015**
Gender					0.7	0.9
	F	7 (32%)	1,037 (41%)	11 (39%)		
	M	15 (68%)	1,516 (59%)	17 (61%)		
Ethnicity					>0.9	>0.9
	ASIAN	0 (0%)	49 (2.2%)	0 (0%)		
	BLACK	1 (5.9%)	123 (5.6%)	1 (4.0%)		
	HISPANIC	0 (0%)	98 (4.4%)	0 (0%)		
	MULTI RACE	0 (0%)	6 (0.3%)	0 (0%)		
	OTHER	1 (5.9%)	104 (4.7%)	1 (4.0%)		
	WHITE	15 (88%)	1,834 (83%)	23 (92%)		
	Unknown	5	339	3		
Cohort					0.015	**0.026**
	SCI Fracture	5 (23%)	368 (14%)	9 (32%)		
	SCI noFracture	0 (0%)	122 (4.8%)	3 (11%)		
	Spine Trauma	17 (77%)	2,063 (81%)	16 (57%)		
Length of stay (days)		7 (3, 13)	7 (5, 13)	7 (4, 9)	0.3	0.4
	Unknown	0	1	0		
Died in hospital		8 (36%)	172 (6.7%)	9 (32%)	<0.001	**0.002**
Number of diagnostics		26 (17, 33)	13 (9, 19)	15 (11, 21)	<0.001	**<0.001**

Annexed Table 33. GBMT model selection results

k	poly	npar	BIC	APPA	%class1	%class2	%class3	%class4	%class5	%class6	%class7	%class8
1	p1	24	246211.5	1.00	100.00	NA	NA	NA	NA	NA	NA	NA
1	p2	29	245562.0	1.00	100.00	NA	NA	NA	NA	NA	NA	NA
1	p3	33	243556.1	1.00	100.00	NA	NA	NA	NA	NA	NA	NA
1	p4	40	242531.2	1.00	100.00	NA	NA	NA	NA	NA	NA	NA
2	p1	51	243007.2	0.96	93.31	6.69	NA	NA	NA	NA	NA	NA
2	p2	59	241137.3	0.98	91.43	8.57	NA	NA	NA	NA	NA	NA
2	p3	67	237817.9	0.98	92.12	7.88	NA	NA	NA	NA	NA	NA
2	p4	81	240144.7	0.87	76.33	23.67	NA	NA	NA	NA	NA	NA
3	p1	77	241091.0	0.97	90.75	5.77	3.48	NA	NA	NA	NA	NA
3	p2	92	239446.4	0.96	89.79	6.77	3.44	NA	NA	NA	NA	NA
3	p3	103	236305.4	0.98	90.63	6.65	2.72	NA	NA	NA	NA	NA
3	**p4**	**119**	**235009.0**	**0.98**	**90.55**	**6.62**	**2.83**	**NA**	**NA**	**NA**	**NA**	**NA**
4	p1	101	240594.0	0.91	68.57	23.02	5.12	3.29	NA	NA	NA	NA
4	p2	122	238982.6	0.96	89.22	1.64	6.08	3.06	NA	NA	NA	NA
4	p3	139	235961.1	0.97	90.25	0.46	6.62	2.68	NA	NA	NA	NA
4	p4	119	235009.9	0.98	90.44	6.65	2.91	NA	NA	NA	NA	NA
5	p1	128	240052.6	0.91	71.43	3.48	18.47	4.97	1.64	NA	NA	NA
5	p2	152	238271.4	0.97	88.07	1.53	1.19	6.42	2.79	NA	NA	NA

Annexed Table 33. GBMT model selection results

k	poly	npar	BIC	APPA	%class1	%class2	%class3	%class4	%class5	%class6	%class7	%class8
5	p3	174	235656.6	0.96	86.35	7.23	0.31	3.82	2.29	NA	NA	NA
5	p4	162	234486.4	0.93	85.70	4.78	4.51	5.01	NA	NA	NA	NA
6	p1	154	239443.8	0.86	48.22	36.02	2.79	6.92	4.70	1.34	NA	NA
6	p2	184	237811.8	0.95	85.20	3.63	1.34	0.69	6.62	2.52	NA	NA
6	p3	210	235009.4	0.95	86.85	2.41	4.86	0.46	3.82	1.61	NA	NA
6	p4	203	234244.5	0.93	85.12	1.64	4.74	3.21	5.28	NA	NA	NA
7	p1	180	238831.9	0.89	31.09	6.62	16.98	1.84	40.54	2.07	0.88	NA
7	p2	215	237363.0	0.95	8.18	82.22	2.14	1.45	0.80	4.32	0.88	NA
7	p3	213	235041.9	0.93	4.28	87.88	1.19	1.03	5.12	0.50	NA	NA
7	p4	243	234138.6	0.93	2.03	84.67	0.69	4.05	3.52	5.05	NA	NA
8	p1	206	238459.1	0.86	26.08	39.20	2.37	2.72	6.04	17.86	4.09	1.64
8	p2	245	237425.1	0.93	7.80	82.45	0.92	1.30	0.61	1.41	4.02	1.49
8	p3	249	234932.0	0.95	4.21	87.76	1.22	0.84	0.73	4.67	0.57	NA
8	p4	242	233832.0	0.93	2.18	84.17	0.99	5.39	2.56	4.70	NA	NA