technische universität
dortmund

Leveraging real-world biomarker data: Statistical methods for investigating missingness and longitudinal information for patient risk assessment

Berit Hunsdieck

A dissertation submitted towards the degree
Doctor of Natural Sciences (Dr. rer. nat.)
of the Faculty of Statistics
of Technische Universität Dortmund

Advisor and referee: Prof Dr. Katja Ickstadt
Additional external advisors: Dr. Johanna Mielke, Dr. Christian Bender
Referee: Prof. Dr. Jörg Rahnenführer
Date of Defense: 07.07.2025

Abstract

The increasing reliance on big data within the pharmaceutical industry underscores significant challenges related to noise and missingness, which can adversely impact data quality and the interpretation of patient outcomes. Noise arises from various sources across different data types, with genomic and transcriptomic data influenced by genetic and environmental factors, while proteomic and metabolomic data are affected by a wider array of variables, complicating the extraction of meaningful insights. Additionally, missing data—whether classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR)—further complicates analyses, particularly in longitudinal studies. To address these challenges, this thesis emphasises the importance of replicating real-world conditions, enabling researchers to better understand data behaviour and develop robust statistical methodologies.

The thesis comprises three papers that tackle these challenges from multiple perspectives. The first paper focusses on missingness in metabolomics data, proposing a novel clustering method that integrates missingness information, rather than relying solely on imputed values, to enhance clustering accuracy. This two-step clustering procedure aims to improve patient clustering outcomes, particularly when data are MNAR, and demonstrates superior performance compared to standard methods as confirmed by external validation measures.

The second paper addresses the complexities of utilising longitudinal Electronic Health Record (EHR) data for health risk assessment by employing joint models that integrate longitudinal and survival data. This study simulates realistic longitudinal EHR data, incorporating noise, sample size, and cohort homogeneity, to analyse how various data quality characteristics impact model performance. The findings reveal conditions under which joint models outperform traditional Cox models in risk prediction.

The third paper explores the modelling and prediction of blood pressure trajectories using data from wearable devices in hypertensive patients. For that, a framework is developed for simulating realistic blood pressure trajectories. This framework is used for evaluating the performance of novel statistical approaches for the prediction of treatment effects of antihypertensive drugs.

Together, these studies contribute to a deeper understanding of the challenges posed by noise

and missingness in pharmaceutical big data, while offering innovative methodologies to enhance data analysis and patient outcomes.

Acknowledgements

First of all, I would like to thank my supervisor, Johanna, for her support and optimism throughout my time at Bayer. Thank you for always supporting and guiding me when I needed it, while also allowing me a lot of freedom to explore and find my way. I would like to thank my second advisor, Christian, for acting as a supervisor when needed and for consistently finding time for me. Without your support I would not have been able to finish my dissertation in time! I am grateful for the opportunity to work on my Ph.D. within the computational biology team. It was a great experience working with such a diverse international team in which I felt welcomed right from the beginning and for making my time as a PhD student so fun.

I am equally thankful to Katja Ickstadt for acting as my internal advisor at TU Dortmund, who always freed up time for me while providing guidance and support throughout my academic path early on. Furthermore, thanks to Joerg Rahnenfuehrer for refereeing the thesis and Uwe Ligges and Guido Knapp for being part of my thesis committee.

I would like to thank all my PhD colleagues, particularly Jan, Patryk, and Afrah. Thank you for sharing not only the challenges that come with pursuing a Ph.D. but also for helping me navigate through them. Your support has been a source of relief, allowing me to clear my mind and focus on what truly matters.

Furthermore, I express my gratitude to Eren Elci for assisting me in my decision to pursue a Ph.D. and for offering valuable guidance throughout that journey.

Finally, I would like to thank my family, especially my sister Randi, for always being available for coffee breaks during my home office sessions, and friends who have greatly supported me throughout all the ups and downs of this academic journey. Without your support, I would not have made it through the difficult times of the last few years. Thank you for believing in me and for being my greatest source of strength.

List of contributed papers

This cumulative thesis is based on the following three manuscripts:

Article 1: Hunsdieck, B., Bender, C., Ickstadt, K., GCKD Investigators, Mielke, J. (2025). *Leveraging missing information in clustering: A novel method integrating continuous values and informative missingness.* Submitted to Biometrical Journal.

Article 2: Hunsdieck, B., Bender, C., Ickstadt, K., Mielke, J. (2025). *Joint models in big data: Simulation-based guidelines for required data quality in longitudinal electronic health records.* In review for BMC BioData Mining. Preprint available at `https://doi.org/10.21203/rs.3.rs-6031358/v1`

Article 3: Hunsdieck, B., Mielke, J., Ickstadt, K., Elçi, E. (2025). *A simulation-based framework for modeling and prediction of personalized blood pressure trajectories in hypertensive patients after antihypertensive treatment.* PLOS ONE 20(4): e0318549. `https://doi.org/10.1371/journal.pone.0318549`

Further work:

Patent 1: J. I. Mielke, J., Freitag, D., Hunsdieck, B., De Zoete, J.: MONITORING INTAKE OF ANTIHYPERTENSIVE DRUGS, WO2024146753A1

Patent 2: Hunsdieck, B., Elci, E., Mielke, J.: MONITORING THE TREATMENT OF HYPERTENSION, WO2024126178A1

Contents

Chapter 1

Introduction

The phrase "precision medicine" has gained significant popularity in recent years, driven by scientific findings (König et al., 2017). The idea of precision medicine is to identify relevant disease subtypes, as well as patient subgroups, while taking into account genetic, environmental, and lifestyle factors to provide more effective and personalised care, improve results, and minimise adverse effects. This concept is not new; Hippocrates (460-370 BCE) once said "It is far more important to know what person the disease has than what disease the person has". In the past, the problem was that this concept of precision medicine could not be realised due to limited data sources, restricted to clinical trials. Starting with the completion of the human genome project in 2011, the implementation of genomic-based precision medicine has evolved (Gibbs, 2020). As more and more sources of big data related to health became available, more and more data types became available and promoted the implementation of precision medicine. In recent decades, data availability has increased rapidly, driven by public and private data generation consortia, particularly real-world big data. These data can offer previously unknown information on patient characteristics, treatment outcomes, and disease progression in various populations and subgroups. In 2024, the number of patent applications related to big data in the pharmaceutical industry increased by 33% compared to Q2 2023 (Pharmaceutical Technology, 2024). Real-World big data can be sourced from different sources like large cohort studies or so-called biobanks. A biobank is defined as a structured collection of biological samples and associated data, stored for the purposes of current and future research (Parodi, 2015). The potential of biobanks was recognised by the Time magazine already in 2009 as one of the top ten ideas that transform the world (Park, 2009). Biobanks are now essential for pharmaceutical research, helping in the discovery of human target-disease links and gene-environment interactions. They provide information on disease pathogenesis, support risk assessment, diagnostics, pharmacogenomics, and drug development, all crucial for advancing personalised medicine (Olson et al., 2014).

In the realm of pharmaceutical development, the initial emphasis of big data was on genomic data, which played a crucial role in revealing disease mechanisms and informing drug development strategies, demonstrated by resources such as genome databases, for example, the human genome project (Gibbs, 2020). However, recent years have witnessed a significant shift towards incorporating a broader range of data sources, including multiple omics layers and real-world data, particularly from electronic health record (EHR) data, a digital version of a patient's paper chart (Yang et al., 2020). Omics data, including genomics, transcriptomics, proteomics, and metabolomics, have emerged as a key component in the pursuit of personalised medicine, providing personalised insights of individual patient profiles. Genomics involves examining the complete set of genes -the genome- within an organism, focussing on their structure, function, evolution, and mapping, playing a crucial role in enhancing our understanding of genetic disorders, developmental processes, and evolutionary biology (Vailati-Riboni et al., 2017). Transcriptomics focusses on the complete set of RNA transcripts produced by the genome -the transcriptome- analysing gene expression patterns to understand gene regulation (Vailati-Riboni et al., 2017). Proteomics is the study of proteins, particularly their functions and structures, with the aim of understanding the protein composition of cells and tissues, their interactions, and their roles in biological processes (Vailati-Riboni et al., 2017). Lastly, metabolomics studies the complete set of metabolites (small molecules) within a biological sample, providing insights into metabolic processes, identifying biomarkers for diseases, and revealing the effects of drugs or environmental changes on metabolism (Vailati-Riboni et al., 2017). Metabolomics technology is rapidly advancing, enabling researchers to investigate the complexities of the human metabolome, which is estimated to include more than 19.000 small molecules derived from various biological sources (Wishart et al., 2012). These omics datatypes are interconnected and collectively contribute to a comprehensive understanding of biological systems. Additionally, there is primary care data. This data provides information collected during routine healthcare visits. It can include, for example, general practices, medical history, diagnostic tests, clinical findings, and patient outcomes, but also patient demographics. Primary care data can be accessed through EHR data. It can be used to improve quality, plan health services, and perform epidemiological studies and research (de Lusignan and van Weel, 2006).

However, while offering detailed insights into biological processes, omics data is often not readily available on a large scale in biobanks. The application of omics measurements as a screening tool in clinical data is challenging due to the lack of availability of established assays. Taking these two aspects together, both EHR and omics data are considered valuable tools for precision medicine approaches. Although the focus of this thesis is on EHR and, as a representative of the class of omics measurements, on metabolomics data, it should be noted that all proposed methodologies can in principle also be applied to other types of data. In the following, we will use the term real-world

evidence (RWE) as an umbrella term covering both EHR and omics data.

Using available RWE data, researchers and clinicians can improve their understanding of disease processes and improve the precision of patient management strategies, ultimately leading to more effective and tailored therapeutic interventions. This approach not only streamlines the research process but also supports the integration of omics data into routine clinical practice, paving the way for advancements in personalised medicine (Wu et al., 2016).

The growing wealth of real-world big data presents both opportunities and challenges for the pharmaceutical industry. Historically, the most important source of human data in the pharmaceutical industry was given by randomised controlled trials that have generated high-quality data due to strict study protocols (Kendall, 2003). The underlying standards for data quality are very high, i.e. typically one would expect a standardised protocol with measurements taken at the same time for all patients. This reduces the noise of the measurements. In addition, a high number of missing observations is typically not observed since the study protocols ensure a data collection that is as complete as possible.

In RWE data, the situation is completely different: firstly, one is faced with a substantially increased heterogeneity of patients since those are not only carefully selected patients but cover a broad spectrum of individuals with varying diagnoses and characteristics. This patient diversity improves the richness of the data, but also introduces challenges, such as greater heterogeneity and mixed signals when it comes to information extraction (Burden, 2019).

Furthermore, it is important to recognise that all biological systems inherently exhibit variability, often referred to as noise (Ning and Lo, 2010). This noise becomes increasingly significant when working with larger and more inclusive databases. In the context of RWE, noise can be caused by various factors, including measurement variability, data entry inconsistencies, and overall data quality. For example, in omics data, sources of noise may include individual genetic predispositions, environmental influences, technical variability during sample collection and processing, sequencing errors, and inconsistencies in bioinformatics pipelines (Lay Jr et al., 2006). In processes that are less directly related to biological mechanisms, such as proteomics, metabolomics, and phenotyping, additional variables contribute to noise. These include regulatory processes (e.g. protein activation and inactivation), metabolic states, cell cycle phases, and interactions with external factors such as diet, lifestyle, stress, pharmaceuticals, and radiation (Moseley, 2013).When it comes to phenotyping, measurements are particularly susceptible to a broad range of error sources beyond those previously mentioned. These may include measurement errors, environmental factors, and other influences that can further complicate data interpretation.

This variability, which arises from differences between individuals and experimental conditions, can complicate the interpretation of the results. Understanding and mitigating this noise is crucial

for improving the reliability and validity of findings derived from real-world evidence.

But not only unknown influences causing noise can distort disease pathways and signals when it comes to data analysis. Another key feature we want to focus on is the case of missing data. The correct handling and interpretation of missing information is crucial while working with RWE. A distinction is made between three types of missingness (Rubin, 1976). Missingness is called missing completely at random (MCAR) when it is independent of the data. When missingness relies on the observed data—and given these data, it is independent of what has not been observed—the mechanism is termed missing at random (MAR). If missingness is influenced by the unobserved data, possibly along with the observed data, it is known as not missing at random (NMAR).

In general, missing data can be bypassed by using imputation techniques such as multiple imputation, Expectation-Maximisation (EM) imputation, or regression imputation. These methods are valid if the underlying missingness mechanism is not NMAR and the percentage of missing data is not too great (Scheffer, 2002). If the underlying missingness mechanism is NMAR, common imputation methods will cause biases and are therefore not recommended. To incorporate the underlying mechanism, weighted adjustments can be used, which require knowledge about missing data mechanisms (Little and Rubin, 2019).

Another significant issue, especially in primary care data, is the varying frequency and unequal spacing of the measurements. In contrast to clinical trial data, where standardised procedures are applied uniformly across all patients, in primary care data the measurements of specific phenotypes occur only when the clinician thinks that it is necessary. This practice not only introduces selection bias regarding biomarkers, but may also lead to informative missingness when analysing the frequency and temporal spacing of individual measurements. As a result, the absence of data points can reflect underlying patterns related to the clinical decision-making process, further complicating the interpretation of the data (Haneuse and Daniels, 2016). Ibrahim et al. performed a comprehensive review of missing data methods in longitudinal studies (Ibrahim and Molenberghs, 2009). They highlighted that each approach is based to some extent on unverifiable assumptions and therefore has its strengths and weaknesses, so that the method of choice should be guided by the nature of the missing data and the specific research context.

The presence of these unknown distorting factors underscores the need for stringent quality control and advanced analytical techniques to guarantee accurate data interpretation and reliable conclusions when it comes to pharmaceutical research. As already highlighted, RWE come with a lot of problems that have to be tackled with a suitable statistical methodology when it comes to data analysis.

In order to tackle these challenges and address key objectives in precision medicine, it is essential to combine the fields of biology, computer science, and statistics, while considering various data-

related issues. A suitable statistical method is crucial for performing robust analyses and taking advantage of large biodata to improve understanding of diseases and health outcomes, highlighting the need for innovative approaches in data analysis and interpretation (Goh and Wong, 2020).

In the realm of real-world data analysis, managing diverse types of data presents numerous challenges. To effectively assess the impact of different characteristics of the data set and compare the performance of statistical methodologies, creating a controlled environment is essential. This can be achieved through the use of simulation studies, which offer a structured approach to understanding and mitigating potential impacts. Simulation modelling techniques, increasingly used in healthcare applications with various software tools and data sources, have gained growing interest as indicated by the number of published reviews (Salleh et al., 2017). In healthcare, simulation is gaining importance as a methodology for enhancing system improvement initiatives. For any given project, the simulation methodology to be used is highly application-dependent. By simulating the dynamics of an actual health system or procedure over time, one can more effectively understand and evaluate the impact of specific data parameters without having additional effort and costs for data collection (Alvarado et al., 2016). For example, by introducing varying levels of noise or simulating missing data points, researchers can explore how these factors affect the estimation of trends, treatment effects, and patient outcomes. This approach allows for a better understanding of the robustness of the analytical methods used in real-world data scenarios, ultimately leading to improved strategies for managing and interpreting complex healthcare data.

Based on simulated data, algorithms can be developed and improved to address critical research questions in the pharmaceutical industry. Once these algorithms have been developed, it is crucial to assess the real-world performance of the algorithms. In this work, two main real-world data sources have been used for validation, namely the UK Biobank and the German Chronic Kidney Disease (GCKD) study.

First, we use the UK Biobank (Sudlow et al., 2015). It consists of patient data from approximately 500.000 individuals between 40 and 69 years, who were recruited throughout the UK between 2006 and 2010. All participants provided their informed written consent to participate in the study, which was approved by the National Research Ethics Service (11/NW/0382). In this study, a variety of data types were collected, including, for example, omics data, imaging, and EHR data. Alternative sources of more dense longitudinal health data are wearables, collecting, for example, pulse and blood pressure measurements. Although this type of data might be more dense, it is also more susceptible to sources of error caused by activity or device inaccuracies.

Second, we use the German Chronic Kidney Disease study (GCKD). The GCKD study is a prospective observational cohort study of 5217 patients between the ages of 18 and 74 years with chronic kidney disease treated by nephrologists. Patients were included if they had an eGFR between

30 and 60 $ml/min/1.73\ m^2$ or a urinary albumin-to-creatinine ratio (UACR) greater than $300mg/g$. Patients completed a medical assessment, including plasma, serum, and spot urine measurements, every two years (Titze et al., 2015).

This thesis focusses on metabolomics and longitudinal phenotype data, with the goal of extracting valuable information despite challenges in data availability and noise.

The evolving field of metabolomics presents new opportunities for discovery, as metabolites serve as critical indicators for diseases and can often be linked to underlying pathogenic mechanisms. For example, earlier studies have revealed early metabolic indicators for conditions such as pancreatic cancer and type 2 diabetes, often detectable years before clinical symptoms appear (Mayers et al., 2014; Rhee et al., 2011).

When dealing with metabolomics, addressing missing data, for example, caused by device detection limits, is crucial for accurate analysis and interpretation. Studies have applied various imputation methods, but often overlook the impact of missingness in clustering, highlighting a gap in methods that account for informative missingness in metabolomic insights.

Given longitudinal primary care data, for example, phenotype trajectories over several years of GP visits, these data can be leveraged to gain new insights into disease mechanisms and subtypes, for example, modelling the individual patient courses and drawing conclusions on the patients' disease risk by using various statistical approaches. Recent insights derived from longitudinal EHR data indicate that quantitative phenotypes and disease trajectories in women reveal new critical pathways linked to various adverse outcomes, illuminating potential targets for the early detection and prevention of these diseases (Yang et al., 2022).

The problem of unknown distorting factors and missingness can also be observed here. For primary care data, these data are regularly collected by general practitioners (GPs), so there is not only an extensive and diverse patient pool, but also a wealth of data of varying quality accumulated over the years. This supports the opportunity of longitudinal analyses over years and offering a valuable means to track disease risks within populations. However, it is crucial to recognise the biases that can arise from EHR datasets, as this understanding is essential to design epidemiological studies and accurately interpret their results. Implementing strategies to mitigate bias in the context of EHR data can improve the quality and applicability of these data sets (Bower et al., 2017).

In summary, the growing reliance on big data in the pharmaceutical industry highlights significant challenges posed by noise and missingness, which can adversely affect the quality of the data and interpretation of patient outcomes. Noise originates from various sources across data types, with genomic and transcriptomics data influenced by genetic and environmental factors, while proteomics and metabolomics data are affected by a broader range of variables, complicating the extraction of meaningful insights. Furthermore, missing data, whether classified as MCAR, MAR, or NMAR,

further complicate the analyses, particularly in longitudinal studies.

This thesis addresses the aforementioned challenges from multiple viewpoints. The first paper focusses on missingness in metabolomics data. The paper discusses a novel clustering method that addresses missing information in the data. The approach explicitly integrates missingness information, not just imputed values, to improve clustering accuracy. This is especially significant if the data is not missing at random. A two-step clustering procedure is proposed based both on the values and the missingness pattern, with the aim of improving patient clustering outcomes. The method's effectiveness is evaluated using simulated datasets and is shown to outperform standard clustering methods.

The second paper addresses the challenges of using longitudinal EHR data for health risk assessment by employing joint models that integrate longitudinal and survival data. The goal of the study is to develop simulation-based checklists for required data quality to understand when the more complex joint models are superior to a cox regression model, i.e. a well-established survival model that only uses cross-sectional data only. Thus, the study focusses on the simulation of realistic longitudinal EHR data to represent patient health records, incorporating noise, sample size, and cohort homogeneity. By including noise terms and measurement inaccuracies, the model performance dependent on the characteristics of the data is studied. In total, the study compares the performance of a joint model that combines longitudinal and survival information with Cox models in different data scenarios, identifying conditions for superior risk predictions using longitudinal data.

The third paper discusses modelling and predicting blood pressure trajectories using data from wearable devices in hypertensive patients. A simulation framework for blood pressure profiles is introduced through Pharmacokinetic-Pharmacodynamic modelling, which incorporates individual daily rhythms, patient characteristics, and medication effects. Based on this, we propose and evaluate two models for steady-state prediction under antihypertensive therapy, a Gaussian process and a nonlinear mixed-effect model.

The methods developed for the individual papers are introduced in the corresponding section. Each chapter ends with an outlook for possible further research in its topic. All full-length papers are attached thereafter.

Chapter 2

A novel method integrating continuous values and informative missingness in clustering

Contents

2.1 Contributed materials

Hunsdieck, B., Bender, C., Ickstadt, K., GCKD Investigators, Mielke, J. (2025). *Leveraging missing information in clustering: A novel method integrating continuous values and informative missingness.* Submitted to Biometrical Journal.

Authors contribution Berit Hunsdieck developed all the methods, designed and executed the simulation studies, interpreted the results, and wrote the manuscript. Johanna Mielke contributed the initial idea of the project. Johanna Mielke and Christian Bender contributed ideas for the design of the simulation study and evaluation, and corrected and approved the manuscript. Katja Ickstadt supervised the project, contributed to the design of the simulation study and to the interpretation of the results, and corrected and approved the manuscript.

2.2 Summary

One of the major challenges when dealing with RWE is the large amount of missing data. Missing values are typically imputed, i.e., replaced by plausible candidates. Although the imputation method may differ, ultimately, there is an information loss because the fact that a value was initially missing is not used. This may be negligible, especially when the missingness is either completely at random (MCAR) or at random (MAR). However, in cases where data are not missing at random (MNAR) but is missing for a reason, working only with imputed data can introduce a significant bias (Kang, 2013). An example of different types and causes of missingness can be found in dealing with metabolomics data. Although missing data may arise randomly, for example due to problems during sample preparation, it may also result from an inherent mechanism that provides information, such as when values fall below the limit of detection (LOD) (Do et al., 2018). In recent research, several approaches for missing value imputation have been studied in depth with application in metabolite data. A study evaluated 31 different methods for imputing missing values (Do et al., 2018), while another study focused on eight imputation techniques (Wei et al., 2018). However, it is important to note that these studies rarely consider the information of missingness itself, particularly in the context of cluster analysis. Traditional clustering methods, such as fuzzy k-means clustering (Čuperlović-Culf et al., 2009), self-organising maps (SOMs) (Beckonert et al., 2003), and genetic algorithms (Petricoin et al., 2002), typically rely solely on available values and do not integrate missing information at all.

We hypothesise that excluding missing information can reduce clustering effectiveness, insights and reliability. This paper proposes that taking into account missing data can improve patient clustering and help detect patient clusters at risk, highlighting the need for methods that address missing data, particularly those targeting informative missingness. The goal is to integrate the presence of missing data into the clustering process, thereby considering both the actual values and the absence of data during clustering, rather than solely focusing on the imputed values.

2.2.1 Overview of methods and simulation

Figure 2.1 shows an overview of the proposed procedure. It consists of two steps: first, an initial clustering is performed based on the imputed values with a subsequent merging of similar clusters based on the corresponding within- and between-sample distances; second, a subclustering is performed based on missing data information, again followed by a merging step.

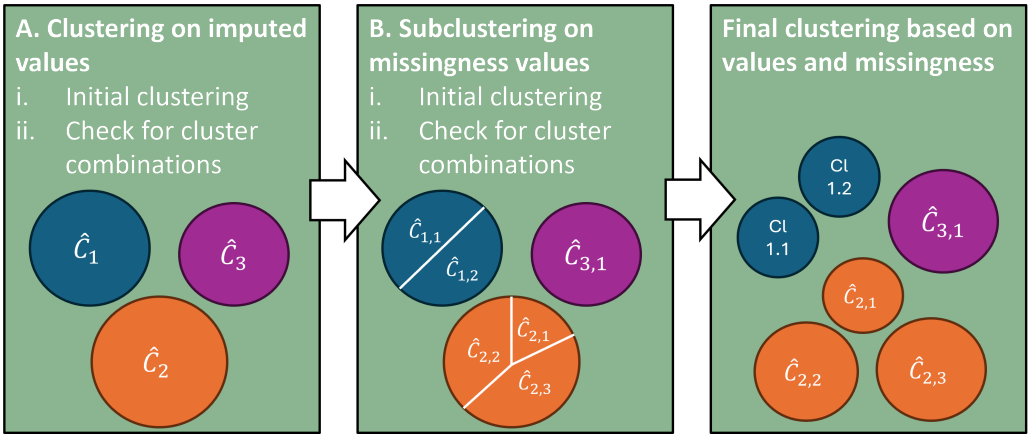

Figure 2.1: Algorithm Scheme: Given prerequisits, the new clustering scheme is devided into to steps, focussing on the value and the missingness information separately. Extracted from Hunsdieck et al. (2025b).

For introducing the single steps in more detail, we define

n : Number of Individuals

M : Number of variables/ markers, $M = m_1 + m_2$

m_1 : Number of cluster-specific markers

m_2 : Number of cluster-unspecific markers

$v_m^{(val)}$: Observed or imputed value of variable/marker m with $v_m^{(val)} \in \mathbb{R}^n,\ m \in \{1, \dots, M\}$

$v_m^{(mis)}$: Missingness information of variable/marker m with $v_m^{(mis)} \in \{0,1\}^n,\ m \in \{1, \dots, M\}$.

For our proposed approach, we first perform an *Initial clustering* on the observed or imputed values, then a *Cluster combination* step. Afterwards obtained clusters are subdivided by an *Initial clustering* based on the missingness information followed by a *Cluster combination* step. The two concepts are introduced below.

Initial clustering Initial clustering is done on measurements (Step A) or missingness (Step B), using k-means. In Step B, subclusters are derived from Step A's clusters. The number of initial clusters, k_0, can either be predetermined or calculated using Maechler's gap statistics method, which builds on Tibshirani et al.'s work (Tibshirani et al., 2001). Using k_0 clusters, the k-means algorithm is applied to each variable set (A or B) to assign each observation to a cluster $C_k^{(A/B)}$ for $k = 1, \dots, k_0$. In the context of the paper, k-means clustering is preferred over methods such as PAM and hierarchical clustering, as it was shown in the simulation study that k-means outperforms the other approaches for the scenarios considered.

Cluster combination Given the clusters $\hat{C}_1, \ldots, \hat{C}_{\hat{k}}$ resulting from the initial clustering, these clusters will be redefined based on their "uniqueness", taking into account the cluster compactness as well as the cluster distinction. This is achieved by incorporating the average distance of the clusters, defined by

$$d_{ij} = \frac{1}{n_1 \cdot n_2} \sum_{s=1}^{n_1} \sum_{s=1}^{n_2} d(x_s, y_t) \tag{2.1}$$

given two clusters $\hat{C}_i = \{x_1, \ldots, x_{n_1}\}$ and $\hat{C}_j = \{y_1, \ldots, y_{n_2}\}$ with

$$d(x_s, y_t) = \begin{cases} d(v_m^{(val)}(s), v_m^{(val)}(t)) & \text{, when looking at (imputed) values (A)} \\ d(v_m^{(mis)}(s), v_m^{(mis)}(t)) & \text{, when looking at missingness (B) .} \end{cases} \tag{2.2}$$

The corresponding individual distances can either be calculated using the Euclidean distance or other similarity measures, for example, Pearson's correlation coefficient.

The main concept involves utilising these distances in such a way that if the distance between clusters substantially exceeds the distance within the single clusters, the clusters will then be combined, so that in the end cohesive and well-defined clusters are remaining. To achieve this, we calculate the distance ratio between two clusters i and j. This ratio is estimated by dividing the minimum of the average within-cluster distances $min(d_{i,i}, d_{j,j})$ by the average distance between the clusters $d_{i,j}$, namely

$$r_{i,j} = \frac{min(d_{i,i}, d_{j,j})}{d_{i,j}}. \tag{2.3}$$

To derive a confidence interval

$$PI_{i,j} = [q_{0.05,\hat{R}_{i,j}}, q_{0.95,\hat{R}_{i,j}}] \tag{2.4}$$

for the empirical ratio, a resampling approach is used. The decision to merge is based on the prediction interval $PI_{i,j}$. The clusters i and j merge when $PI_{i,j}$ is included in the 'equivalence interval' $[1-\alpha, 1+\alpha]$. If not, they will remain distinct. If multiple merging steps are possible, the merging process will be performed iteratively, beginning with the combination of the two clusters having the maximum empirical ratio.

To objectively evaluate the performance of the algorithm, simulated data with known ground truth clusters are used. To assess whether the clustering method is performing well, several external validation measures, namely the adjusted rand index (ARI) (Hubert and Arabie, 1985), purity (Kim and Park, 2007), entropy (Kim and Park, 2007), and mutual information (Cover, 1999), are evaluated given the ground truth.

For the simulated data sets, a population with n individuals is assumed with M markers, including both value information and missingness indicators (1 for missing, 0 for present). Missing values

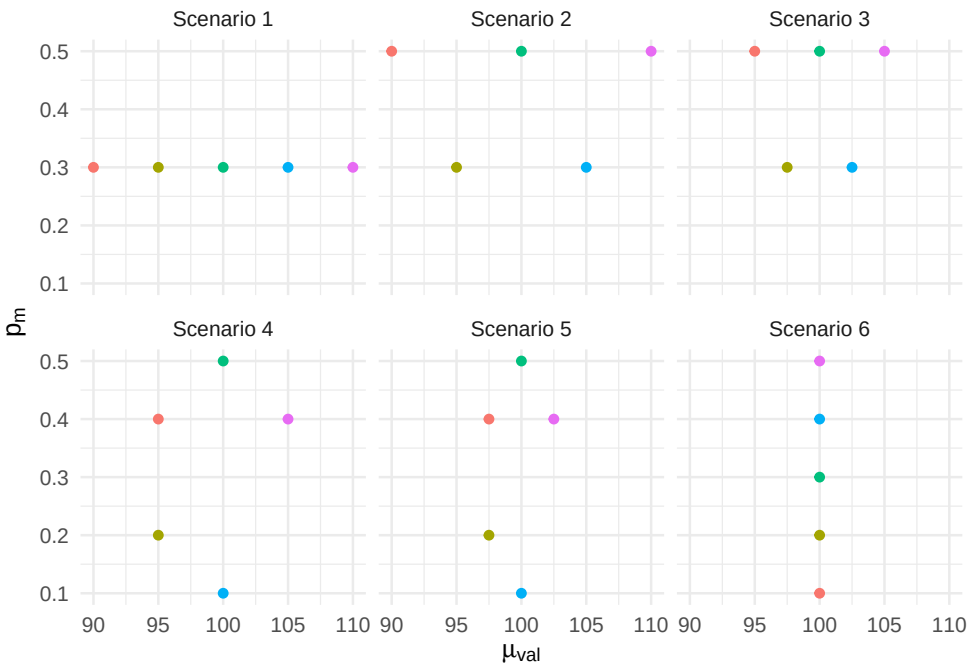

Figure 2.2: Visualisation of cluster means μ_{val} and p_m for seven different scenarios that are analysed further. Extracted from Hunsdieck et al. (2025b).

are imputed using predictive mean matching. Markers are divided into m_1 cluster-specific and m_2 cluster-unspecific types such that $M = m_1 + m_2$. To simulate a cluster of size n_A, we generate a mean value μ_{val} for the cluster-specific markers and a probability of missingness p_m. To be able to evaluate the performance of the clustering algorithm, different scenarios are considered for the choice of μ_{val} and p_m.

Individual values are assumed to be normally distributed around μ_{val} with a standard deviation of 5, and the missingness is modelled from a Bernoulli distribution with probability p_m.

The performance of the algorithm is analysed in six scenarios that vary in the configuration of the clusters, given by cluster-specific mean values μ_{val} and missingness probabilities p_m (see Figure 2.2). Scenario 1 uses value information alone for cluster formation, scenario 6 uses missingness information alone, and scenarios 2 to 5 use both value and missingness information with differing gradations of the value and missingness influence.

2.2.2 Key findings and their broader research implications

Among the established methods (k-means, hierarchical clustering, PAM), k-means performs comparably or better than the others based on various validation measures (ARI, purity, entropy, mutual information). To evaluate the performance of the novel approach, the influence of the three hyper-

parameters α, k_0 and the distance measure is analysed by evaluating the final number of derived clusters as well as established cluster evaluation criteria, namely the adjusted rand index, purity, entropy and mutual information.

The choice of α significantly affects the number of clusters. Lower values of α tend to produce more clusters than exist, while higher values typically produce fewer clusters. Setting k_0 to a reasonable value (e.g., 6) or deriving it using the gap statistic is recommended to for the initial number of clusters.

The new method using Euclidean distance generally performs at least as well as when using Pearson correlation across different hyperparameter combinations.

For the following analysis, we set $\alpha = 0.1$ and $k_0 = 6$ and the Euclidean distance as distance measure. Compared to k-means across six scenarios, the new approach performs consistently as well or better when clustering information is present in missingness and value variables, with notable performance gains in scenarios 1 and 4 (see Figure 2.3). Even without the influence of missing data on cluster formation, the new approach outperforms k-means. The new method shows increased robustness, indicated by narrower error bars. In scenarios 5 and 6, the lack of clustering information provided by the value variables results in challenges for both the new approach and k-means, often yielding an underestimation of the number of clusters. This leads to reduced performance (according to the performance measurement) due to the frequent identification of only two clusters. These findings highlight the importance of hyperparameter selection.

In general, the proposed clustering method outperforms traditional methods such as k-means, PAM, and hierarchical clustering. In particular, the strength of our proposed methodology lies in the combination of information in both the observed data and the missing values. But even in the considered scenario without missing information, the methodology outperforms established standard approaches.

Applying this statistical approach to a real-world data example given by metabolomics data from the GCKD study (Eckardt et al., 2012) that focus on kidney-specific adverse events demonstrates how beneficial it can be to incorporate missingness data to identify patient risk groups. Including missing data during clustering with the newly introduced clustering procedure helps to identify at-risk patients, and therefore new patient subgroups. The findings are supported by cluster-relevant biomarkers, which turn out to be well-established risk markers. In conclusion, the proposed method can not only include informative missingness within the clustering procedure, but also supports the detection of established as well as new biomarkers for personalised medicine.

The statistical approach comes with different hyperparameters that have to be predefined. To apply the algorithm to the problem at hand, it is important to select a suitable value for the parameter α, which is influenced by the specific clustering goal. To identify unique subgroups, such

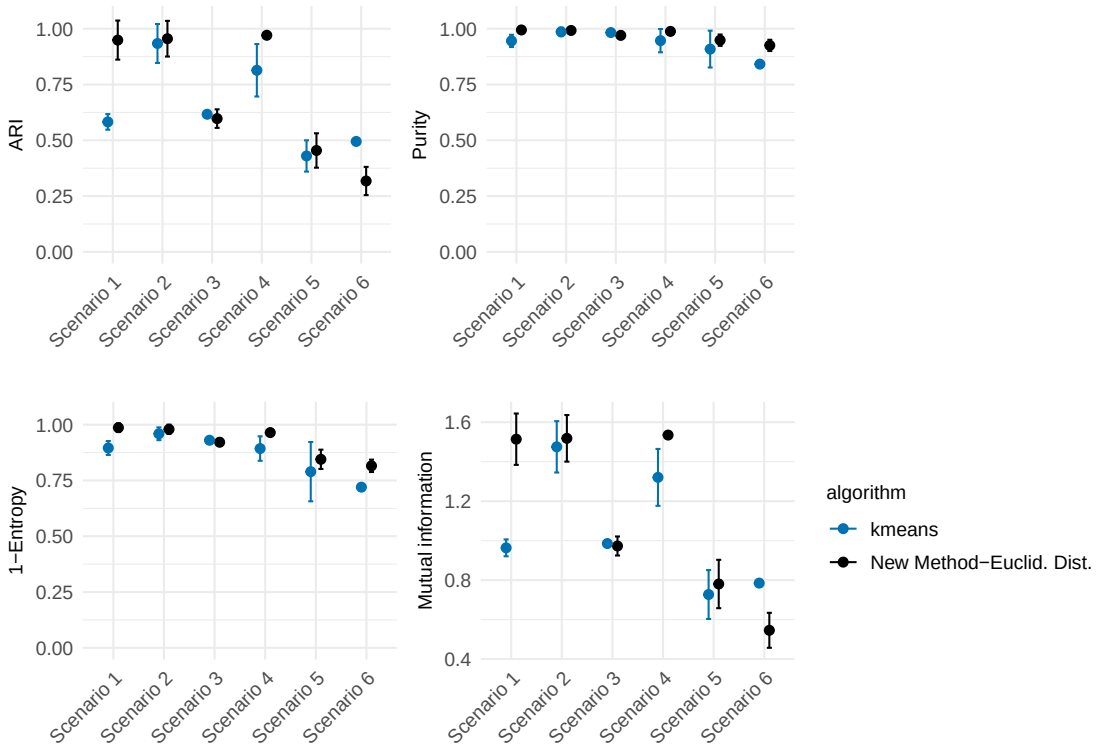

Figure 2.3: Comparison of four different performance measures (ARI, purity, entropy, mutual information) for k-means and new approach given Euclidean distance, $\alpha = 0.1, k_0 = 6$ for six different scenarios; In scenario 1 the information is solely given by the values of the markers. In scenario 6 the information is solely given by the missingness information of the markers. Scenario 2-5 reflect different graduations of missingness and value information. Big dots illustrating the empirical means $\hat{\mu}_{\{\cdot\}}$, error bars are given by $\hat{\mu}_{\{\cdot\}} \pm \hat{\sigma}_{\{\cdot\}}$. Extracted from Hunsdieck et al. (2025b).

as high-risk patients, α should be set at 0.1 or higher. In contrast, to achieve balanced groupings among all patients, α should be decreased, e.g. to 0.05. Additionally, it needs to be considered that in datasets where high noise levels can be assumed, either due to uninformative markers or inaccurate measurements, it is advisable to employ preprocessing techniques, such as principal component analysis, or to incorporate external information such as survival data for a prior selection of clustering variables.

The newly introduced method is particularly beneficial for analysing real-world data, as it helps identifying new patient subgroups that may be at risk or healthier than average. In scenarios with missing data, this method overcomes biases induced by imputation, especially when the missing data are not completely at random, and helps reveal underlying mechanisms that may be obscured

by standard preprocessing techniques.

Chapter 3

Simulation-based guidelines for required data quality in longitudinal electronic health records

Contents

3.1 Contributed materials

Hunsdieck, B., Bender, C., Ickstadt, K., Mielke, J. (2025). *Joint models in big data: Simulation-based guidelines for required data quality in longitudinal electronic health records.* PREPRINT (Version 1) available at Research Square, https://doi.org/10.21203/rs.3.rs-6031358/v1.

Authors contribution Berit Hunsdieck developed all the methods, designed and executed the simulation studies, interpreted the results, and wrote the manuscript. Johanna Mielke and Christian

Bender contributed ideas for the design of the simulation study and evaluation, and corrected and approved the manuscript. Katja Ickstadt supervised the project, contributed to the design of the simulation study and to the interpretation of the results, and corrected and approved the manuscript.

3.2 Summary

Primary care data, for example, accessed through electronic health records (EHR), can contain valuable longitudinal health information, but also pose challenges due to incompleteness, inconsistency, and inaccuracies of the data (Botsis et al., 2010). Under the assumption that small changes in biomarker levels over time can signal health changes (Zhang et al., 2012), the longitudinal information/trajectories can be used to improve the risk stratification of disease diagnoses. To take advantage of the longitudinal data provided by electronic health records, longitudinal and survival data can be combined into a single model that jointly models the outcomes. Joint models, which integrate longitudinal and survival data, have recently been highlighted as an effective tool to leverage comprehensive patient information without losing interpretability, potentially improving parameter estimates compared to traditional static survival data. Implementing a joint model requires having survival and longitudinal data that are of sufficiently high quality.

The contributed paper addresses the topic of the level of data quality, based on longitudinal real-world biomarker data, so that a joint model outperforms simpler and more established survival models like the cox model for risk assessment. These simulations have the goal of developing guidelines for assessing whether the underlying longitudinal data quality is appropriate for gaining information through a joint model.

3.2.1 Overview of methods and simulation

For the development of these guidelines, realistic longitudinal EHR data is simulated by incorporating components that reflect typical data characteristics, such as the amount of noise, the sample size and the homogeneity of the patient cohort, allowing the analysis of the influence of different characteristics on predictions.

As a starting point, we assume a cohort of patients without prior diagnoses at the start time t_{start}. It is assumed that they are recruited with a balanced design with a 50-50 split between healthy patients and patients who will develop the disease during the course of the study. Subsequently, the patients undergo a 5-year monitoring period during which both survival and longitudinal data are gathered, concluding at the time t_{end}. After the monitoring period, a 5-year follow-up period is considered with only survival data given. We assume that the longitudinal measurements are adjusted to fit within the range of $[0, 1]$, which is achievable through min-max normalisation.

For healthy patients, the diagnosis time $t_{i,abs}$ is censored at 120 months. If patients develop the disease, the time point follows a uniform distribution between 10 and 119 months:

$$t_{i,abs} \sim \begin{cases} 120 & \text{, if patient } i \text{ is healthy (censored at 120 months)} \\ U(10, 120) & \text{, if patient } i \text{ becomes sick,} \end{cases} \tag{3.1}$$

To generate meaningful longitudinal data, different aspects that affect the quality of the data in terms of their information content are addressed, namely, the sample size and the number of measurements. The individual number of measurements n_i is simulated from a normal distribution

$$n_i \sim \lfloor \mathcal{N}(n_{abs} \cdot \frac{t_{i,abs}}{12}, 2) \rfloor \tag{3.2}$$

with mean n_{abs} per year and standard deviation 2, reflecting variability in real-world measurement frequency, derived from UK Biobank EHR data.

In addition, parameters that directly affect longitudinal measurement values are considered, namely response rate, underlying noise, years of biomarker change prior to disease diagnosis, baseline measurement difference, and homogeneity within the patient group. Specifically, for subject i with n_i observations, the time-varying longitudinal biomarker measurements, observed at time point t_{ij} are represented as $y_{ij} = \{y_i(t_{ij}), j = 1, \ldots, n_i\}$. We assume, that the longitudinal measurements y_{ij} for patient i at time j are given by

$$y_{ij} = y(t_{ij}) = b_i + m_i \cdot t_{ij} \cdot 1_{t_{ij} \leq t_{i,abs} - 12 \cdot t_m} + \epsilon_{i,j} \tag{3.3}$$

assuming that the expected value remains constant until breakpoint t_m, after which the biomarker begins to change and increases linearly, indicating a forthcoming diagnosis. The individual baseline term b_i is given by

$$b_i \sim \begin{cases} \mathcal{N}(0.5 + \Delta_b, \sigma_b^2) & \text{, if patient } i \text{ is sick} \\ \mathcal{N}(0.5, \sigma_b^2) & \text{, if patient } i \text{ is healthy.} \end{cases}$$

assuming a normal distribution with different means for sick and healthy patients. The individual response term (probability of responding effect before onset) is defined by

$$p_i \sim \begin{cases} 0 & \text{, if patient } i \text{ is healthy} \\ Bern(p_{resp}) & \text{, if patient } i \text{ is sick.} \end{cases}$$

The individual slope is then given by

$$m_i \sim \begin{cases} 0 & \text{, if patient } i \text{ is healthy} \\ p_i \cdot m_i^* & \text{, if patient } i \text{ is sick} \end{cases}$$

with $m_i^* \sim \mathcal{N}(\mu_m, \sigma_m^2)$. An additional noise term $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is further assumed.

This setup allows for an individual and flexible simulation of patient data. To systematically investigate the effects of various characteristics related to the quality and quantity of longitudinal electronic health record (EHR) data on the performance of a joint modelling approach, we compare its risk prediction performance against a standard Cox model (Therneau et al., 2000).

The joint model consists of two submodels, the longitudinal model and the survival model. It can improve the precision of the estimation and the predictive performance by capturing relationships between the longitudinal and survival submodels (Tsiatis and Davidian, 2004).

For the longitudinal model, we assume that the longitudinal outcomes are normally distributed and follow a linear shape. Then, the mixed-effects model is given by

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= x_i^T(t)\beta + z_i^T(t)b_i + \epsilon_i(t) \end{aligned}$$

with fixed effects parts $x_i(t), \beta$, random effect parts $z_i(t), b_i$, and time-dependent error terms $\epsilon_i(t), \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$.

For the survival model, let C_i be the potential censoring time, $T_i^\star$ be the true event time for the i-th subject, and T_i the observed event time with $T_i = min\{C_i, T_i^\star\}$, and $\delta_i = \mathbb{1}(T_i^\star \leq C_i)$ the event indicator. The relative risk is given by

$$h_i(t|\mathcal{M}_i(t), w_i) = h_0(t) exp\left\{\gamma^T w_i + \alpha m_i(t)\right\}, \quad t > 0$$

with history of true unobserved longitudinal process up to time point t, $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$, true unobserved value $m_i(t)$ of covariate at time t, equivalent to respective part in longitudinal model, baseline risk function $h_0(t)$ at time t, (Vector of) baseline covariates w_i with coefficients vector γ of regression coefficients for baseline covariates, and effect α of underlying longitudinal outcome to the risk for an event quantifying association between time-varying covariate and risk of event.

Assuming that these two processes are associated, we can define a model for their joint distribution by assuming that we have full conditional independence; e.g., the random effects explain all interdependencies. This yields to the joint distribution

$$p(y_i, T_i, \delta_i) = \int p(y_i|b_i) \cdot \left\{h(T_i|b_i)^{\delta_i} S(T_i|b_i)\right\} p(b_i) db_i$$

with vector of random effects b_i explaining the interdependencies, and density function $p(\cdot)$ and survival function $S(\cdot)$.

Since we want to evaluate whether the estimated risk is well predicted given a reasonable time frame, the approaches are evaluated using an adjusted time-varying concordance index defined by

$$tvC_{interval}(t) = \frac{\sum_{k \in S_1} \sum_{l \in S_2} 1_{\{r_k(t) > r_l(t)\}}}{n_1 \cdot n_2} \quad \in [0, 1] \tag{3.4}$$

given subset S_1 consists of all individuals with diagnosis in interval $\left[t - interval, t\right]$ and subset S_2 consists of all individuals without diagnosis in interval $\left[0, t + interval\right]$ with $|S_1| = n_1$ and $|S_2| = n_2$.

3.2.2 Key findings and their broader research implications

The key results are based on the mean time-varying C-index during the follow-up period $[t_{end}, t_{end} + \Delta t]$ with $t_{end} = 5$ years and $\Delta t = 5$ years for patients still at risk after the observation period. The analysis examines how different scenarios affect the mean time-varying C-index over a 12-month interval. The influence of various data characteristics on prediction quality is assessed using 100 iterations of simulated data in controlled settings.

The Cox baseline model without any covariates is provided for illustration only. Since this model uses less data, it performs as expected worse than other models. Therefore, the focus is on the joint model's performance using the full longitudinal trajectory of the simulated marker over the five-year observation period, compared to the Cox model, which uses only the last measurement of the simulated marker.

The study investigates the effectiveness of the joint model compared to the Cox model in analysing the simulated EHR data, focussing on various factors that influence model performance.

The joint model outperforms the Cox model when at least one measurement is taken annually, although additional measurements beyond a certain point do not enhance performance. Therefore, at least one annual measurement is recommended for optimal use of the joint model.

Measurement noise significantly affects model performance. The joint model performs better as the noise levels increase, particularly when the noise variance exceeds approximately 0.075. Furthermore, the joint model can detect changes in biomarkers shortly before diagnosis due to its slope estimation capability, whereas the Cox model relies on the most recent measurement.

The joint model benefits more from the increase in the intercept differences, making it suitable when the intercept difference is at least 0.1. A response rate of at least 80% is necessary for the joint model to outperform the Cox model, resulting in more homogeneous subgroups and less variability. A similar conclusion can be drawn from the slope variability, where increased variability leads to less accurate results when it comes to risk prediction using joint models.

The sample size also affects the performance; at smaller sizes ($N = 50$), both models are similar, but the differences become apparent at $N = 200$. A sample size of at least 200 is recommended for reliable joint model predictions, with prediction intervals narrowing as the sample size increases. Overall, the study highlights key factors that make the use of the joint model for EHR data analysis beneficial -particularly when dealing with high level of noise, high sample size, and short-term biomarker changes, while also highlighting the importance of understanding response rates and slope variability over time. A summary of the guidelines for recommended data quality for longitudinal

primary care data is given in Table 3.1.

To apply these guidelines, it is crucial to normalise longitudinal measurements in a range of [0,1], achieving a mean value of approximately 0.5 for scaling purposes. Most parameters can be extracted directly from the relevant real-world data; however, the response rate is often unknown. Generally, to improve response rates, the cohort may be restricted to specific subgroups of patients or diseases.

Parameter	**Superior performance**
Sample Size	$N \geq 200$
Noise Standard Deviation	$\sigma_\epsilon > 0.075$
Percentage of Patients Responding	$p_{perc} \geq 80\%$
Number of measurements per year	$n_{abs} \geq 1$
Intercept difference	$\Delta_b \geq 0.1$
Slope Standard Deviation	$\sigma_m \leq 0.005$

Table 3.1: Guidelines: Criteria for Normalized Electronic Health Record Data That Preferentially Support the Joint Model Over the Cox Model. Extracted from Hunsdieck et al. (2025a).

Applying the derived guidelines to two real-world datasets, namely modelling the impact of serum bilirubin on primary biliary cirrhosis (PBC) (data source: Mayo Clinic, access through the R package JMBayes2 (Rizopoulos et al., 2024)) and the impact of eGFR on chronic kidney disease (CKD) (data source: UK Biobank (Sudlow et al., 2015), access via application 28807), shows that the predicted outcomes align with the final performance of individual models. In the first example, where all criteria are met, the joint model outperforms the Cox model in terms of the time-dependent C-index. However, in the second example, the joint model does not surpass the Cox model due to unmet conditions. Thus, the guidelines not only give a recommendation for individual model performance, but also help identify data characteristics that can be optimised for better outcomes.

Several directions for future research have been identified. First, the interactions between parameters and their mutual influences have yet to be analysed, particularly with regard to the complex interactions between factors such as noise, sample size, and measured effect. In practice, it is reasonable to assume that the measurement frequency will increase near a diagnosis. This increased measurement frequency over time, relevant for certain markers, is not currently modelled due to complexity considerations. Since the analysis was limited by computational time constraints, the results are restricted to the number of variations per parameter. A more thorough examination of individual parameters with a broader range of values could enhance the understanding of their effects and interactions, improving the robustness and reliability of the findings/ guidelines.

Moreover, for specific use cases, the joint model may require modifications as a linear slope after

a specific break point is not always given in real-world data. Furthermore, the complexity of joint models increases with multiple events and multivariate longitudinal observations (Lawrence Gould et al., 2015). Currently, the current predicted value is used as a covariate for the survival submodel, but alternative associations, such as interaction effects or time-dependent slopes, remain an additional exploration aspect (Lawrence Gould et al., 2015).

In conclusion, this work, particularly the guidelines provided, serves as a valuable resource for identifying quality issues in EHR data and determining the data quality needed for the joint model. In addition, it offers practical insights on improving data quality, which can subsequently improve the reliability and validity of joint modelling outcomes.

Chapter 4

A framework for predicting individual blood pressure trajectories post-treatment

Contents

4.1 Contributed materials

Hunsdieck, B., Mielke, J., Ickstadt, K., Elçi, E. (2025) A simulation-based framework for modeling and prediction of personalized blood pressure trajectories in hypertensive patients after antihypertensive treatment. PLOS ONE 20(4): e0318549. https://doi.org/10.1371/journal.pone.0318549

Authors contribution Berit Hunsdieck developed all the methods, designed and executed the simulation studies, interpreted the results, and wrote the manuscript. Johanna Mielke and Eren Elçi contributed ideas for the design of the simulation study and evaluation, and corrected and approved the manuscript. Katja Ickstadt supervised the project, contributed to the design of the simulation

study and to the interpretation of the results, and corrected and approved the manuscript.

4.2 Summary

With technological advances, the variety of data types continues to grow, including those from wearable devices (Eisenkraft et al., 2023; Sola et al., 2021; Kim et al., 2024). Information collected from wearables, for example smart watches, have the unique feature of being accessible not only over a continuous longitudinal time frame but also on demand for immediate read-outs and analysis. Thus, there is the possibility that clinically relevant features (e.g., pulse, blood pressure, oxygen saturation) can be extracted and used in real-time for direct health evaluation, potentially improving immediate patient care. We hypothesise that the use of longitudinal wearable data, specifically blood pressure, can help to manage medications in hypertensive patients and thus combat hypertension, the leading cause of death worldwide.

Currently, blood pressure medications are prescribed based on a static value measured during the doctor's visit. Although this is common clinical practice, this value can be biased due to many factors, such as daytime or stress factors. As the possibility of easily and continuously measuring blood pressure is still very new and the reliability of the measurements is only slowly improving, these measurements are not yet considered further by doctors. However, with technological progress and reliable and easily accessible data, this should change to ensure the best possible care for the patient, not only to overcome wrong diagnoses, but to use these data for early treatment phases, for example, during the adjustment phase of the new blood pressure medication (Schutte, 2024).

In the paper, a novel framework is presented for simulating realistic systolic blood pressure (SBP) profiles that integrate both the modelling of medication effects, as well as daily rhythms and individual patient characteristics. Based on simulated data, long-term effects of antihypertensive treatment should be forecasted by predicting steady-state responses under anti-hypertensive therapy using data collected for just a few days. For this purpose, two algorithms for forecasting the steady-state effect of the medication are introduced, namely a non-linear mixed effects model and an advanced Gaussian process, and compared with respect to their performance.

4.2.1 Overview of methods and simulation

For simulating individual blood pressure profiles, different aspects add up to the final simulated blood pressure profile, namely, the drug effect, the circadian rhythm, the uncertainty of a measurement, and the inclusion of covariate influences. These aspects will be described in detail in the following.

First, the antihypertensive treatment effect is modelled using a pharmacokinetic-pharmacodynamic (PKPD) model (Rowland and Tozer, 2019). The choice of the specific pharmacokinetic and phar-

macodynamic models is the result of a review of the literature that focusses on the identification of commonly used models in drug development for classes of antihypertensive medications. The pharmacokinetic submodel models the plasma concentration dependent on the administered dose. In our case, the model is specified via a two compartment model (Baek et al., 2008; Heo et al., 2016; Rijn-Bikker et al., 2013), given by

$$\begin{aligned}
\frac{dC(1)(t)}{dt} &= -K_a \cdot C(1)(t) \\
\frac{dC(2)(t)}{dt} &= \frac{-(K_{10} + K_{21}) \cdot C(2)(t) + K_a \cdot C(1)(t) + K_{12} \cdot C(3)(t)}{VC} \\
\frac{dC(3)(t)}{dt} &= K_{21} \cdot C(2)(t) - K_{12} \cdot C(3)(t)
\end{aligned} \tag{4.1}$$

with

$$\begin{aligned}
&C(1) : \text{Gut} \\
&C(2) : \text{Central (Plasma)} \\
&C(3) : \text{Peripheral (Tissue)} \\
K_a,\ &K_{10} : \text{Rate of absorption} \\
&K_{21} : \text{Rate of transport of the drug from the central to the peripheral compartment} \\
&K_{12} : \text{Rate of transport of the drug from the peripheral to the central compartment} \\
&K_{e0} : \text{Rate constant for transfer to effect compartment .}
\end{aligned} \tag{4.2}$$

The pharmacodynamic submodel focusses on translating the plasma concentration of the drug into the final effect. As suggested by literature (Otani et al., 2021; Larsson et al., 1990), a sigmoid E_{max}-model

$$BP_{eff} = E_{max} \cdot \frac{CE}{CE + EC_{50}} \tag{4.3}$$

is used with E_{max} as maximum effect of a specific drug, CE as the drug concentration, and EC_{50} as the half-life time, e.g. period after which half of the effect is achieved.

The corresponding parameter choices, as well as the parameter ranges, are based on a literature review focused on PKPD modelling of antihypertensive medications.

In order to effectively integrate the PKPD model with the natural fluctuations in blood pressure, it is necessary to capture the natural shape of blood pressure throughout the day. Since the blood pressure is not a static value but varies dependent on the day time, a circadian rhythm is further introduced by

$$SBP(t) = BSL + amp_1 \cdot cos(\frac{2\pi \cdot (t + hor)}{24}) + amp_2 \cdot cos(\frac{2\pi \cdot (t + hor)}{12}) \tag{4.4}$$

with

$$
\begin{aligned}
nadir &= (BSL - \frac{2}{3} \cdot change) \cdot (1 + exp(\nu)) \\
amp_2 &= \frac{1}{3} \cdot (BSL - nadir) - \frac{4}{9} \cdot change \\
&\quad - \frac{2}{9} \cdot \sqrt{6 \cdot change \cdot (nadir - BSL) + 4 \cdot change^2} \\
amp_1 &= BSL - nadir + amp_2 \quad ,
\end{aligned} \tag{4.5}
$$

motivated by Bikker et al. (Van Rijn-Bikker et al., 2013). SBP is the systolic blood pressure and BSL is the baseline systolic blood pressure. The parameters *nadir* and *change* characterise the patient: *nadir* is the minimum systolic blood pressure at night, and *change* is the day-to-night systolic blood pressure difference.

Additional within-day ($noise_{intra}$) and day-to-day variability ($noise_{inter}$) is added by using noise terms

$$
\begin{aligned}
noise_{intra} &\sim \mathcal{N}(0, (5/z_{1.99/2})^2) = \mathcal{N}(0, 1.94^2), \\
noise_{inter} &\sim \mathcal{N}(0, (8/z_{1.99/2})^2) = \mathcal{N}(0, 3.11^2).
\end{aligned}
$$

These values are obtained from literature Chia et al. (2019) combined with an internal study assessing the blood pressure profiles of eleven healthy volunteers using wearables.

Many of those parameters may dependent on individual characteristics of the patients. Therefore, individual covariate-specific impacts on the specific submodels are included based on literature references. Realistic sets of covariates were generated based on data from UK Biobank, focussing on patients with hypertension stage 2 (Whelton et al., 2018), namely a systolic blood pressure of at least 140 mmHg or a diastolic blood pressure of at least 90 mmHg without ongoing medication to reflect age and weight distributions classified by sex and ethnicity.

Based on longitudinal blood pressure data, the objective is to estimate and forecast the course of the treatment effect beginning in the early stages of antihypertensive medication. For that, we use the daily means of the measurements for the first days after treatment initiation and aim to predict the course of treatment response in subsequent days. Therefore, two approaches are introduced, a parametric and a non-parametric approach. Given patient $i \in \{1, \ldots, n_m\}$ and observation day $t \in \{0, \ldots, n_o\}$, the goal is to predict $\hat{y}_{i,t_{max}}(t)$ for a maximum of t_{max} days.

Given specific assumptions about the shape of a curve, a parametric model can be employed to fine-tune parameters of interest, either individually or at the population level. A popular approach for this is given by the non-linear mixed-effects model (Lindstrom and Bates, 1990). For this use case, the likelihood equation is motivated by Lasserson et al. (2011), given by a generalised sigmoidal

E_{max} curve

$$\hat{y}^{(nlme)}_{i,t_{max}}(t)|_{\eta_i} = exp(lR_{max,i,t_{max}}) \cdot \left(1 - exp\left(-\left(\frac{t \cdot log(2)}{exp(lt_{1/2,i,t_{max}})}\right)^{exp(l\omega)}\right)\right), \tag{4.6}$$

$$lR_{\text{max},i,t_{max}} = lR_{\text{max,pop}} + \eta_{i,1,t_{max}}, \quad lt_{1/2,i,t_{max}} = lt_{1/2,\text{pop}} + \eta_{i,2,t_{max}}$$

with individual effects $\boldsymbol{\eta}_{i,t_{max}} = (\eta_{i,1}, \eta_{i,2})^T$. The corresponding knowledge-based priors for the Bayesian framework are given by

$$lR_{\text{max,pop}} \sim \mathcal{N}(1.2, 0.1), \quad lt_{1/2,\text{pop}} \sim \mathcal{N}(0.5, 0.2), \quad l\omega \sim \mathcal{N}(1, 0.1), \tag{4.7}$$

securing reasonable non-negative trajectories and avoiding multi-modality problems.

In the absence of prior knowledge about the shape, a second non-parametric approach may be used. This model type is aiming for estimating and forecasting the curve shape only based on training data. This nonparametric approach uses an advanced Gaussian process consisting of two steps. First, the mean population curve $\hat{y}_{mean,GP}(t)$ is estimated using a Gaussian process (Rasmussen and Williams, 2008). The Gaussian process

$$\hat{y}_{mean,GP}(t)|(y_i(t), t) \sim \mathcal{N}\left(\mu_{mean}(t), \boldsymbol{\Sigma}_{mean}(t)\right) \tag{4.8}$$

given a squared exponential kernel

$$k(x, x_\star) = \sigma^2 \cdot exp\left(-\frac{(x - x_\star)^2}{2l^2}\right) \quad . \tag{4.9}$$

is fitted by optimising the log marginal likelihood

$$log\ p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{1}{2}log\ |\boldsymbol{K} + \sigma_n\boldsymbol{I}| - \frac{n}{2}log(2\pi) \quad , \tag{4.10}$$

given the training data. To estimate the individual trajectories given by the remaining residuals

$$\epsilon_i(t) = y_i(t) - \hat{y}_{mean,GP}(t) \quad , \tag{4.11}$$

an individual Gaussian process

$$\hat{\epsilon}_i(t)|S_{nearest,i,t_{max}} \sim \mathcal{N}(\mu_{indiv}(t), \Sigma_{indiv}(t)) \tag{4.12}$$

is fitted in the second step using a nearest-neighbour approach based on the training data. This nearest-neighbour approach takes into account the five most similar trajectories of the training data, determined by their Euclidean distance

$$dist_{t_{max}}(\epsilon_i, \epsilon_j) = \sqrt{\sum_{t=1}^{t_{max}} (\epsilon_i(t) - \epsilon_j(t))^2} \tag{4.13}$$

limited to the specific time frame of interest $[0, t_{max}]$.

4.2.2 Key findings and their broader research implications

In this paper, the performance of two prediction models is compared using cohorts of patients with sample sizes of $N = 60$, $N = 120$, and $N = 200$. Furthermore, various observation durations are considered, spanning from a single day to five days of measurements. To ensure robust results, we simulated and evaluated a total of 50 datasets, evaluating bias and root mean square error (RMSE). The results are displayed in Figure 4.1. Our analysis reveals that both models exhibit minimal

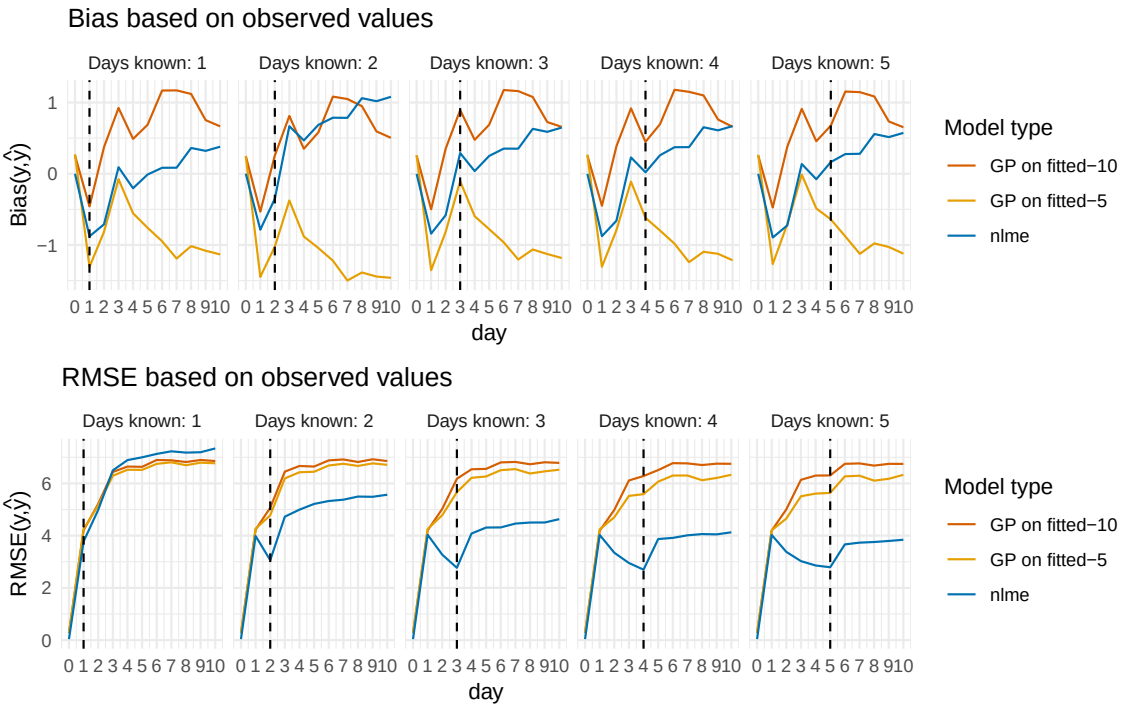

Figure 4.1: Root Mean Squared Error (RMSE) and Bias across days ($t = 0 \ldots, 10$) compared to actually observed values, models (non-linear mixed-effects model (nlme) in blue, Gaussian process (GP) with five resp. ten nearest neighbours in yellow resp. red)($t_{max} = 1, \ldots, 5$, vertical line). Extracted from Hunsdieck et al. (2025c).

bias, with a maximum deviation of approximately 1 mmHg. The NLME model seems to benefit highly from additional data, leading to a reduction in RMSE. This improvement is attributed to the model's ability to utilise both fixed and random effects across the data. In contrast, the performance of the GP model does not improve with increased data availability; it is based mainly on population means and limited data points for individual fits.

Although the GP model shows better performance for single-day data in terms of RMSE, the NLME model shows superior performance if more data is available when considering both bias and RMSE. This suggests that the GP model is preferable only in scenarios with very limited data

availability. The findings remain consistent across various sample sizes, indicating that there are no significant performance gains from adding more subjects. This observation suggests an early saturation of the model performance.

In conclusion, while the GP model is more suitable for minimal data scenarios, the NLME model demonstrates superior performance if more data is available. Overall, both models exhibit increased performance levels as the sample size increases, highlighting important considerations for their application in predictive modelling.

The integration of statistical methods into wearable devices presents a novel solution to monitor medication efficacy in hypertensive patients. It can predict if the desired reduction in blood pressure will be achieved within the next days, allowing the timely adjustment of the medication if necessary. There is increasing interest in forecasting the effects of antihypertensive drugs, supported by studies using machine learning to predict response to therapy (Mroz et al., 2024). This approach improves monitoring, improves treatment effectiveness, and promotes better blood pressure control, marking a significant advancement in personalised healthcare for hypertension management. Future steps involve testing algorithms on real-world data to assess their performance and reliability.

Chapter 5

Discussion and outlook

This thesis addresses challenges that occur when dealing with real-world data, namely metabolomics and phenotype data, interspersed with noise from external sources and missingness. The three articles summarised in this thesis focused on different aspects of the use of real-world biomarker data for patient risk assessment.

The first manuscript proposed a method to take into account missingness mechanisms that are not at random for improving patient clustering. Instead of just imputing missing values, missingness itself is incorporated in the clustering process, to capture both the value and missingness information for clustering. The application of this algorithm to the metabolomics data from the GCKD study highlighted its efficacy in identifying subgroups of patients at risk and well-established as well as new biomarkers. The second manuscript develops guidelines on required data quality for the use of a representative of complex models for longitudinal biomarker data, the joint model. It compares joint models combining longitudinal and survival data with Cox models, identifying conditions for improved risk predictions. Two real-world examples demonstrate the applicability in practice of the guidelines. The third manuscript focusses on the prediction of long-term effects of antihypertensive drugs based on high-frequency measurements during the first days of treatment. For that, a framework is presented to simulate blood pressure profiles, integrating the effects of medications, daily rhythms, and patient characteristics. For the prediction task, we compare two statistical approaches: a nonlinear mixed-effects model (nlme) and a Gaussian process (GP). The GP model excels with limited data, while the nlme model performs better with more available data.

In summary, the need for advanced tailored statistical methodologies is essential to generate insights from real-world data. Innovative approaches, such as those presented in this thesis, have the potential to significantly enhance data analysis and interpretation. Each of the three projects deals with the topic of real-world biomarker data, highlighting that different aspects and issues of

real-world data should be considered, depending on the specific research question.

In this thesis, we study the properties of established and proposed statistical methodologies by simulation. This is the gold standard for statistical method development since the ground truth is known. Although the individual projects introduced earlier include advanced comprehensive simulations, additional aspects can always be taken into account, depending on the assumptions made. For example, when simulating cluster/ marker values, it is possible to not only rely on fixed cluster-specific marker means, but to extend the simulation by incorporating correlation and interdependency terms.

While the proposed methodologies are presented separately in the different contributed papers, it is also possible to combine the approaches. Data-driven clustering of patient cohorts (Chapter 2) could help identify more homogeneous cohorts, which ultimately results in better results when it comes to risk prediction (3) or modelling (Chapter 4). Therefore, it may be beneficial to first apply the clustering approach for integrating informative missingness introduced in Chapter 2 prior to the longitudinal joint modelling approach.

In addition to the individual, paper-specific aspects which were already mentioned in the corresponding chapters, the topic of real-world biomarker data offers several directions for improvements and extensions and future research topics.

It is essential to address specific challenges, particularly in the realm of clustering and longitudinal analysis. In real-world scenarios, researchers often face uncertainty when missingness is present. This uncertainty complicates the interpretation of the clustering results. When no missingness is present, clustering is based completely on the available values, which may lead to a more straightforward analysis. However, when dealing with missing data, the potential for bias increases, as the clustering outcomes may reflect the patterns of the missing data rather than the true underlying structure. Therefore, explicitly stating these limitations not only enhances the transparency of the research but also guides future studies in addressing these challenges.

Especially in the context of precision medicine, a comprehensive understanding of the characteristics of the individual patient is important. By combining different data types, RWE can be used to identify well-defined patient cohorts, new targets for drug development, and personalised treatment strategies. All of the methods discussed have the potential to be transferred to multiple data types to improve the precision of patient evaluation. For example, genetics can be incorporated for drug safety or suitability testing for a patient, while data types other than genetic information also influence therapeutic outcomes, including epigenetic variables, age, medication use, lifestyle choices, and sex of the patient (Duffy, 2015).

Using the example of incorporating continuous values and missingness into clustering analyses, various omics layers (e.g., proteomics, transcriptomics, genomics) can be integrated. This integration

allows for the consideration of not only different omics types during the clustering process but also enables the modelling of the information content of each respective data layer.

Using the example of individual blood pressure predictions, the approach can be extended to multivariate time series. This allows simultaneous analysis of multiple interrelated variables over time, enhancing the ability to capture complex dynamics and interactions among different physiological parameters.

By integrating multiple data types, this can lead to a broader understanding of patient health and disease mechanisms. Currently, there are approaches for the integration of multimodal data, such as evolutionary random forest clustering (Bi et al., 2020). However, these current methods often operate as black boxes, limiting interpretability and understanding of the underlying processes. In future research, interpretable methods for multimodal data could be developed based on the methods introduced in this work.

Moreover, the methodologies proposed in this thesis, are not without limitations. The reliance on simulation studies, although a robust approach, may not fully capture the complexities and variability of real-world data. The effectiveness of the proposed methods can be influenced by the specific characteristics of the data sets used, such as sample size, heterogeneity, and the presence of confounding factors. Future research should focus on validating these methodologies in diverse real-world settings to ensure their generalisability and robustness.

Prior to an application in practice, it is essential to verify that the results are stable and can be validated with external data sources. Therefore, a sufficient sample size is important. While biobanks typically offer large sample sizes, if one is focussing on a homogeneous subpopulation or specific indication of interest, these may still be too small for meaningful analysis. By increasing the sample size, the final results will be more robust. This can be achieved, for example, using upcoming initiatives for EHR recruitment, such as Our Future Health (2022). This not only enables an increase in the overall sample size, but also allows for an indication-specific increase in data density from a longitudinal perspective. Since the results of the data analysis may be highly influenced by the underlying cohort, other data sources and biobanks should be considered for validation. For example, one can consider the Finnish cohort *FinnGen* (FinnGen, 2020) or the American cohort *All of Us* (All of Us Research Program Investigators, 2019) for validation. Although establishing clear quality guidelines across different biobanks has been challenging due to various subsequent use cases, in the past decade the importance of centralised biobank infrastructures has been recognised globally. This led to substantial efforts in creating and managing the global network of biobanks guaranteeing well-defined patient samples of high quality (Hummel and Specht, 2019). This not only improves working with biobanks but will make quality control (e.g., standardisation, harmonisation) much easier for future work, leading to an increased use of biobanks for decision making.

In conclusion, the methodologies offer meaningful solutions for patient risk assessment and can potentially be adapted for other applications. The integration of advanced statistical methodologies and innovative approaches in the analysis of real-world biomarker data not only addresses the inherent challenges of noise and missingness, but also paves the way for better patient evaluation and personalised treatment strategies, ultimately contributing to the advancement of precision medicine and the identification of novel therapeutic targets.

Bibliography

All of Us Research Program Investigators. 2019. The "All of Us" research program. *New England Journal of Medicine* 381, 7 (2019), 668–676.

Michelle Alvarado, Mark Lawley, and Yan Li. 2016. *Healthcare simulation tutorial: Methods, challenges, and opportunities.* IEEE.

In-hwan Baek, Min-hyuk Yun, and Kwang-il Yun, Hwi-yeoland Kwon. 2008. Pharmacokinetic/pharmacodynamic modeling of the cardiovascular effects of beta blockers in humans. *Archives of Pharmacal Research* 31 (2008), 814 – 821.

Olaf Beckonert, Jürgen Monnerjahn, Ulrich Bonk, and Dieter Leibfritz. 2003. Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 16, 1 (2003), 1–11.

Xia-an Bi, Xi Hu, Hao Wu, and Yang Wang. 2020. Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. *IEEE Journal of Biomedical and Health Informatics* 24, 10 (2020), 2973–2983.

Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. 2010. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translational Bioinformatics* 2010 (March 2010), 1–5.

Julie K Bower, Sejal Patel, Joyce E Rudy, and Ashley S Felix. 2017. Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Current epidemiology reports* 4 (2017), 346–352.

Andrea M Burden. 2019. Pharmacoepidemiology and big data analytics: challenges and opportunities when moving towards precision medicine. *Chimia* 73, 12 (2019), 1012–1012.

Yook-Chin Chia, Kazuomi Kario, Naoko Tomitani, Sungha Park, Jinho Shin, Yuda Turana, Jam Chin Tay, Peera Buranakitjaroen, Chen-Huan Chen, Satoshi Hoshide, and et al. 2019. Com-

parison of day-to-day blood pressure variability in hypertensive patients with type 2 diabetes mellitus to those without diabetes: Asia BP@Home study. *The Journal of Clinical Hypertension* 22, 3 (2019), 407–414. https://doi.org/10.1111/jch.13731

Thomas M Cover. 1999. *Elements of information theory.*

Miroslava Čuperlović-Culf, Nabil Belacel, Adrian S Culf, Ian C Chute, Rodney J Ouellette, Ian W Burton, Tobias K Karakach, and John A Walter. 2009. NMR metabolic analysis of samples using fuzzy K-means clustering. *Magnetic Resonance in Chemistry* 47, S1 (2009), S96–S104.

Simon de Lusignan and Chris van Weel. 2006. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family practice* 23, 2 (2006), 253–263.

Kieu Trinh Do, Simone Wahl, Johannes Raffler, Sophie Molnos, Michael Laimighofer, Jerzy Adamski, Karsten Suhre, Konstantin Strauch, Annette Peters, Christian Gieger, et al. 2018. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* 14 (2018), 1–18.

David J. Duffy. 2015. Problems, challenges and promises: perspectives on precision medicine. *Briefings in Bioinformatics* 17, 3 (08 2015), 494–504. https://doi.org/10.1093/bib/bbv060 arXiv:https://academic.oup.com/bib/article-pdf/17/3/494/6687486/bbv060.pdf

Kai-Uwe Eckardt, Barbara Bärthlein, Seema Baid-Agrawal, Andreas Beck, Martin Busch, Frank Eitner, Arif B Ekici, Jürgen Floege, Olaf Gefeller, Hermann Haller, et al. 2012. The German chronic kidney disease (GCKD) study: design and methods. *Nephrology Dialysis Transplantation* 27, 4 (2012), 1454–1460.

Arik Eisenkraft, Nir Goldstein, Roei Merin, Meir Fons, Arik Ben Ishay, Dean Nachman, and Yftach Gepner. 2023. Developing a real-time detection tool and an early warning score using a continuous wearable multi-parameter monitor. *Frontiers in Physiology* 14 (2023), 1138647.

FinnGen. 2020. *FinnGen Documentation of R3 release.* https://finngen.gitbook.io/documentation/

Richard A Gibbs. 2020. The human genome project changed everything. *Nature Reviews Genetics* 21, 10 (2020), 575–576.

Wilson Wen Bin Goh and Limsoon Wong. 2020. The birth of bio-data science: trends, expectations, and applications. *Genomics, proteomics & bioinformatics* 18, 1 (2020), 5–15.

Sebastien Haneuse and Michael Daniels. 2016. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *EGEMs* 4, 1 (2016), 1203.

Young-A Heo, Nick Holford, Yukyung Kim, Mijeong Son, and Kyungsoo Park. 2016. Quantitative model for the blood pressure-lowering interaction of valsartan and amlodipine. *British Journal of Clinical Pharmacology* 82, 6 (2016), 1557–1567. https://doi.org/10.1111/bcp.13082 arXiv:https://bpspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/bcp.13082

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2 (1985), 193–218.

Michael Hummel and Cornelia Specht. 2019. Biobanks for future medicine. *Journal of Laboratory Medicine* 43, 6 (2019), 383–388. https://doi.org/10.1515/labmed-2019-0106

Berit Hunsdieck, Christian Bender, Katja Ickstadt, GCKD Investigators, and Johanna Mielke. 2025b. Leveraging missing information in clustering: A novel method integrating continuous values and informative missingness. *Submitted to Biometrical Journal.* (2025).

Berit Hunsdieck, Christian Bender, Katja Ickstadt, and Johanna Mielke. 2025a. Joint Models in Big Data: Simulation-Based Guidelines for Required Data Quality in Longitudinal Electronic Health Records. *BioData Mining* 18 (2025). https://doi.org/10.1186/s13040-025-00450-z

Berit Hunsdieck, Johanna Mielke, Katja Ickstadt, and Eren Elçi. 2025c. A simulation-based framework for modeling and prediction of personalized blood pressure trajectories in hypertensive patients after antihypertensive treatment. *PLOS ONE* 20, 4 (04 2025), 1–20. https://doi.org/10.1371/journal.pone.0318549

Joseph G Ibrahim and Geert Molenberghs. 2009. Missing data methods in longitudinal studies: a review. *Test* 18, 1 (2009), 1–43.

Hyun Kang. 2013. The prevention and handling of the missing data. *Korean journal of anesthesiology* 64, 5 (2013), 402–406.

JM Kendall. 2003. Designing a research project: randomised controlled trials and their principles. *Emergency medicine journal* 20, 2 (2003), 164–168.

Hyunsoo Kim and Haesun Park. 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23, 12 (2007), 1495–1502.

Jihoon Kim, Sung-A Chang, and Seung Woo Park. 2024. First-in-Human Study for Evaluating the Accuracy of Smart Ring Based Cuffless Blood Pressure Measurement. *Journal of Korean medical science* 39, 2 (2024).

Inke R König, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V Kopp. 2017. What is precision medicine? *European respiratory journal* 50, 4 (2017).

Rutger Larsson, Bengt E. Karlberg, Agneta Gelin, Jan Åberg, and Carl-Gunnar Regårdh. 1990. Acute and steady-state pharmacokinetics and antihypertensive effects of felodipine in patients with normal and impaired renal function. *The Journal of Clinical Pharmacology* 30, 11 (1990), 1020–1030. https://doi.org/10.1002/j.1552-4604.1990.tb03589.x

Daniel S. Lasserson, Thierry Buclin, and Paul Glasziou. 2011. How quickly should we titrate antihypertensive medication? Systematic review modelling blood pressure response from trial data. *Heart* 97, 21 (2011), 1771–1775. https://doi.org/10.1136/hrt.2010.221473

A. Lawrence Gould, Mark Ernest Boye, Michael J. Crowther, Joseph G. Ibrahim, George Quartey, Sandrine Micallef, and Frederic Y. Bois. 2015. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* 34, 14 (2015), 2181–2195. https://doi.org/10.1002/sim.6141 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6141

Jackson O Lay Jr, Rohana Liyanage, Sabine Borgmann, and Charles L Wilkins. 2006. Problems with the "omics". *TrAC Trends in Analytical Chemistry* 25, 11 (2006), 1046–1056.

Mary J. Lindstrom and Douglas M. Bates. 1990. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* 46, 3 (1990), 673–687. http://www.jstor.org/stable/2532087

Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data.*

Jared R Mayers, Chen Wu, Clary B Clish, Peter Kraft, Margaret E Torrence, Brian P Fiske, Chen Yuan, Ying Bao, Mary K Townsend, Shelley S Tworoger, et al. 2014. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nature medicine* 20, 10 (2014), 1193–1198.

Hunter N.B. Moseley. 2013. Error analysis and propagation in metabolomics data analysis. *Computational and Structural Biotechnology Journal* 4, 5 (2013), e201301006. https://doi.org/10.5936/csbj.201301006

Thomas Mroz, Michael Griffin, Richard Cartabuke, Luke Laffin, Giavanna Russo-Alvarez, George Thomas, Nicholas Smedira, Thad Meese, Michael Shost, and Ghaith Habboub. 2024. Predicting hypertension control using machine learning. *PloS one* 19 (03 2024), e0299932. https://doi.org/10.1371/journal.pone.0299932

MingMing Ning and Eng H Lo. 2010. Opportunities and challenges in omics. *Translational stroke research* 1 (2010), 233–237.

Janet E Olson, Suzette J Bielinski, Euijung Ryu, EM Winkler, Paul Y Takahashi, Jyotishman Pathak, and James R Cerhan. 2014. Biobanks and personalized medicine. *Clinical genetics* 86, 1 (2014), 50–55.

Yuki Otani, Hidefumi Kasai, and Yusuke Tanigawara. 2021. Pharmacodynamic analysis of hypertension caused by lenvatinib using real-world postmarketing surveillance data. *CPT: Pharmacometrics & Systems Pharmacology* 10, 3 (2021), 188–198. https://doi.org/10.1002/psp4.12587

Our Future Health. 2022. Our Future Health. https://www.ourfuturehealth.org.uk Accessed: 2025-04-14.

Alice Park. 2009. 10 Ideas Changing the World Right Now. *Time* (2009). https://content.time.com/time/specials/packages/article/0,28804,1884779_1884782_1884766,00.html

Barbara Parodi. 2015. *Biobanks: A Definition.* Springer Netherlands, Dordrecht, 15–19. https://doi.org/10.1007/978-94-017-9573-9_2

Emanuel F Petricoin, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, et al. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *The lancet* 359, 9306 (2002), 572–577.

Pharmaceutical Technology. 2024. *Big data in the pharmaceutical industry: analyzing innovation, investment and hiring trends.* Pharmaceutical Technology. https://www.pharmaceutical-technology.com/data-insights/big-data-in-pharma

Carl Edward Rasmussen and Christopher K Williams. 2008. *Gaussian processes for machine learning.* MIT Press.

Eugene P Rhee, Susan Cheng, Martin G Larson, Geoffrey A Walford, Gregory D Lewis, Elizabeth McCabe, Elaine Yang, Laurie Farrell, Caroline S Fox, Christopher J O'Donnell, et al. 2011. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. *The Journal of clinical investigation* 121, 4 (2011), 1402–1411.

Petra C. Van Rijn-Bikker, Oliver Ackaert, Nelleke Snelder, Reinier M. Van Hest, Bart A. Ploeger, Richard P. Koopmans, and Ron A. A. Mathot. 2013. Pharmacokinetic-pharmacodynamic modeling of the antihypertensive effect of eprosartan in black and white hypertensive patients. *Clinical Pharmacokinetics* 52 (2013), 793–803. https://doi.org/10.1007/s40262-013-0073-6

Dimitris Rizopoulos, Grigorios Papageorgiou, and Pedro Miranda Afonso. 2024. *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data.* https://drizopoulos.github.io/JMbayes2/, https://github.com/drizopoulos/JMbayes2.

Malcolm Rowland and Thomas N Tozer. 2019. *Clinical Pharmacokinetics and Pharmacodynamics* (5 ed.). Wolters Kluwer Lippicott, Philadelphia.

Donald B Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

Syed Salleh, Praveen Thokala, Alan Brennan, Ruby Hughes, and Andrew Booth. 2017. Simulation modelling in healthcare: an umbrella review of systematic literature reviews. *PharmacoEconomics* 35 (2017), 937–949.

Judi Scheffer. 2002. Dealing with missing data. *Research Letters in the Information and Mathematical Sciences* 3 (2002), 153–160.

Aletta E. Schutte. 2024. Wearable cuffless blood pressure tracking: when will they be good enough? *Journal of Human Hypertension* 38, 9 (2024), 669–672. `https://doi.org/10.1038/s41371-024-00932-3`

Josep Sola, Anna Vybornova, Sibylle Fallet, Erietta Polychronopoulou, Arlene Wurzner-Ghajarzadeh, and Gregoire Wuerzner. 2021. Validation of the optical Aktiia bracelet in different body positions for the persistent monitoring of blood pressure. *Scientific Reports* 11, 1 (2021), 20644.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* 12, 3 (03 2015), 1–10. `https://doi.org/10.1371/journal.pmed.1001779`

Terry M Therneau, Patricia M Grambsch, Terry M Therneau, and Patricia M Grambsch. 2000. *The cox model*.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 2 (2001), 411–423.

Stephanie Titze, Matthias Schmid, Anna Köttgen, Martin Busch, Jürgen Floege, Christoph Wanner, Florian Kronenberg, Kai-Uwe Eckardt, GCKD Study Investigators, Kai-Uwe Eckardt, et al. 2015. Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrology Dialysis Transplantation* 30, 3 (2015), 441–451.

Anastasios A Tsiatis and Marie Davidian. 2004. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* (2004), 809–834.

Mario Vailati-Riboni, Valentino Palombo, and Juan J Loor. 2017. What are omics sciences? *Periparturient diseases of dairy cows: a systems biology approach* (2017), 1–7.

Petra C. Van Rijn-Bikker, Nelleke Snelder, Oliver Ackaert, Reinier M. Van Hest, Bart A. Ploeger, Richard P. Koopmans, and Ron A. A. Mathot. 2013. Nonlinear Mixed Effects Modeling of the Diurnal Blood Pressure Profile in a Multiracial Population. *American Journal of Hypertension* 26, 9 (2013), 1103–1113. https://doi.org/10.1093/ajh/hpt088

Runmin Wei, Jingye Wang, Mingming Su, Erik Jia, Shaoqiu Chen, Tianlu Chen, and Yan Ni. 2018. Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific reports* 8, 1 (2018), 663.

Paul K. Whelton, Robert M. Carey, Wilbert S. Aronow, Donald E. Casey, Karen J. Collins, Cheryl Dennison Himmelfarb, Sondra M. DePalma, Samuel Gidding, Kenneth A. Jamerson, Daniel W. Jones, and et al. 2018. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/apha/ash/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* 71, 6 (2018). https://doi.org/10.1161/hyp.0000000000000065

David S Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, et al. 2012. HMDB 3.0—the human metabolome database in 2013. *Nucleic acids research* 41, D1 (2012), D801–D807.

Po-Yen Wu, Chih-Wen Cheng, Chanchala D Kaddi, Janani Venugopalan, Ryan Hoffman, and May D Wang. 2016. –Omic and electronic health record big data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering* 64, 2 (2016), 263–273.

Haomin Yang, Yudi Pawitan, Fang Fang, Kamila Czene, and Weimin Ye. 2022. Biomarkers and disease trajectories influencing women's health: results from the UK biobank cohort. *Phenomics* 2, 3 (2022), 184–193.

Jin Yang, Yuanjie Li, Qingqing Liu, Li Li, Aozi Feng, Tianyi Wang, Shuai Zheng, Anding Xu, and Jun Lyu. 2020. Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine* 13, 1 (2020), 57–69. https://doi.org/10.1111/jebm.12373 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jebm.12373

Daoqiang Zhang, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative. 2012. Predicting Future Clinical Changes of MCI Patients Using Longitudinal and Multimodal Biomarkers. *PLOS ONE* 7, 3 (03 2012), 1–15. https://doi.org/10.1371/journal.pone.0033182

Publications

Leveraging missing information in clustering: A novel method integrating continuous values and informative missingness

Berit Hunsdieck[1,2*], Christian Bender[1], Katja Ickstadt[2,3], GCKD Investigators, Johanna Mielke[1]

[1*]Computational Biology, Bayer AG, Wuppertal, Germany.
[2]Statistics Faculty, TU Dortmund University, Dortmund,Germany.
[3]Lamarr-Institute for Machine Learning and Artificial Intelligence, Dortmund,Germany.

*Corresponding author(s). E-mail(s): berit.hunsdieck@bayer.com;
Contributing authors: christian.bender@bayer.com;
ickstadt@statistik.tu-dortmund.de; johanna.mielke@bayer.com;

Abstract

Clustering approaches can only produce meaningful results if all relevant information is available. Missing data can be an important contribution, though it is often neglected. In this paper, we present a novel method for clustering data where missing data is explicitly considered by integrating both observed values and information about missing data. More concretely, we propose a two-step approach in which the first clusters are derived and optimised based on continuous measurements. Afterwards, obtained clusters may be subdivided based on information obtained from missing data if these are considered informative.

We show in a simulation study that our approach can achieve better results than established standard clustering procedures (k-means, PAM, hierarchical clustering) in the evaluation measures of adjusted rand index (ARI), entropy, purity, and mutual information.

We demonstrate the applicability in practice by clustering patients from the German Chronic Kidney Disease study, a prospective cohort study of patients with chronic kidney disease (CKD), based on metabolite data. Our approach can identify high-risk clusters for CKD disease progression, which would have been overlooked if missing data would have only been imputed and not explicitly included in the data matrix for clustering.

Keywords: Missing data, informative missingness, mixed data type clustering, real-world data

1 Introduction

Patient subtyping, i.e. the identification of subpopulations of patients with a joint disease mechanism, is of high importance for the development of novel treatments. Treatments developed for a specific endotype of interest could more directly treat larger effect sizes and more homogeneous effects [1]. The identification of such subpopulations remains challenging. Diverse data types, such as omics data, e.g. metabolomics, transcriptomics, or proteomics data, lead to a better description of the underlying biological mechanisms [2]. Since these sub-populations are not yet described, hypothesis-free approaches are the method of choice for describing such disease sub-populations. In this paper, we present a novel approach for hypothesis-free discovery of such disease sub-populations.

When experimental data are measured in a high-throughput fashion, for example, untargeted metabolomics data, missing information is typically encountered [3]. It is well known that missingness can have different underlying causes: it can occur at random (MAR) or completely at random (MCAR) if, for example, there were problems with the sample preparation or technical issues [4]. On the other hand, it can also be not at random (MNAR) caused by a systematic issue if, for example, the measured concentration was below the lower limit of detection, LOD [5, 6]. While MCAR occurs when missing data is entirely unrelated to any values (observed or missing), MAR occurs when missingness is related to observed data, but not to the missing values themselves. In mass spectometry applications, Wei et al. discussed various approaches for handling missing values by imputation, i.e. exchanging missing value with plausible replacements [7]. The imputation method can vary for each type of missingness. While for MCAR and MAR random forest imputation is recommended, for MNAR, methods such as quantile regression imputation of left-censored data are preferred [7]. However, the type of missingness is often unknown, or is a mixture of these categories. After imputation, imputed data are typically treated in the same way as observed data, and missigness is not further considered in standard follow-up analysis [8]. Since for some of the experimental data, the rate of missingness can be high ($\geq 50\%$) and the missingness very informative, an imputation strategy may not reflect the biological meaning of the data. We propose to handle the missingness information transparently by explicitly including it in the clustering algorithm. For that, we modify the data set by introducing an indicator variable that reflects the missing status of the biomarker. This results in two layers per biomarker: one that gives the actual observed or imputed value (continuous variable) and the other one that describes the missingness status (binary variable).

There are already some approaches for handling paired variables where one variable is continuous and one variable is binary [9]. If the binary missingness feature is taken into account, the standard approach for handling mixed data types is given by a clustering based on the distance matrix using the Gowers distance [9], where the distances are scaled by the range of each variable. However, this approach has the difficulty that, when including binary and continuous variables, the distance range given by the continuous values differ substantially from the distance range given by the binary variables. This is particularly problematic for the presence of missing data, as this may lead to a bias within the distance matrix. Such an imbalance can skew the

analysis, potentially leading to false conclusions about the relationships and variations among the clusters. Therefore, it is essential to consider both missingness patterns and the range of values in an appropriate manner, to accurately interpret the data and maintain the robustness. Optimising the weighting for clustering, as done in previous studies [10], is not optimal, since it leads to clusters based on missing or non-missing data. However, we aim to make use of both types of information.

We hypothesise that missingness itself can bring additional information gain for the clustering of patients. Our goal is to consider missingness information for clustering, rather than focussing on the imputation. Given a population of patients with measured markers, where some of the measurements are missing, we aim to present a clustering algorithm which not only imputes the missing values, but also takes the original information of missingness into account. This way, in the final clustering the different causal influences of values and missingness on clusters can be modelled. For that, we introduce a stepwise approach where the different layers of information are sequentially included in the clustering by first clustering based on the observed and imputed observations and fine-tuning the clustering afterwards based on the missingness status. The superiority of the proposed approach is demonstrated in a simulation study. We also demonstrate the applicability in practice by clustering patients with Chronic Kidney Disease (CKD) based on metabolomics data obtained from the German Chronic Kidney Disease cohort (GCKD, [11]).

2 Methodology

The proposed methodology is a stepwise approach which is able to adequately take into account both the actual observed (or imputed) data, but also the information on missingness of data. In contrast to classic clustering algorithms like k-means [12], our proposed methodology avoids problems of handling various data types (here: continuous measurement- and binary missingness status-data).

We assume that a data set X is given in which some variables are missing for some subjects. Given M imputed variables accompanied by information on missing values and N observations, every variable $v_m(n) = (v_m^{(val)}, v_m^{(mis)}),\ m = 1, \ldots, M$, for observation $n, n = 1, \ldots, N$, can be split into

1. value information $v_m^{(val)}(n)$: containing the actual or imputed value
2. missingness information $v_m^{(mis)}(n)$: containing information about the missingness status with

$$v_m^{(mis)}(n) = \begin{cases} 0 & \text{, if the actual value is measured} \\ 1 & \text{, if the actual value is missing} \end{cases} \tag{1}$$

Since the specific choice of imputation methodology (e.g., mean imputation, nearest neighbour imputation [13], MICE [14]) is not our focus of research, we assume that the data have already been imputed and $v_m^{(val)}$ no longer contain missing data. We also assume that data has been cleaned and preprocessed and the data is normalised (e.g., 0-1/min-max normalisation).

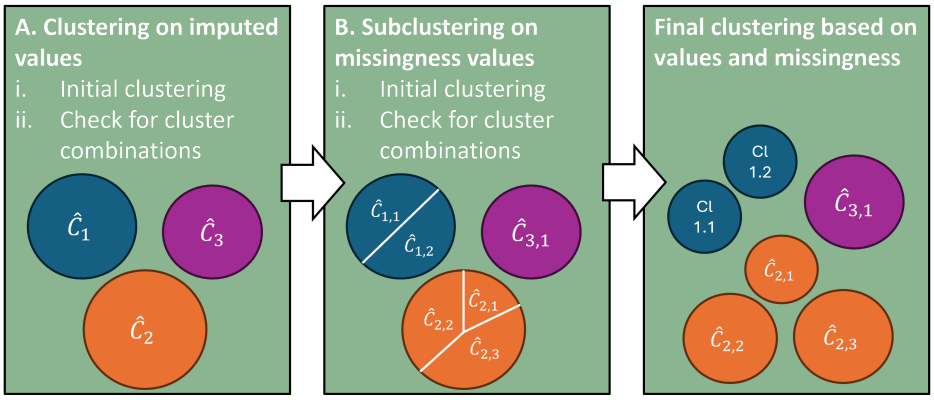

Fig. 1 Algorithm Scheme: The new clustering scheme is devided into two steps, focussing on the value and the missingness information separately.

In the following, we first describe in detail the proposed approach, before we introduce the specifics of the simulation study. It should be emphasised that a simulation study is necessary here, since for hypothesis-free clustering approaches, the ground truth is typically not known. The generation of artificial data facilitates the assessment of the approach. The simulated data are oriented around metabolomics data, where many markers are measured for individuals, but with highly varying amounts of missingness.

2.1 Clustering Procedure

The general scheme is visualised in Figure 1. Details on the individual steps are given in the following sections. In the first step, we focus on the measured or imputed values. An initial clustering is performed on those values (Step A). Afterwards, it is checked if any of the clusters can be merged to improve the compactness and separation of the clusters. Next, we focus on the missingness status: using the clustering obtained within Step A, we focus on each cluster separately, and within the obtained cluster, perform an additional clustering that only takes into account the missingness information (Step B). In this step, we may generate meaningful subclusters of the clusters from Step A that we may have overlooked if we had ignored the missingness information. This is followed by a step in which it is checked if the clustering can be improved by merging parts of the clusters. Steps A and B follow the same structure: first, there is an initial clustering step, and then a cluster combination step. These are described in the following sections in detail. The final clusters will be labelled as $\hat{C}_{x,y}$. Here, x indicates the value-based cluster, while y denotes the subcluster determined by the missingness within the value cluster.

For introducing the different steps, we define variable set A as

$$A = \{v_1^{(val)}, \ldots, v_M^{(val)} : v_m^{(val)} \in \mathbb{R}^n, m = 1, \ldots, M\} \tag{2}$$

as set of M variables containing value information and variable set B as

$$B = \{v_1^{(mis)}, \ldots, v_M^{(mis)} : v_m^{(mis)} \in \{0,1\}^n, m = 1, \ldots, M\} \tag{3}$$

as set of M variables containing missingness information.

2.1.1 Initial Clustering

In this step, an initial clustering is carried out separately on the measurements (Step A) or missingness information (Step B). It should be emphasised that in Step B, multiple subclusters are derived based on the results of Step A, i.e., the clustering is performed separately within each cluster resulting from Step A. In both cases, the k-means algorithm [12] is used due to its status as a standard procedure in cluster analysis. Its performance is demonstrably superior to other standard clustering algorithms like partitioning around medoids (PAM) and hierarchical clustering for this particular use case, as evidenced by the results and the comparison between the new method and the established clustering procedures (see Section 3.2).

For k-means, it is necessary to pre-specify the initial number of clusters k_0. This can either be fixed or derived using a gap statistics approach proposed by Maechler [15]. The implementation is given by the R package *cluster* [15].

2.1.2 Cluster Combination

Starting with a preliminary clustering result according to section 2.1.1, it remains unclear whether this clustering effectively differentiates the data. The primary goal in this step is to assess the derived clusters for their "uniqueness", specifically examining whether the clusters are compact and distinctly separated from one another.

To assess the (internal) validity of the initially derived clusters, we compute the average distances both within the clusters and between them. Given $\hat{k}$ derived clusters, the average distance matrix $D = (d_{ij})_{i,j \in \{1,\ldots,\hat{k}\}}$ will be computed by calculating the average distance between and within the clusters $\hat{C}_1, \ldots, \hat{C}_{\hat{k}}$. The average distance of two clusters $\hat{C}_i = \{x_1, \ldots, x_{n_1}\}$ and $\hat{C}_j = \{y_1, \ldots, y_{n_2}\}$ is given by

$$d_{ij} = \frac{1}{n_1 \cdot n_2} \sum_{s=1}^{n_1} \sum_{t=1}^{n_2} d(x_s, y_t) \tag{4}$$

where the pairwise distance $d(x_s, y_t)$ of two observations $x_s = (x_{s,1}, \ldots, x_{s,M})$ and $y_t = (y_{t,1}, \ldots, x_{t,M})$ can be defined as the euclidean distance

$$d_{L^2}(x_s, y_t) = \sum_{m=1}^{M} (x_{s,m} - y_{t,m})^2 \tag{5}$$

or using the distance based on Pearson correlation coefficient

$$d_{cor}(x_s, y_t) = 1 - |Cor(x_s, y_t)|$$

$$= 1 - \left| \frac{\sum_{m=1}^{M}(x_{s,m} - \bar{x}_s)(y_{t,m} - \bar{y}_t)}{\sqrt{\sum_{m=1}^{M}(x_{s,m} - \bar{x}_s)^2}\sqrt{\sum_{m=1}^{m}(y_{t,m} - \bar{y}_t)^2}} \right| \in [0, 1]. \quad (6)$$

Distances are calculated separately in Step A and B. If the distance between clusters is not greater than the distance within individual clusters, it may be appropriate to merge the clusters.

For that, the ratio

$$r_{i,j} = \frac{min(d_{i,i}, d_{j,j})}{d_{i,j}} \quad (7)$$

of the minimal average within-cluster distance and the average between-cluster distance will be evaluated between cluster $\hat{C}_i$ and $\hat{C}_j$. If there is no difference in the clusters, the corresponding ratio $r_{i,j}$ is numerically close to 1. In order to assess if the observed data is compatible with the hypothesis that the ratio matches 1 (then clusters need to be merged), an equivalence test is performed, i.e. we test

$$H_0 : R_{i,j} \neq 1 \text{ vs } H_1 : R_{i,j} = 1.$$

For this test, a confidence interval is calculated. We determine the empirical distribution $\hat{R}_{i,j}$ of the ratio between two clusters $\hat{C}_i$ and $\hat{C}_j$ by resampling. Resampling has the advantage that no statistical assumptions need to be met, which is particularly beneficial when dealing with real-world data. Given the empirical distribution $\hat{R}_{i,j}$, the confidence interval $PI_{i,j}$ for the ratio of cluster $\hat{C}_i$ and $\hat{C}_j$ is given by

$$PI_{i,j} = [q_{p,\hat{R}_{i,j}}, q_{1-p,\hat{R}_{i,j}}] \quad (8)$$

with $q_{p,\hat{R}_{i,j}}$ denoting the p-quantile of distribution $\hat{R}_{i,j}$.

The clusters $\hat{C}_i$ and $\hat{C}_j$ will merge if the corresponding prediction interval is contained in the equivalence interval $[1 - \alpha, 1 + \alpha]$, $\alpha \in [0, 1]$ (in step A: $\alpha = \alpha_1$; in step B: $\alpha = \alpha_2$). The process will be carried out for all cluster combinations in parallel. If there are multiple options for the merging of the clusters, the process will be carried out iteratively, beginning with the highest mean empirical ratio $\mu_{\hat{R}_{\cdot,\cdot}}$ of the resampling datasets.

When choosing a small α (e.g. $\alpha = 0.05$), the analysis will favour the formation of more unique clusters, as the merging criteria will be more stringent. In contrast, increasing α will tend to produce larger clusters, in which a cluster will only remain distinct if there is a truly significant difference compared to other clusters. This approach allows for a nuanced balance between cluster compactness and distinctiveness, depending on the chosen significance level. Depending on the level of information within the individual sets of variables A and B, α can be chosen according to the set of relevant variables.

2.2 Evaluation via external validation measures

To assess the effectiveness of this clustering method, it is crucial to compare its results with the ground truth (available for simulated datasets). Consequently, we use external validation metrics for evaluation.
Suppose that we are given L (true) subgroups, while the clustering method generates K clusters. For each combination of clusters sets $\hat{C} = \{\hat{C}_1, \ldots, \hat{C}_K\}$ and $C = \{C_1, \ldots, C_L\}$, the corresponding numbers are given in the Appendix, Table 2.

For cluster evaluation, we will focus on four validation measures, namely the adjusted rand index (ARI) [16], purity [17], entropy [17], and mutual information [18]. Our objective is to achieve high ARI, purity, and entropy while maintaining low entropy. For clarity, we will display $1 - entropy$ in all following Figures. Then, a high value is desirable for all measurements. Detailed definitions of the individual evaluation measures are given in in the Appendix, section S1.

2.3 Simulation of clusters

For our simulation study, we assume a population X with n individuals, information on M markers for each individual includes value information and missingness information (1 for missing, 0 for not missing). To simulate clusters, it is assumed that among the M markers, there are m_1 informative markers and m_2 cluster-unspecific markers such that $M = m_1 + m_2$. Cluster-unspecific markers represent non-informative markers (noise). We assume that all cluster-specific markers follow the same underlying distribution within one cluster.

For our simulation study, we select $M_1 = 100$ informative markers and $M_2 = 20$ non-informative markers measured on $n = 1900$ subjects with data organised into $k = 5$ clusters. For the cluster-unspecific marker, a mean value $\mu_{val} = 100$ and a missingness probability of $p_m = 0.5$ is assumed. For the cluster-specific markers, we select six scenarios (see supporting information, Table 4). The corresponding cluster-specific mean values μ_{val} and missingness probabilities p_m are illustrated in Figure 2. In the first scenario, the underlying clusters are only split by value information. In the sixth scenario, the underlying clusters are only split by missingness information. In scenarios 2 to 5, the underlying clusters are split by value *and* missingness information, with rising influence of missingness and descending influence of value information for the formation of the cluster. We simulate 100 datasets and compare the performance of our proposed approaches with three common clustering algorithms k-means, hierarchical clustering using complete linkage method, and PAM [19].

For comparison, k-means, hierarchical clustering, and PAM will have missingness and value information as input. The new clustering approach is evaluated for $\alpha \in \{0.05, 0.1\}$. The initial number of clusters is either derived from the gap statistic or set to a fixed value (6 or 10). The distance measure used for calculating the average distance matrix is either given by the Euclidean distance or using the Pearson correlation coefficient.

For values originally missing, the missing values are imputed using predictive mean matching implemented via the R package "mice" [14].

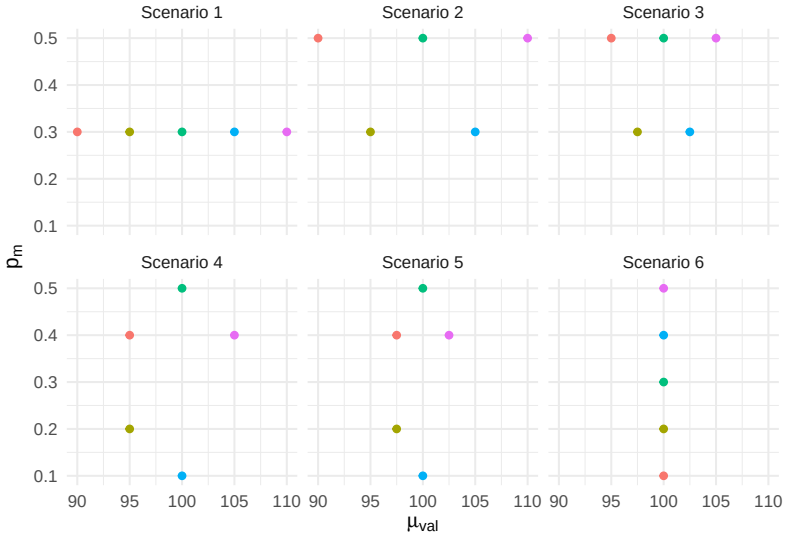

Fig. 2 Visualisation of cluster means μ_{val} and p_m for seven different scenarios that are analysed further, see Table 4 for details.

2.4 Software versions

The following software has been used:
R Version [20]: 4.4.1
Package Versions: dplyr [21]: 1.1.4; ggplot2 [22]: 3.5.1; patchwork [23]: 1.3.0; purrr [24]: 1.0.2; tidyr [25]: 1.3.1; NbClust [26]: 3.0.1; mice [14]: 3.17.0; factoextra [27]: 1.0.7; mclust [28]: 6.1.1

3 Results

First, we give a concrete walk-through example based on simulated data (see Section 3.1) to demonstrate the individual steps of the algorithm. In Section 3.2, we analyse the generalisability of the performance by looking at 100 different simulated datasets, evaluating and comparing the results with other clustering procedures using external validation measures (see Section 2.2). Finally, in section 3.3 we apply the algorithm on the metabolite data given by the German study of chronic kidney disease (GCKD) [11].

3.1 Walk-through example

Assume that we have $N = 1900$ patients divided into $L = 5$ clusters with means and cluster sizes according to the Appendix, Table 3. As outlined in section 2.3, for each cluster, we simulate for each subject $M_1 = 100$ informative markers based on the chosen cluster means (see Appendix, Table 3). The final distribution of the marker means per cluster is visualised in the Appendix, Figure 9, where the dots give the average values per cluster$(\mu_{val}^{(m)}, \mu_{mis}^{(m)})$ of marker $m, m = 1, \ldots, M$.

Missing data is imputed using predictive mean matching. The (imputed) values will be normalised using a Min-Max-Normalisation. The following walk-through is based on $\alpha = 0.1$, i.e., clusters are merged if the prediction interval falls within $[0.9, 1.1]$. Since we assume the same number of informative markers for the value and the missingness, we do not choose a different α for the first and second merging step. The initial numbers of the clusters will be set to $k_0 = 6$ with the Euclidean distance as a distance measure.

Step A: Clustering on values with merging step

In Step A of the clustering procedure (see Figure 1), we perform an initial clustering based on the imputed values. Results are shown in Figure 3. Comparing the ground truth clusters with the initial cluster assignments, we note that a_4 and a_5 mainly include observations from the original cluster C and E. When calculating the corresponding empirical distribution $\hat{R}_{a\_4,a\_5}$, the prediction interval $PI_{\hat{R}_{a\_1,a\_6}} \approx [0.977, 0.981]$ is included in the equality interval $[0.9, 1]$, which results in a merge of these two clusters to the new cluster a'_1. Iteratively following this procedure, the clusters a_2 and a_3 will be merged to the cluster a'_2, a_1 and a_6 will be merged to cluster a'_3, and finally a'_1 and a'_2 will be merged to cluster a'_4, so that in the end the two clusters a'_3 and a'_4 are remaining based on the continuous values. In Figure 3 the distribution of the initial clusters (A) together with their pairwise distances (B) as well as the distribution of the joint clusters (C) and the distribution of the empirical ratio for the hypothesis of merging a'_3 and a'_4 (D) are illustrated.

Step B: Subclustering on missingness with merging step

After clustering based on the values, we proceed with subclustering on the missingness. In Step B of the clustering procedure (see figure 1), we perform an initial sub-cluster given $k_0 = 6$ clusters based on the missingness information within each of the derived clusters a'_3 (displayed on the upper panels in the figure) and a'_4 (displayed on the lower panels in the figure). Analogously to Step A, we merge iteratively, resulting in two subclusters for cluster a'_3 and two subclusters for cluster a'_4. As a final result, we obtain 4 clusters as shown in the right panels of Figure 4. The first cluster $a'\_3 + b'\_1$ contains nearly exclusively observations from original cluster B, the second cluster $(a'\_3 + b'\_4)$ represents the original cluster D, the third one $(a'\_4 + b'\_2)$ is a mixture of cluster C and E. The last cluster $(a'\_4 + b'\_4)$ is mostly representing original cluster A. Considering the Appendix, Figure 9, the lack of separation for cluster C and E is not unexpected, since the mean values of the markers and the missingness patterns also overlap.

It is already clear that the clustering specifications, such as the choice of α, can influence the clustering results. When looking for more distinct clusters, α should be increased. In the next section, the influence of α and other hyperparameters with respect to the clustering performance will be analysed.

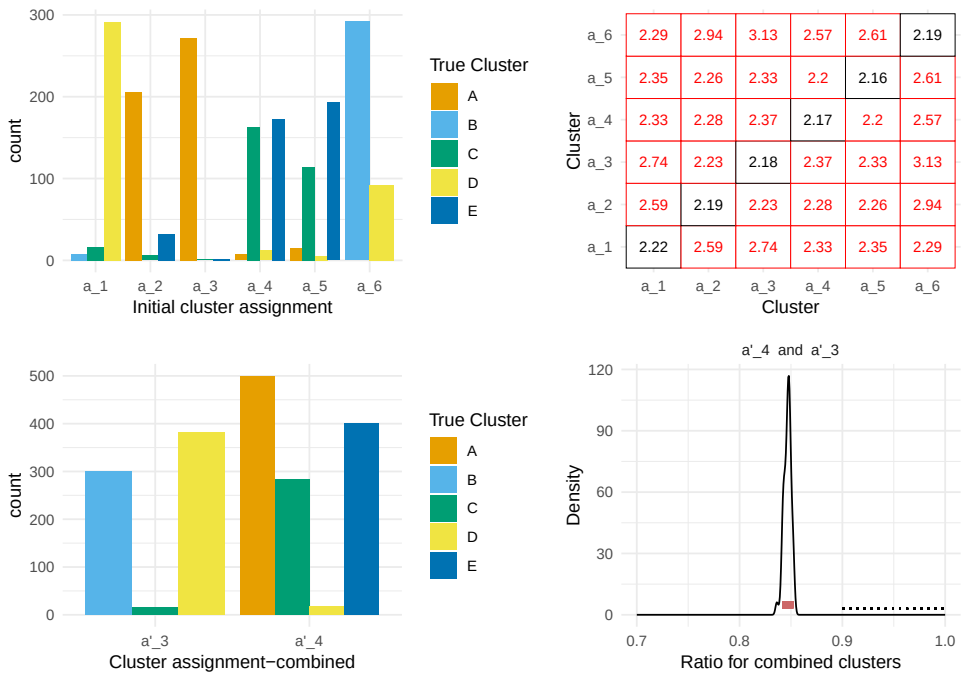

Fig. 3 Distributions of original and derived clusters based on value information: A: Distribution of original clusters within initially derived clusters; B: Within and between distances of initially derived clusters; C: Distribution of original clusters within merged clusters; D: Empirical distribution of pairwise ratios of merged derived clusters

3.2 Evaluation of clustering performance across simulated datasets

The generalisability of the results is evaluated by simulating data sets as explained in Section 2.3 comprising $M_1 = 100$ informative markers and $M_2 = 20$ non-informative markers with five underlying clusters. To facilitate a discussion on the algorithm's performance under the appropriate hyperparameter selection (α, k_0, distance measure), six scenarios are analysed with varying influence of missingness and values on cluster formation (see Section 2.3). The new algorithm will be evaluated given the Euclidean or correlation distance, $\alpha \in \{0.05, 0.1\}$ and $k_0 \in \{6, 10\}$ or derived from the gap statistic.

We first examine the influence of the different settings on the final number of derived clusters, as illustrated in the Appendix, Figure 10. Note that the true number of underlying clusters is 5. Setting the initial number of clusters to a fixed high value (here: $k_0 = 10$), the new clustering approach is no longer able to join clusters and tends to a high number of derived clusters, if the values have less impact on cluster formation (see scenarios 1-5). This can be compensated by using a higher α. For $\alpha = 0.05$, more clusters are generated than actually exist in the dataset, while for

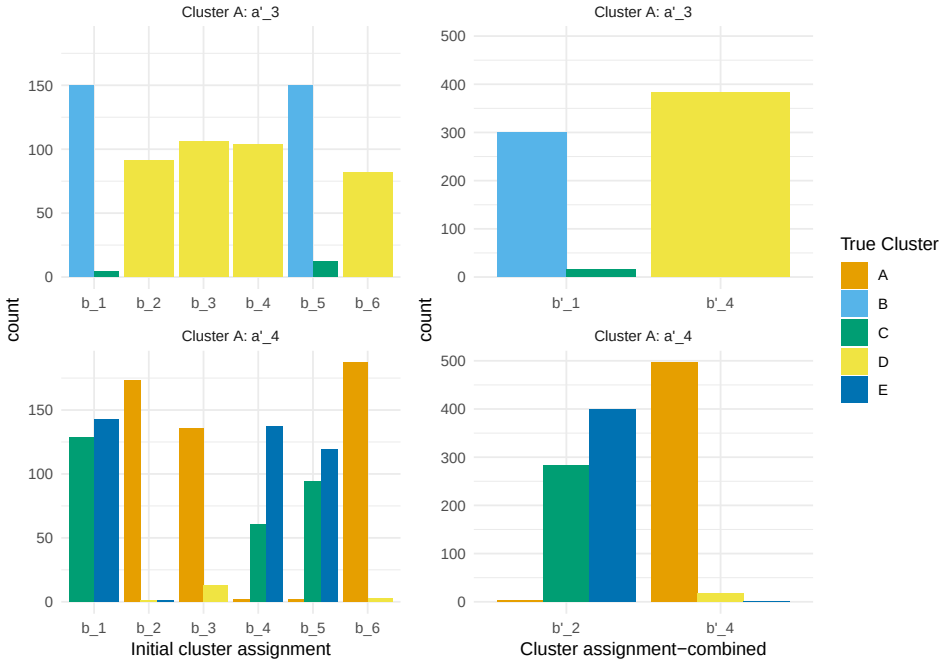

Fig. 4 Distributions of original and derived clusters based on missingness information: Left: Distribution of original clusters within initially derived clusters; Right: Distribution of original clusters within merged clusters. The right side displays the final result of the clustering.

$\alpha = 0.1$, fewer clusters are typically produced. This indicates that α is an important parameter influencing the results. Using the Gap statistic is a good alternative for the preselection of the initial cluster number. If the clusters are solely driven by the missingness information, the algorithms underestimate the correct number of clusters, resulting in two clusters in almost all cases.

The choice of the clustering method should depend on its performance, evaluated using internal and external cluster validation methods to focus on different aspects of the clusters. For example, the Adjusted Rand Index (ARI) is an external performance measure used to assess clustering results by comparing them with true cluster labels. In contrast, purity and entropy focus less on the external performance of the clustering than on the quality of the clusters itself, namely the homogeneity within the clusters and their purity. When comparing the three established clustering approaches, we note that for the chosen data setup, k-means is at least as good as hierarchical clustering and PAM in terms of ARI, purity, entropy and mutual information (see Appendix, Figure 11). Therefore, we will focus on the performance of the introduced approach compared to k-means taking into account values and missingness.

Comparing the performance of the proposed method using the Pearson correlation as the distance matrix versus the Euclidean distance for different choices of α and k_0 (see Figure 5), we note that the Euclidean distance demonstrates at least a

performance similar to the Pearson correlation approach for almost all combinations considered. That is why we focus on the Euclidean distance as distance measurement in the following analysis. It is also evident that the optimal performance is predominantly achieved with the parameters $\alpha = 0.1$ and $k_0 = 6$, leading to the choice of this parameter configuration for further analysis.

Therefore, we will further compare the proposed approach given the Euclidean distance with $\alpha = 0.1$ and $k_0 = 6$ to k-means for the six different scenarios. The direct comparison of the four different evaluation measures is shown in Figure 6. Looking at ARI, the new approach is always performing at least as good as k-means except for scenario 6. We note that even without the effect of missing data on cluster formation, the new approach still outperforms k-means. A large performance gain of at least 0.1 is achieved using the new approach for scenario 1 and 4. When comparing the configuration of various scenarios, it is evident that for some scenarios, the method performs similarly to k-means, especially when the initial cluster centres μ_{val} of the values are closer together. If there is an appropriate distance both in terms of values and missingness, our approach outperforms k-means. Similar observations can be made when looking at the other performance measures. In addition, the narrowing error bars indicate that the new approach is more robust when it comes to clustering. Scenario 5 and 6 seem to be more challenging for the approaches. A more in-depth examination of scenarios 5 and 6 reveals that both algorithms fail to accurately identify the correct number of clusters in these cases, frequently resulting in only two derived clusters. Consequently, the outcomes of the performance indices are not indicative, as in both cases, limited significance is provided by the presence of only two clusters.

Therefore, we conclude that the strength of our proposed methodology lies in the combination of information in both the observed data and the missing values.

3.3 Clustering of chronic kidney disease using metabolomic data

The GCKD study is a prospective observational cohort study of 5217 patients with chronic kidney disease treated by nephrologists. The study has previously been described in detail [29]. In brief, patients were included if they had an eGFR between 30 and 60 $ml/min/1.73\ m^2$ or a urinary albumin-to-creatinine ratio (UACR) greater than $300mg/g$ [29]. Patients completed a medical assessment, including blood and urine sampling, every two years. In addition, urine and plasma metabolites were measured at the beginning of the study using untargeted mass spectrometry. In this paper, we focus on the urine metabolomics study in this paper. This study has previously been described in detail, including information on sample processing and data cleaning [30]. In total, 1020 metabolites with at least one missing value are available.

Since the number of measured metabolites is large compared to the sample size, we first reduce the number of metabolites by only including those with prognostic value for kidney disease progression into the clustering approach (both value and missingness status). For that, we apply a Cox survival model [31], using days to temporary or permanent dialysis or kidney transplantation as the outcome and metabolites as covariates (including missingness and value as features). Observations are censored if no event was observed within a follow-up period of 5 years. More concretely, the

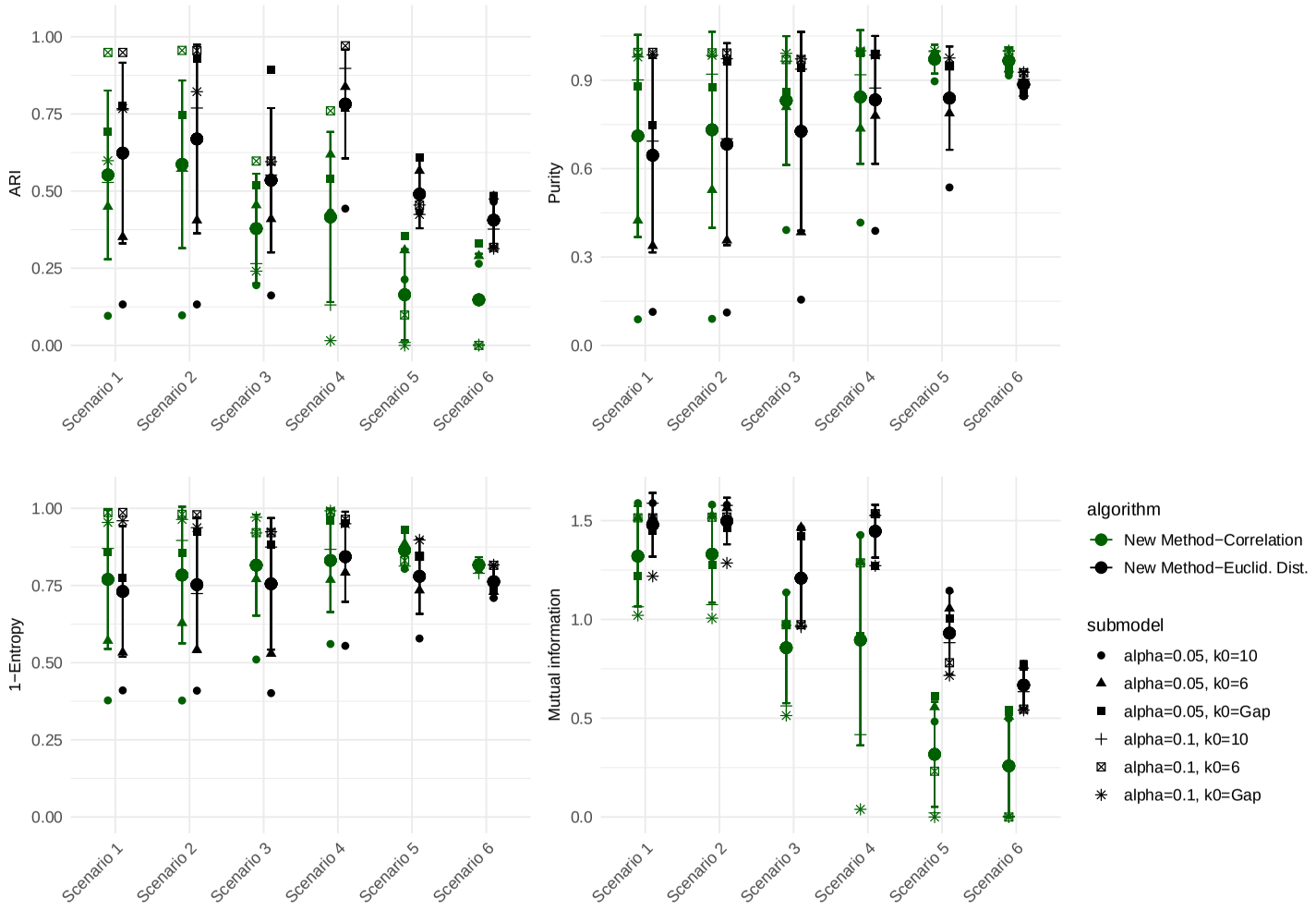

Fig. 5 Comparison of four different performance measures (ARI, purity, entropy, mutual information) for new approach given Euclidean distance vs. correlation distance given different hyperparameter constellations, evaluated for six different scenarios; In scenario 1 the information is solely given by the values of the markers. In scenario 6 the information is solely given by the missingness information of the markers. Scenario 2-5 reflect different graduations of missingness and value information. The big dots illustrating the empirical means $\hat{\mu}_{\{\cdot\}}$ error bars are given by $\hat{\mu}_{\{\cdot\}} \pm \hat{\sigma}_{\{\cdot\}}$.

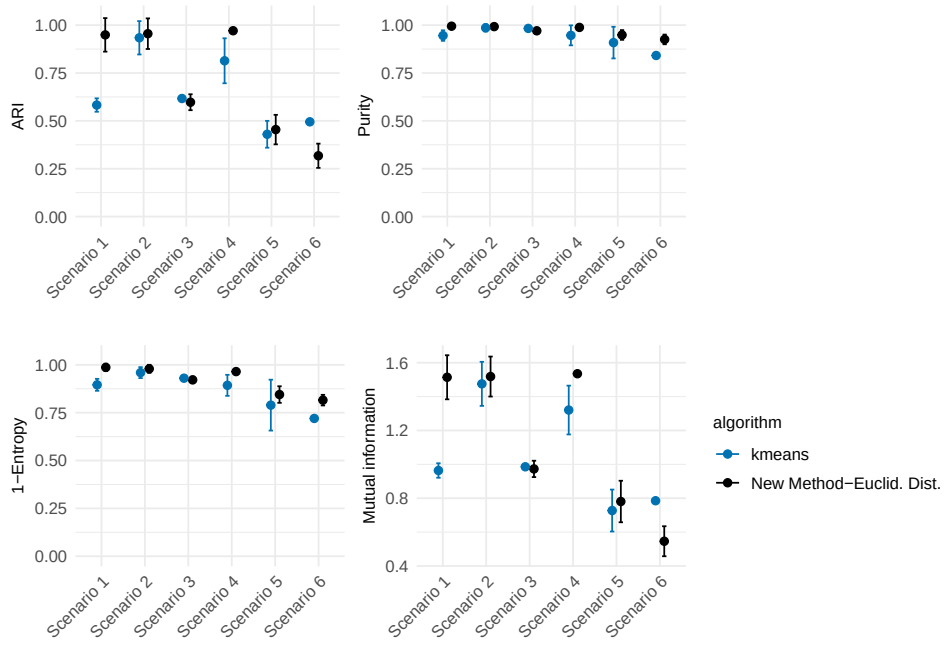

Fig. 6 Comparison of four different performance measures (ARI, purity, entropy, mutual information) for k-means and new approach given Euclidean distance, $\alpha = 0.1, k_0 = 6$ for six different scenarios; In scenario 1 the information is solely given by the values of the markers. In scenario 6 the information is solely given by the missingness information of the markers. Scenario 2-5 reflect different graduations of missingness and value information. Big dots illustrating the empirical means $\hat{\mu}_{\{\cdot\}}$, error bars are given by $\hat{\mu}_{\{\cdot\}} \pm \hat{\sigma}_{\{\cdot\}}$.

following model was used:

$$h(m,t) = h_0(t) \cdot exp\{\gamma_1 \cdot v_m^{(val)}(n) + \gamma_2 \cdot v_m^{(mis)}(n)\} \tag{9}$$

with $h_0(t)$ is the baseline hazard function at time t, $\gamma = (\gamma_1, \gamma_2)$ represents the vector of regression coefficients and $v_m(n) = (v_m^{(val)}(n), v_m^{(mis)}(n))$ are the measurements and missingness status, respectively $(m, m = 1, \ldots, M)$. We select all metabolites with a p-value of 0.05 or lower for the value *and* the missingness association for further clustering. No multiplicity correction was used. Metabolites where the observed values are highly correlated (Pearson correlation greater than 0.8) are excluded by including the metabolite with the highest association, namely the lowest p-value, on survival per correlated group. In total, 122 metabolites remain.

The aim of this case study is to identify distinct groups with medium-to-large size, i.e., we aim to tend to merge early. Therefore, we select $\alpha_1 = 0.1$ and $\alpha_2 = 0.15$. An initial number of $k_0 = 6$ clusters is chosen to maintain a reasonable total number of clusters. As distance measure, the Euclidean distance is chosen.

For step A, three clusters $\hat{C}_1, \hat{C}_2$ and $\hat{C}_3$ of sizes 507, 306 and 3376 are derived. After the sub-clustering step B, in total 10 derived clusters are given. Details on the obtained clustering, including sample sizes per cluster, can be found in Table 5. In the following, we will focus on clusters of a size greater than 50 subjects. The seven final clusters for further characterisation are called $\hat{C}_{1,1}, \hat{C}_{1,2}, \hat{C}_{2,1}, \hat{C}_{2,2}, \hat{C}_{3,1}, \hat{C}_{3,2}$ and $\hat{C}_{3,3}$.

In the following, our objective is to characterise the identified clusters based on available phenotypic information and disease progression trajectories. For that, we first analyse the Kaplan-Meier survival curves [32] with adverse renal events, namely temporary or permanent dialysis or kidney transplantation, as endpoint. This is illustrated in Figure 7. We note that cluster $\hat{C}_{1,2}$ has the highest risk of disease progression.

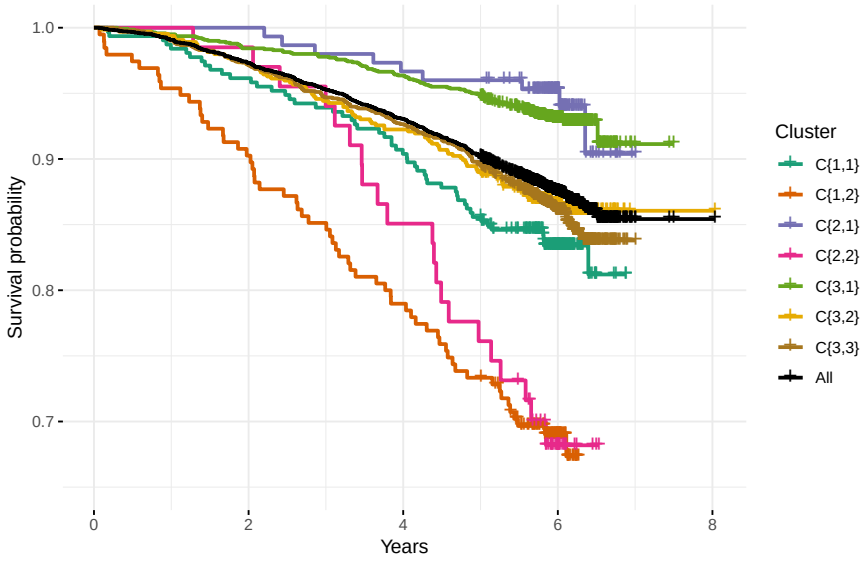

Fig. 7 Survival curves for each derived cluster and for the full population (in black), $C\{x, y\}$ denotes cluster $\hat{C}_{x,y}$

Specifically, the risk is much higher than in cluster $\hat{C}_{1,1}$. If considering the imputed data only and ignoring the missingness pattern, one would not have been able to separate these patients into high and low risk groups. Thus, missingness information substantially improved assignments into biologically diverse and heterogeneous clusters. A similar observation can be made by considering $\hat{C}_{3,1}, \hat{C}_{3,2}$ and $\hat{C}_{3,3}$. All of these fall into one cluster when considering the imputed data only. However, as shown in Figure 7, the risk of $\hat{C}_{2,1}$ and $\hat{C}_{3,1}$ is substantially lower than for the other subgroups. Again, we conclude that the missingness information contains relevant information for subtyping.

To better understand the patient characteristics of the different clusters, we analyse the corresponding biomarker levels based on a heatmap using log_2-fold change given the full population as reference. A selection of markers is shown in Figure 8. The complete heat map containing all biomarkers analysed is given in the Appendix, figure

12. Within the high-risk cluster $\hat{C}_{1,2}$, a range of known biomarkers for kidney disease progression are observed to be up-regulated. For example, C-reactive protein (CRP) levels [33, 34], copeptin [35], galectin-3 levels [36], and urea levels [37] are increased. In addition, the estimated glomerular filtration rate (eGFR) [38] is lower than in the full population.

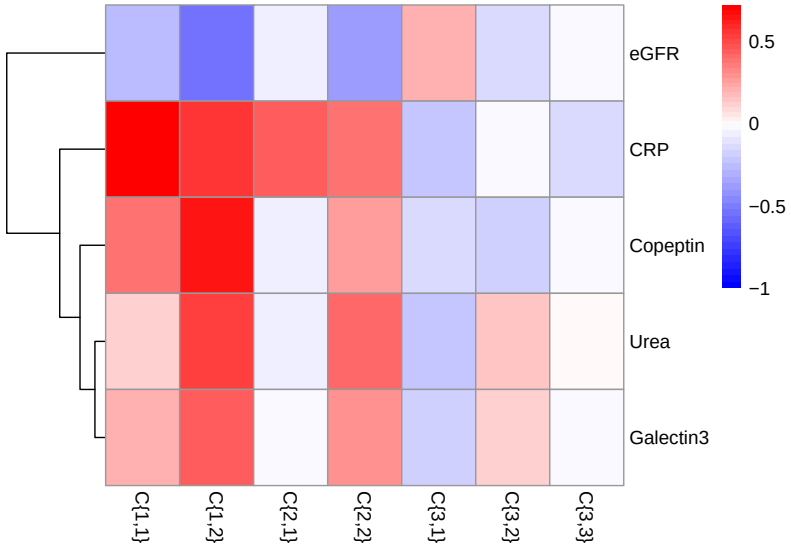

Fig. 8 Heatmap illustrating the log2 fold change of outstanding biomarkers across each cluster. The color intensity represents the magnitude of change, with blue indicating upregulation and red indicating downregulation of biomarkers relative to the full cohort. $C\{x, y\}$ denotes cluster $\hat{C}_{x,y}$.

Analysing metabolite levels in each cluster (see Appendix, figures 13 and 14) suggest that certain markers might be MNAR since the missingness status varies drastically between clusters.

4 Discussion and outlook

In real-world data analysis, e.g., in metabolomics and electronic health records (EHR), missing data is a major challenge. If this missingness is missing at random (MAR), imputation methods help eliminate missing values so that data analysis can be based on imputed values. If missingness is not at random (MNAR), such as when influenced by device limitations or specific measurements, using basic imputation methods can skew the data, leading to biased results [39].

In this paper, we focus on clustering approaches and propose a method that is able to integrate both observed values and information about missing data in an unbiased way. This specifically means that no distinction between MNAR and MAR is needed. This is achieved not by merely aggregating all elements, but through a two-step process that appropriately handles each variable type. When informative missingness is

present, this approach can identify markers with informative missingness and use this knowledge for efficient clustering. Our algorithm is capable of identifying information sources and surpasses traditional clustering methods like k-means, hierarchical clustering, and PAM, particularly when data contains information both in missingness and value.

We also demonstrate the applicability in practice by clustering patients from the German Chronic Kidney Disease Cohort based on metabolite data. In this example, we demonstrate that including the missingness information leads to the identification of high-risk clusters for disease progression that would otherwise have been overlooked.

The method requires, in addition to the selection of an initial number of clusters, the choice of α for the equivalence margins in the cluster combination step, depending on the clustering goal. If the goal is to leave only highly distinct subgroups (such as a specific high-risk patient subgroup, tendency towards large clusters), α should increase to 0.1 or greater. However, if the objective is to keep also small clusters, α should be in $[0, 0.1]$. It should be noted that typically the ground truth is not known. However, the impact of α w.r.t. the obtained subgroups can be analysed considering the phenotype profile and risk trajectories, as demonstrated in the case study.

Another important question is whether data should be pre-filtered to reduce the dimensionality prior to clustering. This certainly depends on the use case, the number of subjects included, and the number of markers measured. However, as a rule of thumb, we recommend using preprocessing or marker selection methods such as principal component analysis or incorporate external information such as survival information when dealing with data sets where many non-informative markers are suspected.

So far, we have assessed the performance of the methodology for specific data situation (e.g., normally distributed data). As an area of future research, it would be interesting to assess the performance in data with different underlying distributions.

It should be noted that while we present the method for the use case of missing data, in principle it also works for other scenarios where paired bivariate data (continuous and binary observation) need to be clustered. The proposed clustering approach could be extended to other types of combination, e.g., count data and continuous data or count data and binary data. Also, conceptually it would be possible to extend it to higher dimensions, i.e. the approach could run iteratively across e.g. three or four different data layers. However, one may need to allow for merging of clusters from the first steps to avoid a large number of small-sized clusters. The construction of such a procedure could be a field of future research.

To summarise, in this paper we present a clustering methodology that can deal with data sets in which variables have a varying degree of missing observation. Since the missingness is explicitly modelled, we account for missingness at random or completely at random - or even a combination. We have demonstrated in a simulation study and a case study that this approach can lead to more meaningful clusters.

Ethics

GCKD Study was approved by the local ethics committees and registered in the national registry for clinical studies (DRKS 00003971). Between 2010 and 2012, 5217 eligible adult patients provided written consent and were enrolled into the study.

Acknowledgements

Katja Ickstadt acknowledges the support of BMBF and MKW.NRW within the Lamarr-Institute for Machine Learning and Artificial Intelligence. We are grateful for the willingness of the patients to participate in the GCKD study. The enormous effort of the study personnel of the various regional centers is highly appreciated. We thank the large number of nephrologists who provide routine care for the patients and collaborate with the GCKD study.

Availability of data and materials

The program codes and data sets for the simulation study and illustrations in section 3 are available on zenodo with the id 15280709 (DOI:10.5281/zenodo.15280709).

Competing interests

Berit Hunsdieck, Christian Bender and Johanna Mielke are employees of Bayer AG.

References

[1] Loh, W.-Y., Cao, L., Zhou, P.: Subgroup identification for precision medicine: A comparative review of 13 methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**(5), 1326 (2019)

[2] Wang, R.C., Wang, Z.: Precision medicine: disease subtyping and tailored treatment. Cancers **15**(15), 3837 (2023)

[3] Fiero, M.H., Huang, S., Oren, E., Bell, M.L.: Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials **17**, 1–10 (2016)

[4] Bennett, D.A.: How can i deal with missing data in my study? Australian and New Zealand journal of public health **25**(5), 464–469 (2001)

[5] Do, K.T., Wahl, S., Raffler, J., Molnos, S., Laimighofer, M., Adamski, J., Suhre, K., Strauch, K., Peters, A., Gieger, C., *et al.*: Characterization of missing values in untargeted ms-based metabolomics data and evaluation of missing data handling strategies. Metabolomics **14**, 1–18 (2018)

[6] Redestig, H., Kobayashi, M., Saito, K., Kusano, M.: Exploring matrix effects and quantification performance in metabolomics experiments using artificial biological gradients. Analytical chemistry **83**(14), 5645–5651 (2011)

[7] Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., Ni, Y.: Missing value imputation approach for mass spectrometry-based metabolomics data. Scientific reports **8**(1), 663 (2018)

[8] Lin, W.-C., Tsai, C.-F.: Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review **53**, 1487–1509 (2020)

[9] Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics, 857–871 (1971)

[10] Bektas, A., Schumann, R.: How to optimize gower distance weights for the k-medoids clustering algorithm to obtain mobility profiles of the swiss population. In: 2019 6th Swiss Conference on Data Science (SDS), pp. 51–56 (2019). IEEE

[11] Eckardt, K.-U., Bärthlein, B., Baid-Agrawal, S., Beck, A., Busch, M., Eitner, F., Ekici, A.B., Floege, J., Gefeller, O., Haller, H., *et al.*: The german chronic kidney disease (gckd) study: design and methods. Nephrology Dialysis Transplantation **27**(4), 1454–1460 (2012)

[12] Ahmed, M., Seraj, R., Islam, S.M.S.: The k-means algorithm: A comprehensive survey and performance evaluation. Electronics **9**(8), 1295 (2020)

[13] Batista, G.E., Monard, M.C., *et al.*: A study of k-nearest neighbour as an imputation method. His **87**(251-260), 48 (2002)

[14] van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of Statistical Software **45**(3), 1–67 (2011) https://doi.org/10.18637/jss.v045.i03

[15] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: Cluster: Cluster Analysis Basics and Extensions. (2023). R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source). https://CRAN.R-project.org/package=cluster

[16] Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**, 193–218 (1985)

[17] Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007)

[18] Cover, T.M.: Elements of Information Theory. (1999)

[19] Kaufman, L.: Partitioning around medoids (program pam). Finding groups in data **344**, 68–125 (1990)

[20] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2024). R Foundation for

Statistical Computing. https://www.R-project.org/

[21] Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D.: Dplyr: A Grammar of Data Manipulation. (2023). R package version 1.1.4. https://CRAN.R-project.org/package=dplyr

[22] Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis. (2016). https://ggplot2.tidyverse.org

[23] Pedersen, T.L.: Patchwork: The Composer of Plots. (2024). R package version 1.3.0. https://CRAN.R-project.org/package=patchwork

[24] Wickham, H., Henry, L.: purrr: Functional Programming Tools. (2023). R package version 1.0.2. https://CRAN.R-project.org/package=purrr

[25] Wickham, H., Vaughan, D., Girlich, M.: Tidyr: Tidy Messy Data. (2024). R package version 1.3.1. https://CRAN.R-project.org/package=tidyr

[26] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software **61**(6), 1–36 (2014)

[27] Kassambara, A., Mundt, F.: Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2020). R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

[28] Scrucca, L., Fraley, C., Murphy, T.B., Raftery, A.E.: Model-Based Clustering, Classification, and Density Estimation Using mclust In R. (2023). https://doi.org/10.1201/9781003277965 . https://mclust-org.github.io/book/

[29] Titze, S., Schmid, M., Köttgen, A., Busch, M., Floege, J., Wanner, C., Kronenberg, F., Eckardt, K.-U., Investigators, G.S., Eckardt, K.-U., *et al.*: Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the german chronic kidney disease (gckd) cohort. Nephrology Dialysis Transplantation **30**(3), 441–451 (2015)

[30] Steinbrenner, I., Schultheiss, U.T., Kotsis, F., Schlosser, P., Stockmann, H., Mohney, R.P., Schmid, M., Oefner, P.J., Eckardt, K.-U., Köttgen, A., *et al.*: Urine metabolite levels, adverse kidney outcomes, and mortality in ckd patients: a metabolome-wide association study. American Journal of Kidney Diseases **78**(5), 669–677 (2021)

[31] Therneau, T.M., Grambsch, P.: The Cox Model. (2000)

[32] Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association **53**(282), 457–481 (1958)

[33] Amdur, R.L., Feldman, H.I., Gupta, J., Yang, W., Kanetsky, P., Shlipak, M.,

Variable	Description
N	Number of observations
M	Number of Variables
L	Number of true clusters
K	Number of derived clusters
$v_m(n)$	Variable m, , $m \in \{1, \ldots, M\}$, with $v_m(n) = (v_m^{(val)}(n), v_m^{(mis)}(n))$, for observation $n, n \in \{1, \ldots, N\}$
$v_m^{(val)}(n)$	Value information of variable m, $v_m^{(val)}(n) \in \mathbb{R}^N$ for observation $n, n \in \{1, \ldots, N\}$
$v_m^{(mis)}(n)$	Missingness information of variable m, $v_m^{(val)}(n) \in \{0, 1\}^N$, for observation $n, n \in \{1, \ldots, N\}$
$C(n)$	True cluster of observation $n, n \in \{1, \ldots, N\}$
$\hat{C}(n)$	Derived cluster of observation $n, n \in \{1, \ldots, N\}$
n_k	Number of samples in derived cluster k
n^l	Number of samples in original cluster l
n_k^l	Number of samples in cluster k belonging to cluster l

Table 1 Annotations of variables utilised in method development

Rahman, M., Lash, J.P., Townsend, R.R., Ojo, A., *et al.*: Inflammation and progression of ckd: the cric study. Clinical journal of the American Society of Nephrology **11**(9), 1546–1556 (2016)

[34] Abraham, G., Sundaram, V., Sundaram, V., Mathew, M., Leslie, N., Sathiah, V.: C-reactive protein, a valuable predictive marker in chronic kidney disease. Saudi Journal of Kidney Diseases and Transplantation **20**(5), 811–815 (2009)

[35] Afsar, B.: Pathophysiology of copeptin in kidney disease and hypertension. Clinical hypertension **23**, 1–8 (2017)

[36] Bellos, I., Marinaki, S., Lagiou, P., Benetou, V.: Galectin-3 in chronic kidney disease. Clinica Chimica Acta, 119727 (2024)

[37] Vanholder, R., Gryp, T., Glorieux, G.: Urea and chronic kidney disease: the comeback of the century?(in uraemia research). Nephrology Dialysis Transplantation **33**(1), 4–12 (2018)

[38] Glassock, R.J., Winearls, C.: Screening for ckd with egfr: doubts and dangers. Clinical Journal of the American Society of Nephrology **3**(5), 1563–1568 (2008)

[39] White, I.R., Carlin, J.B.: Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine **29**(28), 2920–2931 (2010)

5 Appendix

S1 Evaluation via external validation measures

Purity

The purity measures the ability of a clustering method to recover known classes, namely how "pure" the derived clusters are. It is defined by

$$Purity = \frac{1}{N} \sum_{q=1}^{K} max_{1 \leq j \leq L}(n_q^j) \quad \in [0,1] \tag{10}$$

with N total number of samples, n_q^j number of samples in cluster q belonging to original class j $(1 \leq j \leq L)$. The higher the purity, the better the clustering performance.

Entropy

Entropy is a measure that reflects the mean homogeneity within the clusters. It is defined by

$$Entropy = -\frac{1}{N \cdot log_2(l)} \sum_{q=1}^{K} \sum_{j=1}^{L} n_q^j \cdot log_2 \left(\frac{n_q^j}{n_q} \right) \tag{11}$$

with N total number of samples, n_q number of samples in cluster q, n_q^j number of samples in cluster q belonging to original class j $(1 \leq j \leq L)$. The smaller the entropy, the better the clustering performance.

Mutual Information

Mutual Information I is defined as relative entropy between the joint distribution of the true and derived clusters and the product distribution $p(x)p(y)$,

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \cdot log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{12}$$

$\hat{C}$ \C	C_1	C_2	$\cdots$	C_L	sum
$\hat{C}_1$	n_1^1	n_1^2	$\cdots$	n_1^l	n_1
$\hat{C}_2$	n_1^2	n_2^2	$\cdots$	n_2^l	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\hat{C}_K$	n_k^1	n_k^2	$\cdots$	n_K^L	n_K
sum	n^1	n^2	$\cdots$	n^L	

Table 2 Scheme: Contingency Table given two clusterings $\hat{C} = \{\hat{C}_1, \ldots, \hat{C}_K\}$ and $C = \{C_1, \ldots, C_L\}$

given the probability distributions $p(x)$ of the true clusters and $p(y)$ of the derived clusters. It computes the entropy of the empirical probability distribution. The higher the mutual information, the better the clustering performance.

Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) measures the degree of agreement between two sets of cluster assignments: the denominator is generated by a statistical method independent of the label groups, and the numerator reflects the true classification. Referring to the Contingency Table 2, the adjusted rand index (ARI) is defined by

$$ARI = \frac{\sum_{k,l}\binom{n_k^j}{2} - \left[\sum_k \binom{n_k}{2}\sum_l \binom{n^l}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_k \binom{n_k}{2} + \sum_l \binom{n^l}{2}\right] - \left[\sum_k \binom{n_k}{2}\sum_l \binom{n^l}{2}\right]/\binom{n}{2}} \quad \in [0,1]. \tag{13}$$

An increased ARI indicates superior clustering outcomes. It is calculated using the R package *mclust* [28].

	μ_{val}	p_m	Cluster Size
Cluster A	96.55	0.48	500
Cluster B	104.68	0.24	300
Cluster C	99.68	0.23	300
Cluster D	102.77	0.65	400
Cluster E	99.08	0.24	400

Table 3 Characteristics used for simulation of five clusters, given by mean value μ_{val}, average missingness proportion p_m and cluster size.

	Scenario 1		**Scenario 2**		**Scenario 3**		
	μ_{val}	p_m	μ_{val}	p_m	μ_{val}	p_m	Cluster Size
Cluster A	90	0.3	90	0.5	90	0.5	500
Cluster B	95	0.3	95	0.3	100	0.5	300
Cluster C	100	0.3	100	0.5	90	0.1	300
Cluster D	105	0.3	105	0.3	100	0.1	400
Cluster E	110	0.3	110	0.5	110	0.5	400
	Scenario 4		**Scenario 5**		**Scenario 6**		
	μ_{val}	p_m	μ_{val}	p_m	μ_{val}	p_m	Cluster Size
Cluster A	95	0.4	97.5	0.4	100	0.1	500
Cluster B	95	0.2	97.5	0.2	100	0.2	300
Cluster C	100	0.5	100	0.5	100	0.3	300
Cluster D	100	0.1	100	0.1	100	0.4	400
Cluster E	105	0.4	102.5	0.4	100	0.5	400

Table 4 Characteristics used for six different scenarios of information (value and missingness): Simulation of five clusters, given by mean value μ_{val}, average missingness proportion p_m and cluster size.

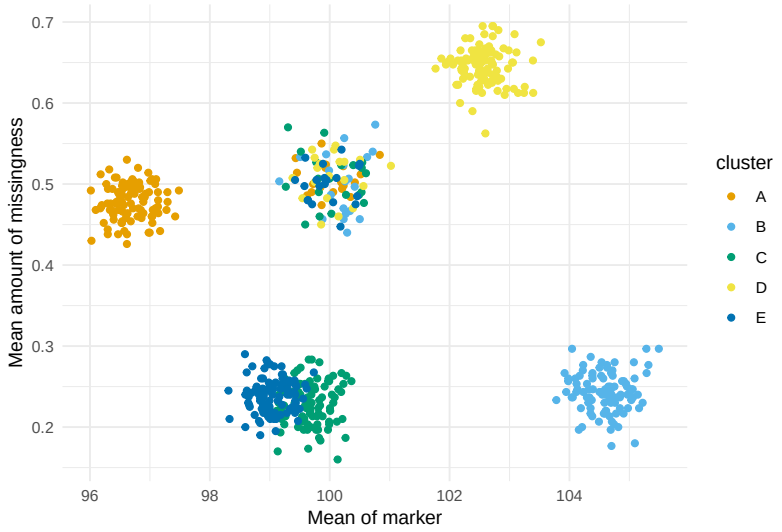

Fig. 9 Distribution of the marker means per cluster: Dots are given as the marker means $(\mu_{val}^{(m)}, \mu_{mis}^{(m)})$, of marker m, $m = 1, \ldots, 120$, of the corresponding cluster.

Value-Cluster	Missingness-Subcluster	Cluster size
C_1	$C_{1,1}$	312
	$C_{1,2}$	195
C_2	$C_{2,1}$	150
	$C_{2,2}$	67
	$C_{2,3}$	43
	$C_{2,4}$	28
	$C_{2,5}$	18
C_3	$C_{3,1}$	1401
	$C_{3,2}$	645
	$C_{3,3}$	1430

Table 5 Distribution of cluster sizes, $\hat{C}_{x,y}$. Here, x indicates the value-based cluster, while y denotes the subcluster determined by the missingness within the value cluster.

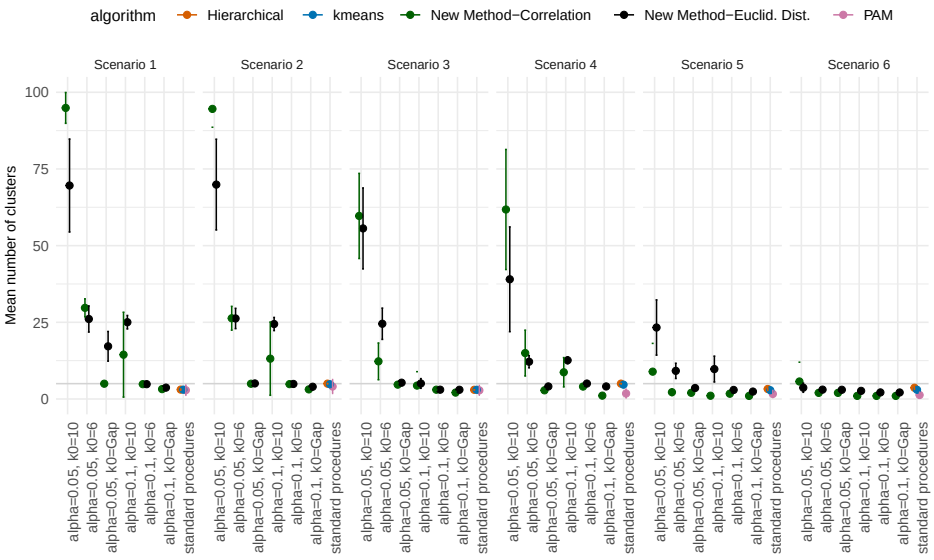

Fig. 10 Final number n of derived clusters for different clustering algorithms. Dots are representing the mean number of derived clusters. Error bars are given by $\hat{\mu}_n \pm \hat{\sigma}_n$. Grey horizontal line is marking the true number of clusters.

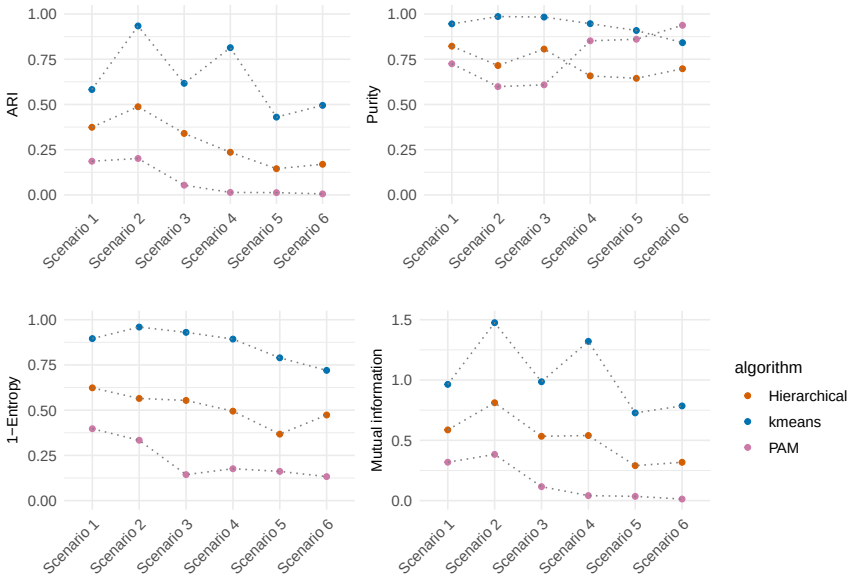

Fig. 11 Comparison of four different performance measures (ARI, purity, entropy, mutual information) for k-means, PAM and hierarchical clustering, evaluated for six different scenarios; In scenario 1 the information is solely given by the values of the markers. In scenario 6 the information is solely given by the missingness information of the markers. Scenario 2-5 reflect different graduations of missingness and value information. The big dots illustrating the empirical means $\hat{\mu}_{\{\cdot\}}$ error bars are given by $\hat{\mu}_{\{\cdot\}} \pm \hat{\sigma}_{\{\cdot\}}$.

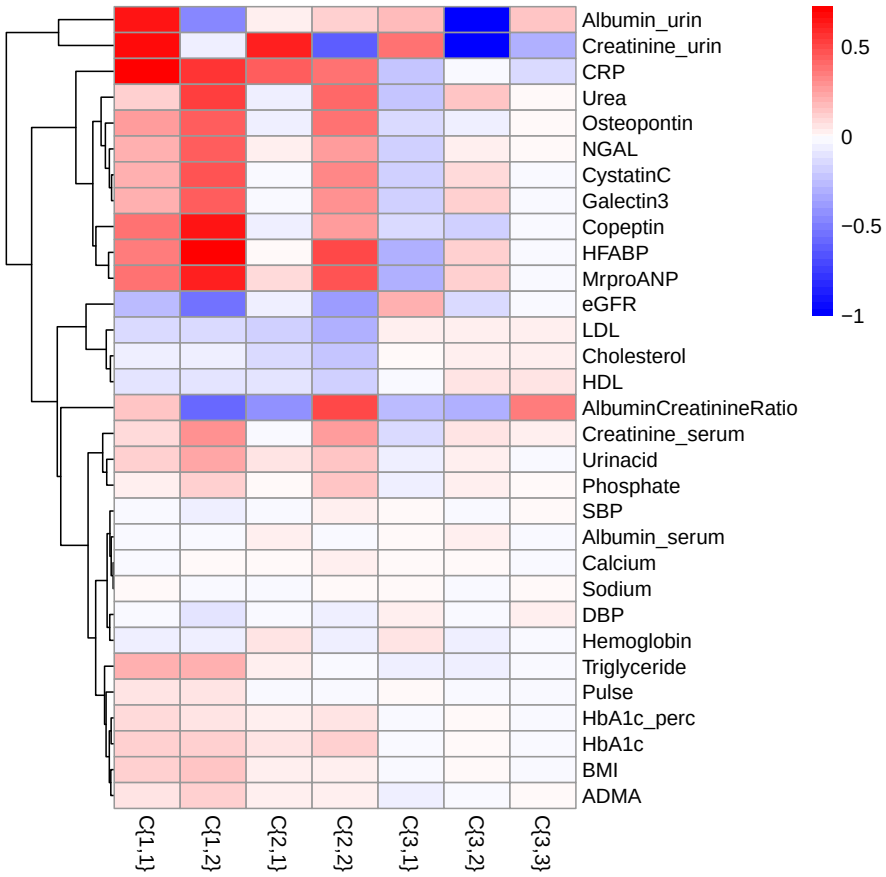

Fig. 12 Heatmap illustrating the log2 fold change of biomarkers across each cluster. The color intensity represents the magnitude of change, with blue indicating upregulation and red indicating downregulation of biomarkers relative to the full cohort. $C\{x, y\}$ denotes cluster $\hat{C}_{x,y}$.

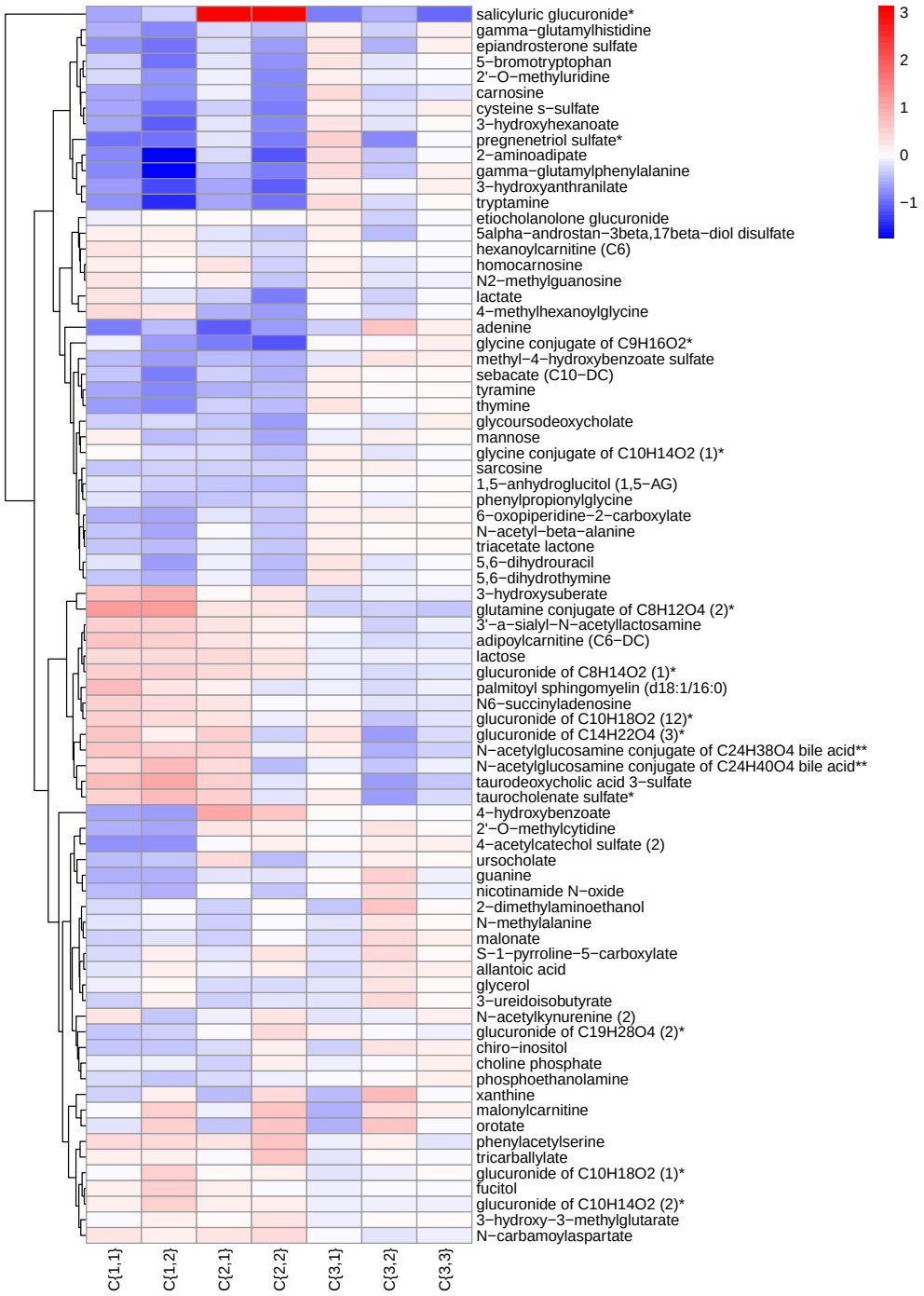

Fig. 13 Heatmap illustrating the log2 fold change of metabolite values across each cluster. The color intensity represents the magnitude of change, with blue indicating upregulation and red indicating downregulation of metabolites relative to the full cohort. $C\{x, y\}$ denotes cluster $\hat{C}_{x,y}$.

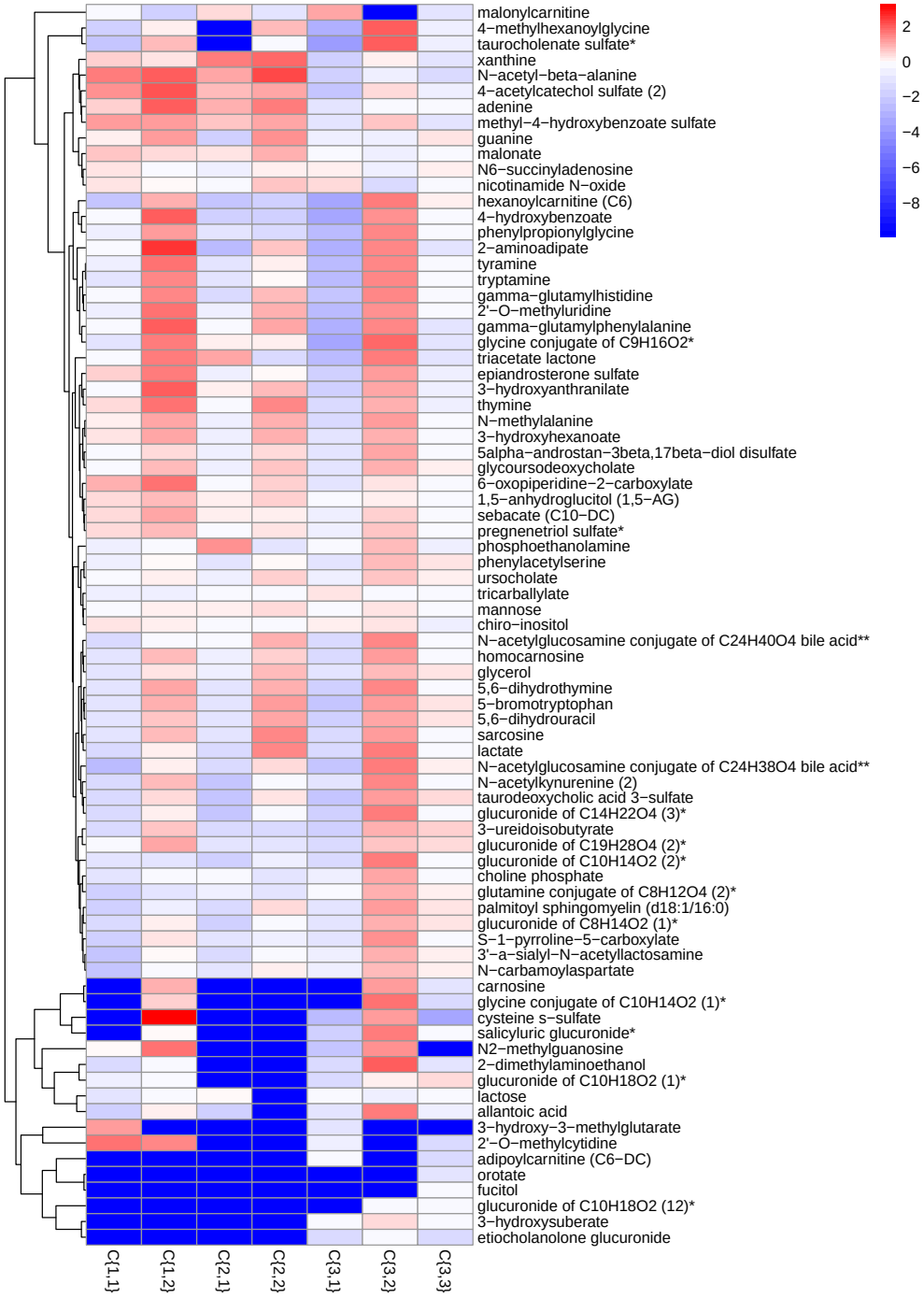

Fig. 14 Heatmap illustrating the log2-fold change of missingness information of metabolites across each cluster. The color intensity represents the magnitude of change, with blue indicating upregulation and red indicating downregulation of metabolites relative to the full cohort. $C\{x, y\}$ denotes cluster $\hat{C}_{x,y}$.

Joint Models in Big Data: Simulation-Based Guidelines for Required Data Quality in Longitudinal Electronic Health Records

Berit Hunsdieck
berit.hunsdieck@bayer.com
Bayer (Germany)
Christian Bender
Bayer (Germany)
Katja Ickstadt
TU Dortmund University
Johanna Mielke
Bayer (Germany)

Method Article

Keywords: Joint Modelling, Longitudinal Data Application, Primary Care Data, Simulation Study, Chronic Kidney Disease

Posted Date: March 27th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6031358/v1

Additional Declarations: Competing interest reported. Berit Hunsdieck, Christian Bender and Johanna Mielke are employees of Bayer AG.

Joint Models in Big Data: Simulation-Based Guidelines for Required Data Quality in Longitudinal Electronic Health Records

Berit Hunsdieck[1,2*], Christian Bender[1†], Katja Ickstadt[2,3†], Johanna Mielke[1†]

[1*]Computational Biology, Bayer AG, Wuppertal, Germany.
[2]Department of Statistics, TU Dortmund University, Dortmund,Germany.
[3]Lamarr-Institute for Machine Learning and Artificial Intelligence, Dortmund,Germany.

*Corresponding author(s). E-mail(s): berit.hunsdieck@bayer.com;
Contributing authors: christian.bender@bayer.com;
ickstadt@statistik.tu-dortmund.de; johanna.mielke@bayer.com;
†These authors contributed equally to this work.

Abstract

Background: Over the past decade an increase in usage of electronic health data (EHR) by office-based physicians and hospitals has been reported. However, these data types come with challenge regarding completeness and data quality and it is, especially for more complex models, unclear how these characteristics influence the performance.

Methods: In this paper, we focus on joint models which combines longitudinal modelling with survival modelling to incorporate all available information. The aim of this paper is to establish simulation-based guidelines for the necessary quality of longitudinal EHR data so that joint models perform better than cox models. We conducted an extensive simulation study by systematically and transparently varying different characteristics of data quality, e.g., measurement frequency, noise, and heterogeneity between patients. We apply the joint models and evaluate their performance relative to traditional Cox survival modelling techniques.

Results: Key findings suggest that biomarker changes before disease onset must be consistent within similar patient groups. With increasing noise and a higher measurement density, the joint model surpasses the traditional Cox regression

model in terms of model performance. We illustrate the usefulness and limitations of the guidelines with two real-world examples, namely the influence of serum bilirubin on primary biliary liver cirrhosis and the influence of the estimated glomerular filtration rate on chronic kidney disease.

Keywords: Joint Modelling, Longitudinal Data Application, Primary Care Data, Simulation Study, Chronic Kidney Disease

1 Background

Identifying patients at high risk for clinical diagnosis as early as possible is increasingly important [1, 2]. Statistical models can be used for generating this early risk prediction of various health conditions. In this paper, we focus on the task of identifying patients at high risk for disease based on (longitudinal) biomarker levels. This builds on the hypothesis that small changes in biomarker levels can indicate changes in health, which ultimately lead to diagnosis of a disease at a later time point [3].

Joint models, which combine longitudinal and survival data in a unified framework, present a compelling approach to take advantage of all available longitudinal information within a single model [4]. Research has shown that these models can enhance our understanding and yield improved parameter estimates compared to static survival data [5]. To implement joint models, it is essential to have a dataset that encompasses both types of data (survival and longitudinal data) of reasonably high quality.

Electronic health record (EHR) data, which typically includes longitudinal primary care and hospital information, offer a rich repository of patient health details, including laboratory results, diagnostic tests, treatments, symptoms and results [6]. However, working with EHR data presents numerous challenges, particularly in the pre-processing phase and in the analysis of processed data. The primary quality issues associated with the EHR data include incompleteness, inconsistency, and inaccuracy [7]. In the context of primary care data, specific challenges arise, such as missing data, noise, irregular data patterns, and the difficulty in accurately identifying relevant data points.

So far, it has not been systematically assessed how joint models perform in noisy real-world data as expected in EHR data sets, that is, what level of data quality is required so that joint models still offer an advantage in terms of precision and bias compared to other more established approaches such as Cox regression [8].

The primary objective of this paper is to establish guidelines for the necessary quality of longitudinal data for joint models through simulations. More concretely, this paper conducts an extensive simulation study by systematically and transparently varying different characteristics of longitudinal data, including measurement frequency, noise, and heterogeneity between patients. Utilizing the simulated data, we apply the joint models and evaluate their performance compared to traditional Cox survival modelling techniques. Insights gained in this simulation study are summarised in guidelines for data quality that other researchers can apply when making the decision whether a joint model or another model should be fitted.

We illustrate the usefulness of the guidelines with two practical examples to evaluate whether long-term biomarker records are of sufficient quality to extract insights from these trajectories. In particular, we examine how Bilirubin impacts primary biliary cirrhosis and how the estimated glomerular filtration rate (eGFR) affects chronic kidney disease (CKD).

2 Methods

2.1 Framework for Simulating Longitudinal Primary Care Data

We present a simulation framework to generate realistic primary care and hospital data for joint modelling that will be used to investigate the impact of various characteristics of data quality on disease progression prediction models.

For our simulation study, we assume that patients enter the study at time t_{start} without prior diagnosis of the relevant disease. The patients are then followed over a 5-year observation period in which both survival data and longitudinal data are collected, ending at the time point t_{end}. Subsequently, starting after t_{end}, a 5-year follow-up period is considered, where no longitudinal data is observed and only survival data are recorded (see Figure 1). Time is measured in months. We assume that the longitudinal data (the EHR or biomarker data) are scaled to lie within an interval of $[0, 1]$, which can be achieved by min-max normalisation. We also assume a balanced design with a 50-50 split between healthy and diseased patients.

We generate the following types of data in our simulation:

1. Survival outcome, i.e., information if a patient is diagnosed or not during follow-up period
2. Longitudinal EHR data, e.g. measurements of biomarkers (from t_{start} to t_{end})
3. Fixed baseline characteristics of patients, such as sex and age (see Appendix, Simulation of the Baseline characteristics), that affect survival outcomes and are chosen differently between cases and controls so that they indirectly influence the survival outcomes.

In the following, we define P_i as a patient i, $i = 1, \ldots, N$. For simplicity, we consider univariate longitudinal data, i.e., the availability of a single measured biomarker. Details of the data generation process are described in the following sections.

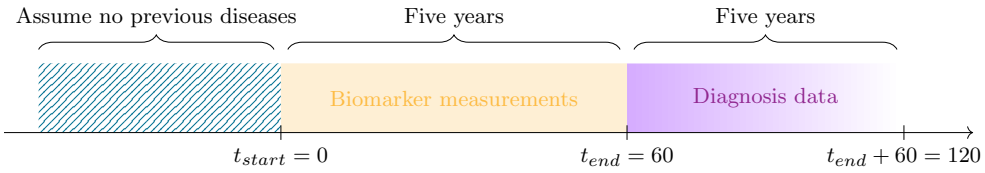

Fig. 1 Scheme: Designated observation periods given in months for simulated data

2.1.1 Survival Outcome

The simulation of the diagnostic time point (time point of event), denoted as $t_{i,abs}$, was carried out within a specified range of 10 to 119 months. This means that we require *some* longitudinal prior (i.e., between month 0 and month 10) to the event and that if a patient is to be diagnosed with the disease, this event would occur no later than five years after the biomarker observation period. For healthy patients, the diagnosis time point was right-censored at 120 months, indicating that no diagnosis was made within the observation period. In contrast, for patients who developed the disease, the diagnosis time point was randomly assigned based on a uniform distribution between 10 and 119 months,

$$t_{i,abs} \sim \begin{cases} 120 & \text{, if patient } i \text{ is healthy patient (censored after 120 months)} \\ U(10, 120) & \text{, if patient } i \text{ becomes sick,} \end{cases} \tag{1}$$

where U represents the uniform distribution. This shows that there is no specific increase or decrease in risk during the time window in which the GP data are modelled.

2.1.2 Longitudinal Data

With the increasing availability of biobanks and EHR data, these resources can be utilized to enhance the development of predictive models. Especially longitudinal EHR data come with irregularities such as varying numbers of measurements and differing precisions of these measurements. To understand the impact of these distinct characteristics on predictive models, it is essential to simulate such data in a realistic and verifiable manner.
In EHR data, biomarker measurements are typically not taken at prespecified time points, but varying between patients (that is, taken on the decision of the physician). That is why, for the longitudinal data, we need to both simulate the data frequency and the actual data time and the values.

Number of Measurements

The number of measurements for each patient, denoted as n_i, is simulated using the absolute value of a normal distribution realisation with the mean number of measurements n_{abs} per year and the standard deviation 2. This distribution is chosen to reflect the variability in the frequency of measurements observed in real-world EHR data and can help to analyse the influence of the number of measurement frequency on prediction. Figure 2 (in the Appendix) shows the mean number of measurements and the standard deviation for 66 biomarkers in the UK Biobank data (see Appendix, UK Biobank), focussing on data up to five years before the baseline visit. Based on this, a mean standard deviation of 2 and 1 measurement every two years on average is most likely for a typical biomarker. As we are interested in examining longitudinal data, the UK Biobank dataset can be tailored to include only those patients with a minimum of two to three measurements. This filtering increases the average number

of measurements per year but leads to a reduced sample size and potential biases. The number of measurements n_i are simulated such that

$$n_i \sim \lfloor \mathcal{N}(n_{abs} \cdot \frac{t_{i,abs}}{12}, 2) \rfloor \tag{2}$$

The diagnostic time point $t_{i,abs}$ is included in the expectation of the number of measurements because the total count of measurements depends on the duration up to the diagnosis, that is, if a longer time until diagnosis is observed, more longitudinal data can be measured.

Simulation of the Distribution of Measurement Dates

The measurement time points for the patient i, denoted as $t_{i,j}$, are simulated as random integers $\{t_{i,1}, \ldots, t_{i,n_i}\}$ with

$$t_{i,j} \sim U(0, t_{i,abs}), \quad j \in \{1, \ldots, n_i\}, \tag{3}$$

where U reflects the uniform distribution. The chosen parameter reflects that longitudinal markers are observed for 60 months (from t_{start} to t_{end}). The measurement dates for each patient were generated independently, resulting in a set of random time points for the patient.

Simulation of Measurement Values

The final goal of this simulation is to generate longitudinal EHR data, denoted as $y(t_{ij})$, prior to the disease onset. For simplicity, we focus on continuous measurements here. We adopt a linear mixed-effects model to represent the underlying trend in longitudinal data preceding the onset of the disease. Linear mixed-effects models, as introduced by [9], are widely recognised for their efficacy in modelling longitudinal trajectories (e.g., see [10]).

More concretely, we assume

$$y_{ij} = y(t_{ij}) = b_i + m_i \cdot t_{ij} \cdot \mathbb{1}_{t_{ij} \leq t_{i,abs} - 12 \cdot t_m} + \epsilon_{i,j} \tag{4}$$

for patient i, where b_i reflects a patient-specific intercept and m_i reflects a patient-specific slope that is added to the data starting from a breakpoint t_m. The error term $\epsilon_{i,j}$ is time-point specific. This means that the expected value of the longitudinal data is assumed to be constant until the break point t_m, where the biomarker will start to change and already give an indication of the later diagnosis, and then will increase linearly thereafter.

The slope parameter m_i is simulated in a two-step procedure: For patients who are in the process of developing the disease, the probability of showing an effect before the onset of the disease based on longitudinal data (responding) is modelled using a Bernoulli distribution with a probability of p_{resp}, that is,

$$p_i \sim \begin{cases} 0 & \text{, if patient } i \text{ is healthy} \\ Bern(p_{resp}) & \text{, if patient } i \text{ is sick} \end{cases} .$$

For healthy patients, we assume that the slope is 0 in all cases. Then, the slope m_i is calculated based on

$$m_i = p_i \cdot m_i^*,$$

with

$$m_i^* \sim \mathcal{N}(\mu_m, \sigma_m^2),$$

where μ_m and σ_m represent the mean value and standard deviation, respectively. The values for μ_m and σ_m are discussed in Section 3. This approach takes into account the expected heterogeneity between patients, as not all patients are expected to show an association between longitudinal and survival data.

For the intercept, we assume that patients developing the disease already exhibit different baseline levels and define the difference as Δ_b (w.l.o.g.: $\Delta_b \in [0, 0.5]$). Therefore, we simulate the intercept b_i with a normal distribution with mean 0.5 and standard deviation σ_b^2, i.e.,

$$b_i \sim \begin{cases} \mathcal{N}(0.5 + \Delta_b, \sigma_b^2) & \text{, if patient } i \text{ is sick} \\ \mathcal{N}(0.5, \sigma_b^2) & \text{, if patient } i \text{ is healthy.} \end{cases} \tag{5}$$

As measurements are consistently affected by noise, additional time-independent noise is included, represented by

$$\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_\epsilon^2) \quad . \tag{6}$$

2.1.3 Parameter Choices and Practical Example

This simulation approach allows an individual generation of realistic longitudinal data that reflects the different trajectories of healthy and diseased patients prior to the disease, allowing for adjusting different data quality parameters.
Since we want to examine how individual data quality parameters influence the performance of models, we vary the corresponding parameters. The parameter choices given in Table 1 are selected to explore various aspects of patient response and measurement variability with respect to their influence on model performance. For example, the number of measurements is selected to align with their distribution within the UK Biobank (see Figure 2 in the Appendix).
To keep the number of settings under evaluation in a manageable range, we use a reference setting per parameter (in bold in Table 1) and vary one parameter at a time. Parameter selections are designed to be based on Min-Max normalised values so that the values are within a $[0, 1]$ range, facilitating transferability to real-world data settings. Since the range of the standard deviation depends on the range of

the mean, only one parameter is varied for both the slope m and the intercept b (only one scenario for the other parameters). An example of a simulated patient

Parameter	Parameter Annotation	Parameter Choices
Sample Size	N	$\{50, 200, \mathbf{500}, 5000\}$
Noise Standard Deviation	σ_ϵ	$\{0.05, 0.075, \mathbf{0.15}, 0.3\}$
Percentage of Patients Responding	p_{perc}	$\{0, 0.2, 0.5, 0.8, \mathbf{1}\}$
Years of Assumed Slope	t_m	$\{1, \mathbf{3}, 5\}$
Number of Measurements per Year	n_{abs}	$\{1, \mathbf{2}, 3\}$
Intercept Difference	Δ_b	$0, \{\mathbf{0.1}, 0.2\}$
Intercept Standard Deviation	σ_b	$\{\mathbf{0.05}\}$
Slope Mean	μ_m	$\{\mathbf{0.005}\}$
Slope Standard Deviation	σ_m	$\{0.001, \mathbf{0.005}, 0.01\}$

Table 1 Parameter selections for further simulation and investigation of the impact of data quality metrics on risk prediction (In bold: Reference values). Since the range of the standard deviation depends on the range of the mean, only one parameter is varied for both the slope m and the intercept b.

trajectory for a patient diagnosed after 75 months is given in Figure 2. It is evident that approximately five years before diagnosis, the values start to change exhibiting an increasing trend over time until the diagnostic time point at $t = 75$. During the observation period $[0, 60]$, a total of 12 measurements are available.

2.2 Theoretical Foundations of Joint Models and Time-varying Evaluation Metrics

To systematically investigate the effects of different characteristics of quality and quantity of longitudinal EHR data on the predictive power of a joint modelling approach (see Section 2.2.1), we compare the prediction performance to a standard

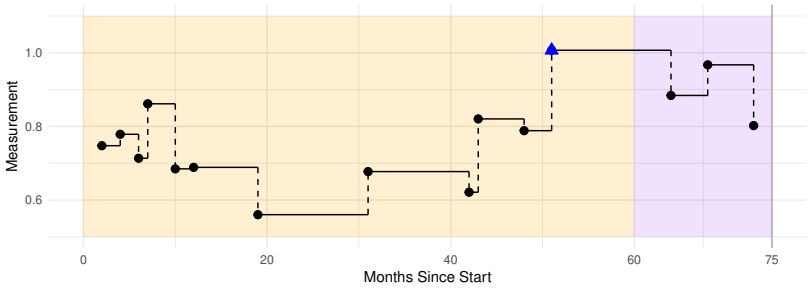

Fig. 2 Example of a simulated trajectory of a sick patient: The Patient is getting sick after 75 months with twelve measurements during observation period $[0, 60]$. The patient is male, not smoking and 66 years old. The blue triangle marks the last (simulated) measurements within the observation period.

Cox model. The approaches are evaluated by a version of the time-varying concordance index (see 2.2.3). This allows for a comprehensive analysis of the influence of longitudinal data of varying quality on the prediction of the risk of disease progression.

2.2.1 Joint Model

Joint modelling of longitudinal and time-to-event processes enhances the precision of the estimation and predictive performance by effectively capturing the intrinsic relationships between the submodels. This approach is particularly advantageous in longitudinal studies, as it characterises the association between a longitudinal response process and a time-to-event outcome [4]. Consequently, these models have become increasingly popular in recent years and, as a widely utilized class of models, will serve as the foundation for developing the guidelines.
The joint model can be split into two submodels, the longitudinal model and the survival model. For the endogenous time-dependent longitudinal covariate (e.g., biomarker measurements), let $y_{ij}(t)$ be the observed value of the i-th subject at time point t_{ij},

$$y_{ij} = \{y_i(t_{ij}), j = 1, \ldots, n_i\} \quad .$$

Let $T_i^\star$ be the true event time for the i-th subject and T_i the observed event time with $T_i = min\{C_i, T_i^\star\}$, C_i the potential censoring time, and $\delta_i = \mathbb{1}(T_i^\star \leq C_i)$ the event indicator.

Longitudinal Model

For the longitudinal model, we assume that the longitudinal outcomes are normally distributed and follow a linear shape. Then, the mixed-effects model is given by

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= x_i^T(t)\beta + z_i^T(t)b_i + \epsilon_i(t) \end{aligned}$$

with

$$\begin{aligned} x_i(t), \beta &: \text{Fixed effects parts} \\ z_i(t), b_i &: \text{Random effects parts, } b_i \sim \mathcal{N}(0, D) \\ \epsilon_i(t) &: \text{Time-dependent error terms, } \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

with variance-covariance matrix D. The longitudinal model is implemented using the R package nlme [11] that includes the estimation of the covariance matrix of the random effects.

Survival Model

For the survival model, the relative risk is given by

$$h_i(t|\mathcal{M}_i(t), w_i) = h_0(t)exp\left\{\gamma^T w_i + \alpha m_i(t)\right\}, \quad t > 0$$

with

$\mathcal{M}_i(t)$: $\{m_i(s), 0 \leq s < t\}$
: History of true unobserved longitudinal process up to time point t
$m_i(t)$: True and unobserved value of covariate at time t (equivalent to respective part in longitudinal model)
$h_0(t)$: Baseline risk function at time t
w_i : (Vector of) baseline covariates with coefficients vector γ
γ : Vector of regression coefficients for baseline covariates
α : Effect of underlying longitudinal outcome to the risk for an event
: Quantifies association between time-varying covariate and risk of event

The survival model can be implemented using the R package survival [12].

Joint Distribution

Assuming that the two processes are associated, we can define a model for their joint distribution by assuming that we have full conditional independence, e.g., the random effects explain all interdependencies. This yields to the joint distribution

$$p(y_i, T_i, \delta_i) = \int p(y_i|b_i) \cdot \left\{h(T_i|b_i)^{\delta_i} S(T_i|b_i)\right\} p(b_i) db_i$$

with

b_i : Vector of random effects explaining interdependencies
$p(\cdot)$: Density function
$S(\cdot)$: Survival function

Taking into account the longitudinal submodel (see section Longitudinal Model) as well as the survival submodel (see section Survival Model). The models are jointly optimised with the EM algorithm through Bayesian approaches using MCMC techniques [13]. The prior distributions are defined using frequentist univariate regression models fitted separately for each outcome. The mean of the Gaussian prior is defined as the maximum likelihood estimate (MLE) and the precision is defined as the inverse of 10 times the variance of the estimate from the univariate model. Regarding the prior distributions for the variance and covariance parameters of Gaussian random effects, JMbayes2 uses gamma priors with mean defined as the MLE of univariate models [14]. It is assumed that, based on the observed history, the mechanisms of censoring and the process of visiting are independent of the actual event times and future longitudinal measurements. This implies that the decision on the withdrawal of a subject from

the study or their attendance at the clinic for a longitudinal evaluation is influenced by their past history, without additional causalities of the underlying latent characteristics of subjects that may be related to his prognosis.
In the following, the joint model will be consistently implemented and fitted using the R package *JMbayes2* [15].

2.2.2 Candidate Models for Comparison of Performance

In this section, we compare the performance of the joint model in simulated data with different characteristics (as outlined in Section 2.1) with the Cox model, as the standard approach for modelling survival data. We aim to identify settings in which the more complex joint model outperforms the Cox model. The models to be compared are given by

1. Joint model incorporating biomarker measurements from the past 5 years along with covariates such as sex, age, and smoking status (represented in green in the result section). More concretely, the submodels are given by:

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= \beta_0 + \beta_1 t + \beta_2 Sex_i + \beta_3 Age_i + \beta_4 SmokingStatus_i + b_{i0} + b_{i1} t + \epsilon_i(t) \\ h_i(t) &= h_0(t) \cdot exp\{\gamma_1 Sex_i + \gamma_2 Age_i + \gamma_3 SmokingStatus_i + \alpha m_i(t)\} \end{aligned}$$

   where the parameters $b_{i\cdot}$ mark the individual subject-specific effects.
2. Cox model including covariates such as age and sex, but no EHR data ("baseline model", shown in blue in the result section).

$$h_i(t) = h_0(t) \cdot exp\{\gamma_1 Sex_i + \gamma_2 Age_i + \gamma_3 SmokingStatus_i\}$$

3. Cox model that incorporates the covariates of age and sex along with the most recent measurement $\tilde{y}_i$ of the biomarker in the 5-year observation period (represented in orange in the result section).

$$h_i(t) = h_0(t) \cdot exp\{\gamma_1 Sex_i + \gamma_2 Age_i + \gamma_3 SmokingStatus_i + \gamma_4 \tilde{y}_i\}$$

It is important to note that the Cox model uses biomarkers measured at t_{end}, while the additional longitudinal data for the joint model is obtained prior to t_{end}. With this selection, we ensure a fair comparison between the models.
Table 1 lists the various scenarios that will be examined.

2.2.3 Evaluation of Risk Prediction Models: Adjusted Time-varying Concordance Index

The goal is to evaluate the precision and reliability of risk models. More concretely, we aim to identify with highest accuracy those who are at increased risk of developing a disease. Commonly used scores to evaluate the performance of the model are the C index to evaluate the model's ability to rank the risk of the subject and the integrated

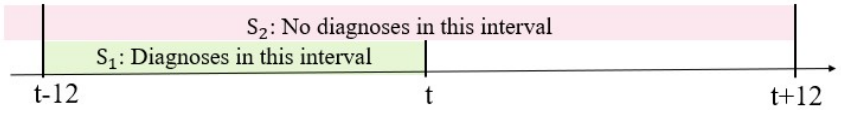

Fig. 3 Illustration of subgroups S_1 and S_2 for the time-varying C-Index definition

Brier score incorporating the discrimination and calibration aspect [16]. Forecasting risk is largely affected by the specific time period being examined (for instance, a Cox model can "*more readily*" evaluate risk profiles over a brief period like a month rather than spanning several years). For further evaluation, we introduce the time-varying C-index, derived from the conventional C-Index (see [17]). This metric can be readily understood in terms of time and adapted according to varying risks over different periods. Given the estimated individual risk $r_i(t)$ at time t of patient i, we define two subsets as illustrated in Figure 3 by

$$S_1 = \left\{\text{Individuals with diagnosis in interval } \left[t - interval, t\right]\right\}$$

and

$$S_2 = \left\{\text{Individuals without diagnosis in interval } \left[0, t + interval\right]\right\}$$

with $|S_1| = n_1$ and $|S_2| = n_2$. The time-varying C-index is then defined as

$$tvC_{interval}(t) = \frac{\sum_{k \in S_1} \sum_{l \in S_2} \mathbb{1}_{\{r_k(t) > r_l(t)\}}}{n_1 \cdot n_2} \quad \in [0, 1] \tag{7}$$

which expresses the ratio of pairs in which the predicted risk for a pre-disease patient (subset S_1) exceeds the predicted risk for a temporarily healthy patient (subset S_2), relative to all possible pairs. The closer the time-varying C-index is to 1, the better the model's performance. Compared to the established C-index, this measure allows us to observe changes in the performance over time.
For a observation period of five years ($t_{end} = 60$) and a follow-up period of five years ($\Delta t = 60$), the mean time-varying C-index of the follow-up period $[t_{end}, t_{end} + \Delta t]$ is given by

$$mean(tvC_{interval}(t)) = \frac{1}{\Delta t} \sum_{t=t_{end}}^{t_{end}+\Delta t} tvC_{interval}(t) = \frac{1}{60} \sum_{t=60}^{120} tvC_{interval}(t) \tag{8}$$

with t given in months.

2.2.4 Set-up and Evaluation of Simulation Study

Models are compared with the introduced performance measure across 100 iterations of simulated data with the parameter settings described in Table 1. From these, the mean and the 0.05 and 0.95 quantiles are computed to establish the prediction interval for one dataset. Note that in the result section, the prediction intervals are only

depicted when a change in the prediction interval is given by different parameter choices; otherwise, they are omitted for the sake of clarity.

In the following, we consider the mean time-varying C-index for patients who remain at risk by t_{end} (those not diagnosed until t_{end}) in three different models during the 5-year follow-up period. Therefore, patients with an early diagnosis, which can technically not be included in the Cox model analysis, are excluded from the evaluation to ensure fairness while comparing the joint model and the Cox model (see Section 2.2.2).

2.3 Software Versions

The following software has been used:
R Version [18]: 4.2.1
Package Versions: dplyr [19]: 1.1.4; ggplot2 [20]: 3.5.1; JMbayes2 [15]: 0.5-0; patchwork [21]: 1.2.0; purrr [22]: 1.0.2; tidyr [23]: 1.3.1; MatchIt [24]: 4.5.5

3 Results

3.1 Derivation of Simulation-Based Guidelines for Longitudinal Primary Care Data

In the following section, we examine how individual data quality parameters influence the performance of the joint model compared to the standard Cox model [8], see Section 2.2.2 for details. For that, we discuss the impact of different parameter choices on the performance given by the 5-year mean time-varying C index (short: mean C-Index) after the last possible measurement time point $t_{end} = 60$ with an interval length of 12 months. An interval of 12 months is selected to thoroughly examine the significant distinctions in risk.
Additionally, we evaluated two additional parameters, namely the response rate and the slope variance, over time (see Section 3.1.6) to demonstrate how performance differences evolve over time.

Figure 4 illustrates the performance given by the mean C index of the three different models in the scenarios described in Table 1. We first note in general that the Cox model and the joint model outperform -as expected- the Cox model without biomarkers in all scenarios. Thus, we use the result from the Cox model without biomarkers for illustration of the baseline performance, but do not discuss it in detail and focus on the two other models. For clarity of the visualisations, we only provide prediction intervals for one scenario, but note that comparable variability is also present in the other cases.

3.1.1 Sample Size

The available sample size in practice depends greatly on the specific use case. For example, without taking into account the availability of longitudinal biomarker data, the number of patients who are diagnosed with stroke (approx. 22.200 patients) is by far larger than the number of patients who are diagnosed with encephalitis (approx.

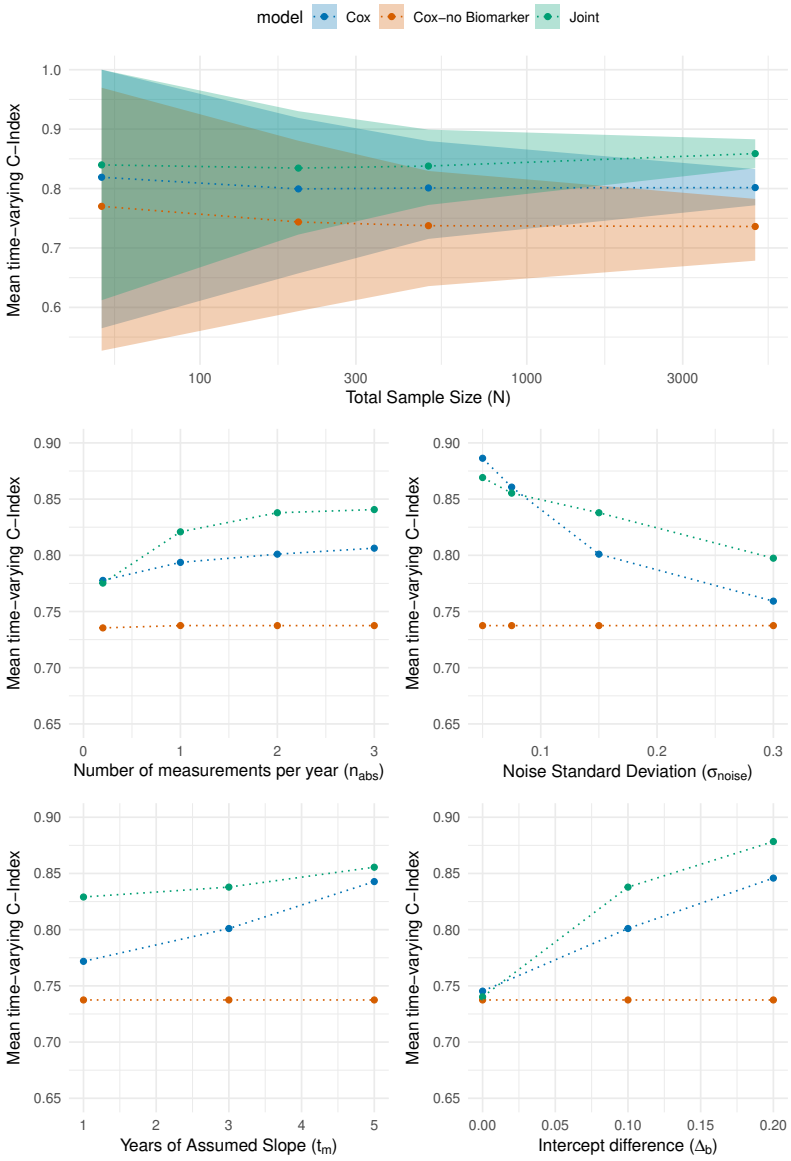

Fig. 4 Comparison of the mean time-varying C-index over five years with an interval of 12 months for different parameter choices for the total sample size (a), the number of measurements per year (b), the noise standard deviation (c), the years of assumed slope (d), and the intercept difference (e).

990 patients) within the UK Biobank. In our simulation study, we thus investigate the time-varying C-index for sample sizes between 50 and 5000 subjects. Figure 4 (a) illustrates that in small sample sizes ($N = 50$), the Cox model (orange) performs very similar to the joint model (green). However, the differences get more prominent when increasing the sample size to $N = 200$. We notice also that there are further small gains in performance for the joint model compared to the Cox model - however, these differences are rather small. Therefore, we recommend a sample size of $N \geq 200$ for a robust prediction.

We also note that the prediction intervals for the mean C-index narrow with increasing sample size.

It must be considered that the recommended sample size is influenced by other factors and can therefore vary, for example, with respect to the homogeneity of the longitudinal data within the cohort under consideration.

3.1.2 Number of Measurements (per year)

It is recognised that the number of measurements in the electronic health record (EHR) data can vary considerably depending on the specific marker and the effort required for measurement. Figure 4 (b) illustrates the performance of the models with respect to different quantities of measurements. When there is at least one measurement per year, the joint model exhibits superior performance. However, after a certain threshold, additional data points do not provide further information and thus do not enhance performance. Consequently, it can be concluded that a greater number of measurements facilitates improved longitudinal trajectory modelling, thus enhancing risk prediction. We recommend at least 1 measurement per year if a joint model is to be used.

3.1.3 Varying Noise Variance

Next, the impact of the noise of the measurements itself is analysed. External factors, such as varying conditions that affect blood pressure readings, can significantly influence measurement noise, affecting model performance. In Figure 4 (c), the importance of noise variance as a critical parameter is highlighted. It is observed that as the level of noise increases, the joint model demonstrates a greater advantage in the C-index over the Cox model. In contrast, at low noise levels (0.05), there is a minimal difference in performance between the Cox model and the joint model. The joint model is capable of filtering out higher levels of noise, yielding more accurate predictions in comparison to the Cox model. Starting at a noise variance of approx. $\sigma_e = 0.075$, the joint model starts to perform better compared to the Cox model, therefore we recommend the joint model specifically in those scenarios with at least $\sigma_e > 0.075$.

3.1.4 Years of Slope

Typically small changes in biomarker level proceeds the clinical diagnosis. However, it is not necessarily clear *how* these changes are present much earlier. As shown in Figure 4 (d), when the slope is given for a shorter time period (i.e., the biomarker starts to change only for a short period prior to the actual diagnosis), it is detected by the joint

model but not by the Cox model. In contrast, when the slope is assumed to be over a longer duration (e.g., 5 years), there is only a minimal difference in the performance between the models. This can be explained by considering that the Cox model uses only the most recent measurement. If the slope begins to increase only recently prior to the most recent measurement used, the absolute difference in biomarker levels is smaller. Since the joint model is estimating the slope, even small differences can be retrieved. It is important to note that for practical application, this value is difficult to derive.

3.1.5 Intercept/ Baseline Difference

Lastly, the impact of a baseline difference is analysed. Like covariates, a specific baseline difference in the (simulated) marker itself can occur within the risk cohort, independently of a specific time frame slope [25]. As illustrated in Figure 4 (e), in scenarios without a difference in intercept, the Cox model and the joint model perform similarly. The joint model benefits more than the Cox model from an increase in the intercept. Therefore, the joint model appears specifically suitable if the intercept difference is greater than $\Delta_b \geq 0.1$.

3.1.6 Response Rate and Slope Variance over Time

For specific parameters, not only the parameter itself, but also time plays a role in its influence on performance, e.g., the statistical model may be very good in predicting the two-year risk of diagnosis, but may fail to predict the five-year risk of diagnosis. Therefore, we look at the time-varying concordance index $tvC_{12}(t)$ over a period of time of 5 years after t_{end} for varying choices of the fraction of patients with a relationship between EHR data and survival outcome (p_{perc}) and slope variance (σ_m). Figure 5 shows the performance of the different models based on the percentage of subjects who actually demonstrate the hypothesised relationship between the biomarker and the diagnosis. We note that only if at least a majority of patients (approximately 80%) demonstrate a response, the joint model or the Cox model (with biomarker) are superior to the baseline Cox model. Therefore, a highly heterogeneous relationship between the biomarker and the diagnosis makes it difficult to detect this pattern.

In Figure 6, we compare different levels of variability of individual slopes. We note that the advantage of the Cox model is highest if the variability of the slope is small, i.e., the slope is similar across patients. However, as heterogeneity increases, the performance of the Cox model, including (bio)marker information, becomes more comparable.

When we consider the dependency of the performance on the chosen time points for the time-varying concordance index $tvC_{12}(t)$, we note that generally, for both parameters, the models are more similar for earlier time points (small values of t). This means that if the aim is to identify subjects at risk for a diagnosis at a time point that is close to the time of the last measurement, there is no strong advantage of the joint model compared to the Cox model. However, when comparisons are made at later time points, the joint model may begin to diverge from the Cox model because of its ability to predict and fit the longitudinal trajectory. This results in a more pronounced distinction between the joint and Cox curves, as illustrated in Figures 5 and 6. The

"bend" after approximately three years can be explained by the simulation of data that accounts for a biomarker change starting three years before the disease diagnosis, allowing the risk prediction to benefit from biomarker data within these three years.

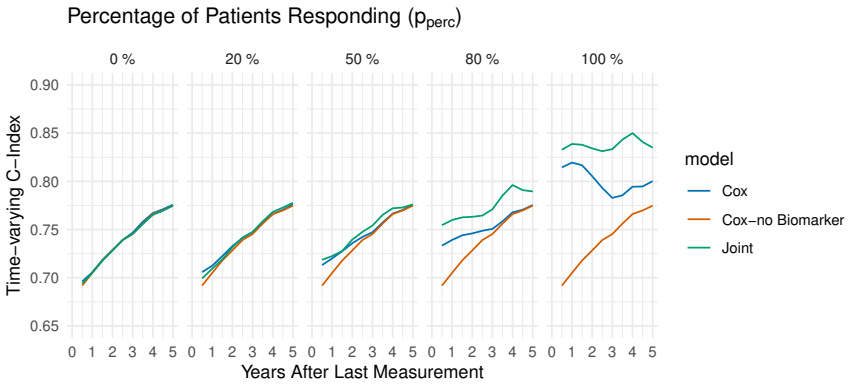

Fig. 5 Comparison of the effect of different parameter choices for the fraction of patients responding p_{resp} in simulated data on the time-varying C-index with an interval of 12 months using a joint model and a Cox model with/ without biomarker information: $p_{resp} \in \{0, 0.2, 0.5, 0.8, 1\}$

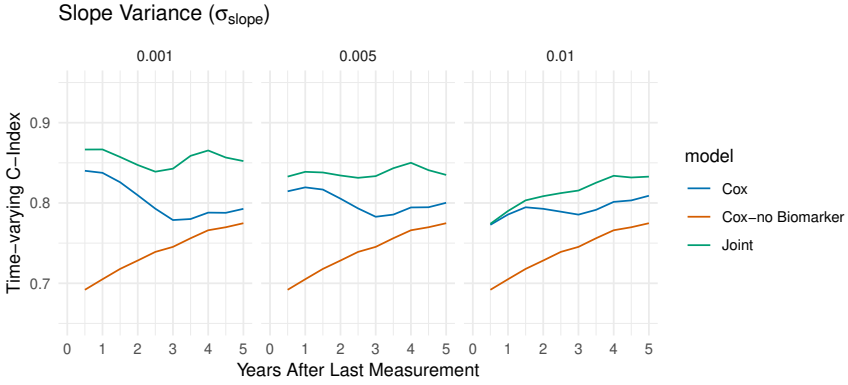

Fig. 6 Comparison of the effect of different parameter choices for the slope standard deviation σ_m in simulated data on the time-varying C-index with an interval of 12 months using a joint model and a Cox model with/ without biomarker information: $\sigma_m \in \{0.001, 0.005, 0.01\}$

3.1.7 Simulation-Based Guidelines for Longitudinal Primary Care Data

The well-established Cox model is much easier to apply and communicate than the more complex joint model. Therefore, we recommend the joint model only in scenarios in which the joint model outperforms the Cox model. Synthesising the findings of the previous sections, guidelines can be formulated that help to determine the utility of the

available longitudinal Primary Care / EHR data for the identification of disease risk. The guidelines are presented in table 2. To adhere to these guidelines, it is essential to normalise the longitudinal measurements to a range of $[0, 1]$, ensuring a mean value of approximately 0.5, primarily for scaling purposes. Almost all of the listed parameters can be directly extracted from the available real-world data of interest. It is important to note that the response rate is typically unknown. To achieve a high response rate, the cohort can be restricted to specific subgroups of patients and / or disease. In the following section, two real-world data examples are analysed using the derived guidelines.

Parameter	**Superior performance of joint model compared to Cox model**
Sample Size	$N \geq 200$
Noise Standard Deviation	$\sigma_\epsilon > 0.075$
Percentage of Patients Responding	$p_{perc} \geq 80\%$
Number of measurements per year	$n_{abs} \geq 1$
Intercept difference	$\Delta_b \geq 0.1$
Slope Standard Deviation	$\sigma_m \leq 0.005$

Table 2 Guidelines: Criteria for Normalized Electronic Health Record Data That Preferentially Support the Joint Model Over the Cox Model

3.2 Case Study: Real-World Applications of the Derived Guidelines

Table 2 presents guidelines formulated from simulated data. To demonstrate their applicability in real-world scenarios, two different datasets were subsequently examined using these guidelines.

3.2.1 Serum Bilirubin and Primary Biliary Cirrhosis (Mayo Clinic)

Primary biliary cirrhosis of the liver (PBC) is considered a progressive disease. The progression is slow and the resulting inflammation progressively results in cirrhosis, damage to the liver's bile ducts, and ultimately death of the patient [26]. More information is available in Dickson et al. [27] and Markus et al. [28]. The data used for further modelling were collected from the Mayo Clinic trial on PBC that took place from 1974 to 1984. A total of 424 patients with PBC, referred to the Mayo Clinic during that 10-year interval, met the eligibility criteria for the randomised placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomised trial and contain largely complete data [12]. The data set for the randomised trial, consisting of 312 participants, is provided by the `JMbayes2` R package [15].
The diseased subgroup comprises all patients who received an initial diagnosis of PBC within the 10-year interval. We used a 5-year period in which longitudinal data is collected, followed by a 5-year follow-up period in which only survival outcome data are recorded. Given these data, the guidelines of Table 2 are used to assess if better performance of the joint model compared to the Cox model can be expected in this scenario. Bilirubin values were initially converted using a logarithmic transformation, followed by a min-max transformation to normalize the distribution within a range of 0 to 1, targeting an average of approximately 0.5 for optimal application in guidelines. The normalised bilirubin value at time t is denoted as $bil_{norm}(t)$. The appendix outlines the process of deriving parameters using real-world data (see " Derivation of Parameter Choices Based on Real-World Data"). For applying this approach to derive slope and noise parameters, it is assumed that $t_m = 3$ years since PBC is progressing slowly. A summary of the derived parameters is given in Table 3.
All requirements given in the guidelines are fulfilled. Therefore, we expect a gain in performance for the joint model compared to the Cox model.
For the joint model, the corresponding submodels are given by

Longitudinal submodel:

$$bil_{norm}(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot Sex + \beta_3 \cdot Age + b_{i0} + b_{i1} \cdot t + \epsilon_i(t) \quad (9)$$

Survival submodel:

$$h_i(t) = h_0(t) \cdot exp\{\gamma_1 Sex_i + \gamma_2 Age_i + \alpha m_i(t)\} \quad (10)$$

with fixed and random effects parameters of the longitudinal model similar to section Longitudinal Model. The Cox models are given by

Cox model with biomarker information:

$$h_{i,Cox_{biom}}(t) = h_0(t) \cdot exp\{\gamma_1 Sex_i + \gamma_2 Age_i + bil_{norm}(\tilde{t})\} \quad (11)$$

Cox model without biomarker information:

$$h_{i,Cox}(t) = h_0(t) \cdot exp\{\gamma_1 Sex_i + \gamma_2 Age_i\} \quad (12)$$

with $\tilde{t}$ denoting the last observation time point of a Bilirubin value within the observation period.

As a result, we first note that the analysis reveals that the inclusion of biomarkers in the model provides a relevant information gain (Cox - no biomarker, illustrated in orange, substantially worse performance to the other models). Comparing the performance of the cox model to the joint model, it is confirmed, as illustrated in Figure 7, that there is a gain in prediction accuracy for the joint model compared to the cox model if the prediction interval is larger than 1 year. This is in line with our expectations since all criteria in the guidelines were met.

Parameter	Estimate	Comparison to Guidelines
Sample Size N	312	Requirement fulfilled
Noise Standard Deviation σ_ϵ	0.078	Requirement fulfilled
Percentage of Patients Responding p_{perc}	unclear	unclear
Number of measurements per year n_{abs}	$1.6(sick), 1.4(healthy)$	Requirement fulfilled
Intercept difference Δ_b	0.116	Requirement fulfilled
Intercept standard deviation σ_b	0.148 (0.124)	-
Slope Mean μ_m	0.007	-
Slope Standard Deviation σ_m	0.003	Requirement fulfilled
Years of *Assumed* Slope t_m	3	unclear

Table 3 PBC Dataset (See R package *jmbayes2* for more information): Estimation of Parameters, derived by a Mixed-Effects Model (see Appendix, Derivation of Parameter Choices Based on Real-World Data), and Comparison with guidelines given by table 2

3.2.2 eGFR and Chronic Kidney Disease (UK Biobank)

Data provided by the UK Biobank (UKB) (see Appendix, UK Biobank) provide a valuable resource for research on chronic kidney disease (CKD). Estimated glomerular filtration rate (eGFR) measurements are well established to predict the risk of CKD, and eGFR is often used for diagnostic purposes in this context [29, 30]. It is of particular interest to investigate whether eGFR trajectories provide additional information about the diagnosis of CKD. To this end, UKB data will be used to implement a joint model and analyse data quality using the derived guidelines (see table 2).

The initial step in the data preparation process involves general preprocessing, which includes joining units, converting measurements, and addressing implausible values and outliers (see Appendix, Preprocessing of eGFR trajectories within UK Biobank GP data). Chronic kidney disease (CKD) is diagnosed according to the ICD-10 codes detailed in the appendix, as shown in table 2, as specified by [31]. The included data

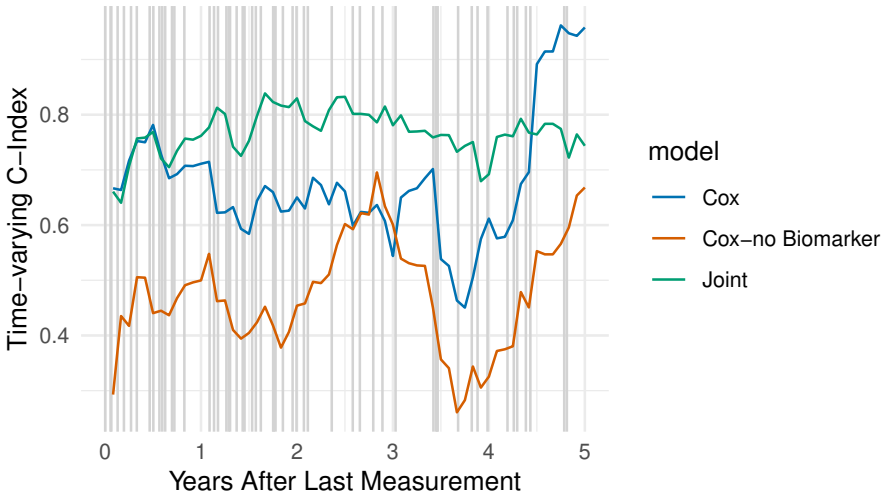

Fig. 7 Assessment of Serum Bilirubin and Primary Biliary Cirrhosis (Mayo Clinic): Evaluation of different models (joint model, and Cox model with/without biomarker information) using the time-varying C-index over a period of five years with an interval of 12 months. Vertical grey lines mark diagnosis time points.

are restricted to subjects with available covariate information and at least one measurement per year on average and a total of three measurements available during the 3-year period.

The diseased subgroup comprises all patients who received an initial diagnosis of chronic kidney disease (CKD) within three years prior to and up to five years after their visit to the assessment centre (visit V0 in UKBB). To create a comparable, similar-sized subgroup within the healthy cohort, propensity score matching is applied using the R package *MatchIt*[24], utilising Mahalanobis distance and the covariates age (UK Biobank Field ID: 21022), sex (UK Biobank Field ID: 31), and smoking status (UK Biobank Field ID: 20116). As a result, no covariates were incorporated into the longitudinal and survival submodel.

In this example, we use a 3-year period in which longitudinal data is collected, followed by a 5-year follow-up period in which only survival outcome data are recorded. The eGFR values are transformed using a Min-Max normalisation to obtain a distribution within a range of 0 to 1, targeting an average of approximately 0.5 for optimal application in guidelines. The normalised eGFR value at time t is denoted as $eGFR_{norm}(t)$.

Given these data, the guidelines of Table 2 are used to assess if better performance of the joint model compared to the Cox model can be expected in this scenario. For applying this approach to derive slope and noise parameters, it is assumed that $t_m = 1$ year. A summary of the derived parameters is given in Table 4. We note that some requirements are fulfilled (e.g. sample size) whereas other requirements are not

Parameter	Estimate	Comparison to Guidelines
Sample Size N	5272	Requirement fulfilled
Noise Standard Deviation σ_ϵ	0.042	Requirement not fulfilled
Percentage of Patients Responding p_{perc}	unclear	unclear
Number of measurements per year n_{abs}	$1.2(sick), 0.8(healthy)$	Requirement not fulfilled
Intercept difference Δ_b	0.119	Requirement fulfilled
Intercept standard deviation σ_b	$0.082, 0.073$	-
Slope Mean μ_m	-0.004	-
Slope Standard Deviation σ_m	0.004	Requirement fulfilled
Years of Assumed Slope t_m	1	unclear

Table 4 eGFR and chronic kidney disease: Estimation of parameters, derived by a mixed-effects model (see Appendix, Derivation of Parameter Choices Based on Real-World Data), and comparison with the guidelines given by table 2

fulfilled (e.g., noise is low, low number of measurements per year). Thus, it is unclear if a gain in performance of the joint model is expected.
For the joint model, the corresponding submodels are given by

Longitudinal submodel:

$$eGFR_{norm}(t) = \beta_0 + \beta_1 \cdot t + b_{i0} + b_{i1} \cdot t + \epsilon_i(t) \tag{13}$$

Survival submodel:

$$h_i(t) = h_0(t) \cdot exp\{\alpha m_i(t)\} \tag{14}$$

The Cox model with biomarker information is given by

$$h_{i,Cox_{biom}}(t) = h_0(t) \cdot exp\{eGFR_{norm}(\tilde{t})\} \tag{15}$$

with $\tilde{t}$ denoting the last observation time point of a eGFR value within the observation period. Since propensity score matching was performed, no covariates are included in the model.
We note that there is no gain in performance for the joint model compared to the Cox model (Figure 8). Therefore, in this example, there is no additional value for the more complex joint model.

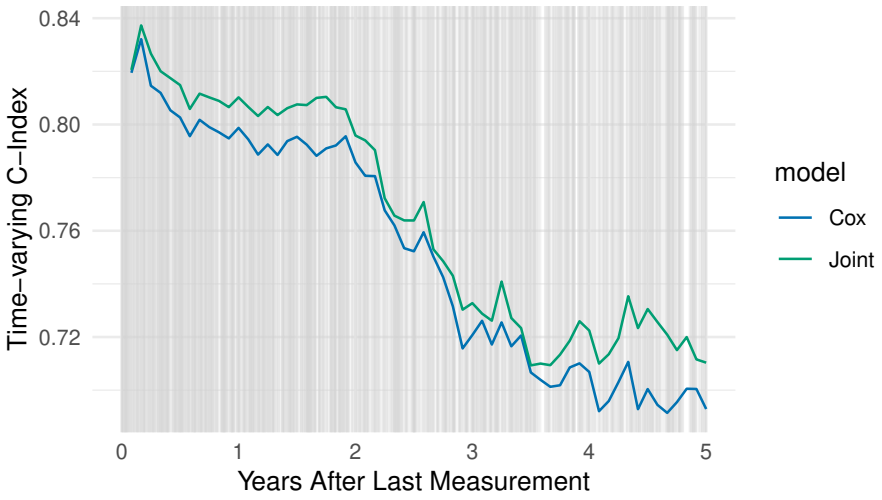

Fig. 8 Assessment of eGFR and Chronic Kidney Disease in the UK Biobank: Evaluation of different models (joint model, and Cox model with biomarker information) using the time-varying C-index over a period of five years with an interval of 12 months. Vertical grey lines mark diagnosis time points.

4 Discussion and Outlook

Identifying patients at risk for disease as early as possible is crucial. More and more statistical methodology is published to perform this task as accurately as possible. One example is a joint model which combines longitudinal and survival data into a single model. Furthermore, the availability of primary care data is continually expanding, providing richer datasets for analysis [32]. However, it was unclear how the quality characteristics of real-life EHR data influences the performance of the joint models. In this study, we identified several critical parameters that relevantly influence the performance of the joint model. The joint model outperforms the conventional Cox regression model in scenarios where data contain noise and when the frequency of longitudinal measurements is increased. The homogeneity of the patients' progress significantly influences the performance of the joint model, affected by both the count of patients showing an observable response and the consistency of the slope in the pre-disease phase. An increase in sample size substantially reduces the width of the prediction intervals derived from the iterated results in all models, and therefore increases the accuracy. In total, there are certainly scenarios in which the joint model outperforms the cox model - but also many scenarios in which no performance gain is observed. Therefore, the application of the more complex joint model - which is also using the full longitudinal trajectories, i.e., more information - needs to be considered carefully. For this purpose, we provide a checklist to allow the analyst to consider if the characteristics of the data offer the potential for a worthwhile application of the joint model.

When using real-world data to assess the predictive power of (longitudinal) measurements (bilirubin on PBC, eGFR on CKD) using the derived data quality-guidelines, the expected outcome following the guidelines match the performance outcomes. As all aspects are fulfilled for the first example, the joint model outperforms the Cox model in terms of the time-varying C-index, whereas for the second example not all requirements are fulfilled, causing that the joint model is not superior over the Cox model.
There are several further direction for research: the way in which the parameters interact with each other and how they may influence each other has not yet been analysed. For example, there is the possibility of a complex interplay between factors such as noise, sample size, and the effect being measured. Due to complexity considerations, an increased frequency of measurements over time, as is sometimes the case for specific markers, is not modelled. Additionally, due to computational time constraints, only a limited number of variations per parameter were considered in the analysis. This restriction suggests that a more thorough examination of a single parameter could benefit from simulating a larger number of values. Expanding the range of values for each parameter could lead to a more nuanced understanding of their effects and interactions, ultimately enhancing the robustness and reliability of the findings.
The joint model can be further modified, as a linear slope after a specific break point is rarely observed in real-world data. Looking at complex real-world data, joint models can quickly become complex, especially when multiple events and multivariate longitudinal observations are involved [33]. For example, the parametrisation of the current value is used [5], i.e. directly including the longitudinal fit as a covariate for the survival submodel. However, alternative associations could be explored, such as interaction effects or time-dependent slopes [33].
It may also be interested to get a better understanding of the quality of EHR data across data sources and health care system: so far, the data in the UK Biobank [34] has been used. There, data quality seems to be a challenge for the joint model. However, it is unclear how representative the EHR data is compared to other biobanks like BioVU [35] and Singapore Precise [36]. Furthermore, the application of this approach could extend to other areas of biostatistics, such as clinical trials that lead to a more consistent data collection and reduced variability, so that the trajectories are likely to be more homogeneous.
Ultimately, it may be interesting to establish similar guidelines for EHR data analysis for other types of complex modelling approaches.
This work and specifically the guidelines provided are valuable tools for identifying quality issues in EHR data and addressing the question of what data quality is required for the joint model. In addition, it can help in practical applications by providing insights on how to improve data quality, thereby improving the reliability and validity of joint modelling outcomes.

5 Conclusions

The research indicates that joint models, which combine longitudinal and survival data, often surpass traditional Cox regression models, especially in situations with high levels of data noise and frequent longitudinal measurements. The efficacy of these joint

models relies heavily on the uniformity of patient progress, as indicated by consistent observable responses among patients. Moreover, larger sample sizes contribute to narrower prediction interval widths, improving accuracy across all models. Although joint models may excel in certain scenarios, there are cases where they do not outperform Cox models, requiring careful consideration before use. Furthermore, the research formulated guidelines for evaluating the quality of real-world data, demonstrating that following these guidelines can accurately forecast joint model performance.
The study offers valuable guidelines to ensure data quality, aiding researchers and clinicians in selecting suitable models. By evaluating joint models alongside traditional methods, the research enhances statistical methodologies in clinical studies, guiding analysts in choosing the right techniques based on dataset characteristics. Ultimately, this study boosts the reliability of healthcare data analysis, contributing to improved health outcomes.

Declarations

Ethics approval and consent to participate: Data obtained from the UK Biobank received ethical approval from the North West Multi-Centre Research Ethics Committee (MREC). All participants provided informed consent to participate in the study and for their data to be used for research purposes.
Consent for publication: Not applicable.
Availability of data and materials: The simulated datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. The real-world data used can be accessed via the R-package JMbayes2 [15] and by the UK Biobank [34]. The UK Biobank data are not publicly available. Access to the UK Biobank data is available to researchers who apply for permission through the UK Biobank's application process. For more information on accessing the data, please visit the UK Biobank website at https://www.ukbiobank.ac.uk.
Competing interests: Berit Hunsdieck, Christian Bender and Johanna Mielke are employees of Bayer AG.
Funding: The authors declare that no funding was received for this research.
Authors' contributions: BH made substantial contributions to the conception, the analysis,and interpretation of data. CB made substantial contributions to the conception and substantively revised the work. KI made substantial contributions to the conception and substantively revised the work. JM made substantial contributions to the conception and substantively revised the work.
All authors read and approved the final manuscript.
Acknowledgements: Katja Ickstadt acknowledges the support of BMBF and MKW.NRW within the Lamarr-Institute for Machine Learning and Artificial Intelligence. This research has been conducted using the UK Biobank Resource (Application 28807).

Appendix

Additional details on methodology

Simulation of the Baseline characteristics

Baseline covariates are generated based on existing literature to mimic real-world scenarios. Specifically, the covariates included in the simulation were sex, age, and smoking status. These standard factors were chosen due to their common causalities with a broad range of diagnoses. For example, it is well established that smoking increases the risk of certain diseases, such as lung cancer [37]. By incorporating these covariates into the simulation, the study aims to better reflect the complexity of real patient data and to assess how these factors influence the onset and progression of the disease.
The distribution assumptions for the covariates given in Table 1 were used to evaluate the different influences of the baseline characteristics on the status of the disease.

Covariable	Distribution in healthy patients	Distribution in sick patients
Age	$58 + \mathcal{N}(2, 6)$	$62 + \mathcal{N}(2, 6)$
Sex	$Bin\,(1/3)$	$Bin\,(2/3)$
Smoking Status	$Bin\,(2/5)$	$Bin\,(2/3)$

Table 1 Distribution assumptions for simulated covariates; For Sex: 0=Female, 1=Male; For Smoking Status: 1=Currently, 0=Never (Patients who stopped smoking ignored)

UK Biobank

The UK Biobank consists of approximately 500,000 community-based participants aged 40 to 69 years, who were recruited throughout the UK between 2006 and 2010. All participants provided their informed written consent to participate in the study, which was approved by the National Research Ethics Service (11/NW/0382). Our study was conducted under access approval number 28807. In this study a variety of data were collected, including hospital inpatient data for the entire cohort and primary care records (including prescribed medications and coded clinical events, including consultations, diagnoses, procedures, and laboratory tests) for approximately 230.000 participants [34].

Derivation of Parameter Choices Based on Real-World Data

Given individual measurements $y(t_{ij})$ with their respective time points t_{ij} and diagnosis time point $t_{i,abs}$, the break point $t_i^{(b)}$ is defined as the specific time at which a significant change occurs in the behaviour of the measurements $y(t_{ij})$ for individual i. For now, it is assumed that the trend after reaching the break point is linear, as illustrated in Figure 1. Given a data set, we can estimate the parameters mentioned in the

chapter 2.1 by fitting a piecewise linear mixed-effects model based on the sick cohort. Therefore, we define $x_{1,i}(t)$ as the piece before the slope starts, resp. before the measurements are influenced by future diagnosis, and $x_{2,i}(t)$ as a piece when the slope is starting, resp. the measurements are influenced by future diagnosis,

$$x_{1,i}(t) = \begin{cases} t & , t < b_i \\ 0 & , t \geq b_i \end{cases}, \quad x_{2,i}(t) = \begin{cases} t - b_i & , t \geq b_i \\ 0 & , t < b_i \end{cases}$$

.

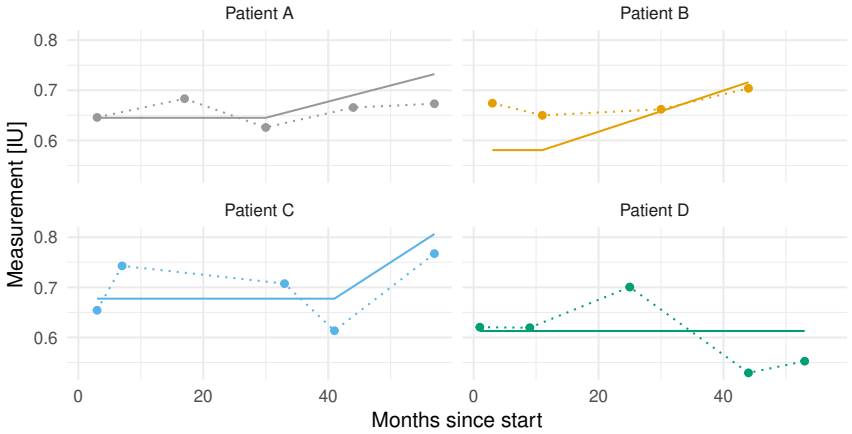

Fig. 1 Example of four trajectories (solid line) assuming a piecewise linear structure with individual break points at $b_A = 30, b_B = 11, b_C = 44, b_D > 60$. Dashed line/dots mark hypothetical measurements.

The piecewise linear mixed-effects model is then defined as

$$\begin{aligned} y_i(t) &= x_{1,i}(t) + \tilde{x}_{2,i}(t) + \epsilon_i(t) \\ \tilde{x}_{2,i}(t) &= x_{2,i}(t) + \eta_i \quad . \end{aligned}$$

An individual effect term, denoted as η_i, is incorporated to capture the individual slope term in the trajectory following the break point. The estimates given real-world data can be utilised for subsequent simulations, e.g. to analyse the data quality and its influence on further predictions. The analysis is carried out using the R package lme4 [38].

Preprocessing of eGFR trajectories within UK Biobank GP data

To effectively utilise the biomarker data of the Electronic Health Record (EHR), the simple extraction of biomarker measurements using read codes proves to be insufficient, and therefore a preprocessing pipeline must be applied.

The eGFR is calculated based on several time-independent characteristics of the patients and time-dependent levels of creatinine. More concretely, for deriving the EGFR value of a patient, the formula

$$\begin{aligned}\text{eGFR} =& 141 \cdot \min\left(\frac{\text{Scr}}{\kappa}, 1\right)^{\alpha} \cdot \max\left(\frac{\text{Scr}}{\kappa}, 1\right)^{-1.209} \cdot 0.993^{\text{Age}} \cdot [1.018 \text{ if female}] \\ & \cdot [1.159 \text{ if Black}] \end{aligned} \tag{1}$$

given by [39] is used. This equation calculates the estimated glomerular filtration rate (eGFR) using the CKD-EPI formula, where Scr is the serum creatinine concentration, κ is 0.7 for women and 0.9 for men, and α is -0.329 for women and -0.411 for men. The serum creatinine concentration within the UK Biobank is given by five different readcodes, containing two different units (umol/L and mmol/L). Units are aligned prior to the calculation of the eGFR. The read codes are given by clinical terms using version 3 (CTV3 or v3). We use the recommended creatinine range for outlier exclusion as given via the UK Biobank Showcase (see https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=30700, accessed: 5.12.2024) by $[10.7, 127.3]$. In order to restrict to the most relevant range of eGFR values (and to allow for a min-max normalisation which is only reasonable after the removal of outliers [40]), we include measurements within the interval $[1, 150]$.

For anonymisation purposes, if the combination of unit and read code appears in fewer than five subject identifiers (SIDs), the associated SIDs should be removed.

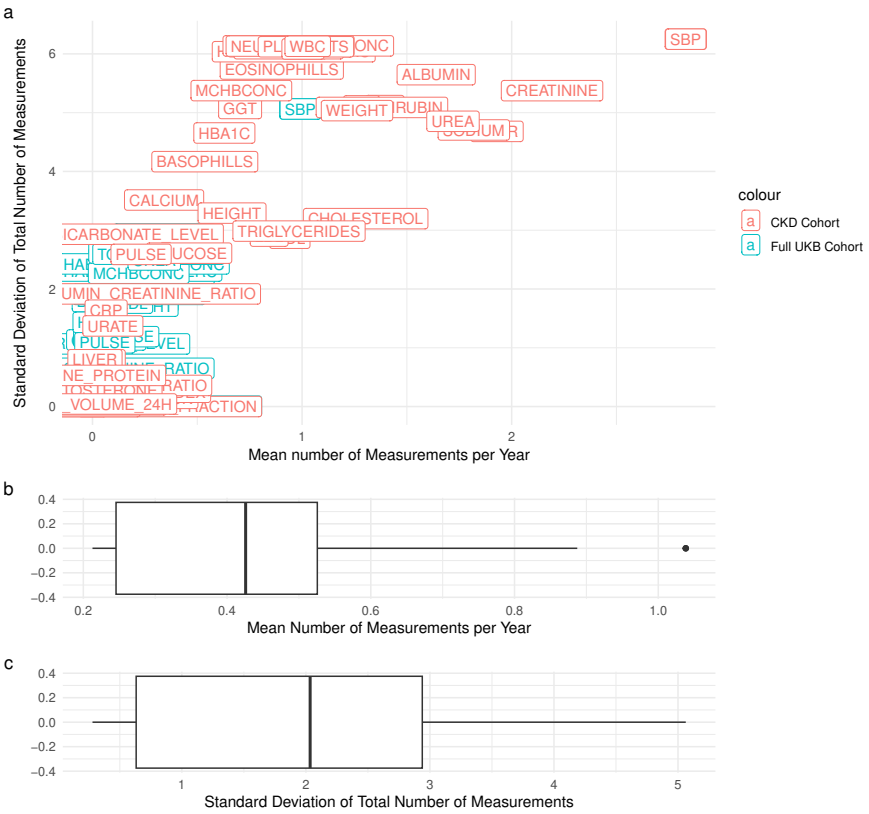

Fig. 2 The mean number of measurements and the corresponding standard deviation for 66 preprocessed biomarkers in the UK Biobank Primary Care data are presented (a), with boxplots providing a summary for the mean (b) and the standard deviation (c). The analysis considers data collected from up to five years prior to the baseline visit through to the baseline visit itself. The data distinguished in red pertain exclusively to individuals with a CKD diagnosis, while the data in blue represent the entire cohort of UK Biobank participants with Primary Care Data.

Supplementary tables

ICD9	585, 5859, 4030, 4031, 4039, 4040, 4041, 4049
ICD10	I120, I131, I132, N18, N180, N181, N182, N183, N184, N185, N188, N189

Table 2 ICD9 and ICD10 Codes Utilized in the Definition of Chronic Kidney Disease Patients

References

[1] Rizopoulos, D.: Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. Biometrics **67**(3), 819–829 (2011) https://doi.org/10.1111/j.1541-0420.2010.01546.x https://academic.oup.com/biometrics/article-pdf/67/3/819/53155811/biometrics_67_3_819.pdf

[2] Scherrer, J.F., Pace, W.D.: Will electronic health record data become the standard resource for clinical research? Family Practice **34**(5), 505–507 (2017) https://doi.org/10.1093/fampra/cmx055 https://academic.oup.com/fampra/article-pdf/34/5/505/20138417/cmx055.pdf

[3] Zhang, D., Shen, D., Initiative, A.D.N.: Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. PLOS ONE **7**(3), 1–15 (2012) https://doi.org/10.1371/journal.pone.0033182

[4] Tsiatis, A.A., Davidian, M.: Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica, 809–834 (2004)

[5] Ibrahim, J.G., Chu, H., Chen, L.M.: Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology (3), 2796–2801 (2010)

[6] Meiman, J., Freund, J.E.: Large data sets in primary care research. The Annals of Family Medicine **10**(5), 473–474 (2012) https://doi.org/10.1370/afm.1441 https://www.annfammed.org/content/10/5/473.full.pdf

[7] Botsis, T., Hartvigsen, G., Chen, F., Weng, C.: Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit on Translational Bioinformatics **2010**, 1–5 (2010)

[8] Therneau, T.M., Grambsch, P.M., Therneau, T.M., Grambsch, P.M.: The Cox Model. (2000)

[9] Albert, P.S.: A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. Statistics in medicine **31**(2), 143–154 (2012)

[10] Zhang, J., Kim, S., Grewal, J., Albert, P.S.: Predicting large fetuses at birth: do multiple ultrasound examinations and longitudinal statistical modelling improve prediction? Paediatric and perinatal epidemiology **26**(3), 199–207 (2012)

[11] Pinheiro, J., Bates, D., R Core Team: Nlme: Linear and Nonlinear Mixed Effects Models. (2022). R package version 3.1-157. https://CRAN.R-project.org/package=nlme

[12] Therneau, T.M.: A Package for Survival Analysis in R. (2022). R package version 3.3-1. https://CRAN.R-project.org/package=survival

[13] Rizopoulos, D.: The *R* Package **JMbayes** for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. Journal of Statistical Software **72**(7) (2016) https://doi.org/10.18637/jss.v072.i07 . Accessed 2024-04-18

[14] Rustand, D., Van Niekerk, J., Krainski, E.T., Rue, H., Proust-Lima, C.: Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested laplace approximations. Biostatistics **25**(2), 429–448 (2024)

[15] Rizopoulos, D., Papageorgiou, G., Miranda Afonso, P.: JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data. (2024). https://drizopoulos.github.io/JMbayes2/, https://github.com/drizopoulos/JMbayes2

[16] Park, S.Y., Park, J.E., Kim, H., Park, S.H.: Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). Korean Journal of Radiology **22**(10), 1697 (2021)

[17] Brentnall, A.R., Cuzick, J.: Use of the concordance index for predictors of censored survival data. Statistical methods in medical research **27**(8), 2359–2373 (2018)

[18] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2022). R Foundation for Statistical Computing. https://www.R-project.org/

[19] Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D.: Dplyr: A Grammar of Data Manipulation. (2023). R package version 1.1.4. https://CRAN.R-project.org/package=dplyr

[20] Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis. (2016). https://ggplot2.tidyverse.org

[21] Pedersen, T.L.: Patchwork: The Composer of Plots. (2024). R package version 1.2.0. https://CRAN.R-project.org/package=patchwork

[22] Wickham, H., Henry, L.: Purrr: Functional Programming Tools. (2023). R package version 1.0.2. https://CRAN.R-project.org/package=purrr

[23] Wickham, H., Vaughan, D., Girlich, M.: Tidyr: Tidy Messy Data. (2024). R package version 1.3.1. https://CRAN.R-project.org/package=tidyr

[24] Ho, D.E., Imai, K., King, G., Stuart, E.A.: MatchIt: Nonparametric preprocessing

for parametric causal inference. Journal of Statistical Software **42**(8), 1–28 (2011) https://doi.org/10.18637/jss.v042.i08

[25] Xie, Y., Bowe, B., Xian, H., Balasubramanian, S., Al-Aly, Z.: Renal function trajectories in patients with prior improved egfr slopes and risk of death. PloS one **11**(2), 0149283 (2016)

[26] Terry M. Therneau, Patricia M. Grambsch: Modeling Survival Data: Extending the Cox Model. Springer, New York (2000)

[27] Dickson, E.R., Grambsch, P.M., Fleming, T.R., Fisher, L.D., Langworthy, A.: Prognosis in primary biliary cirrhosis: model for decision making. Hepatology **10**(1), 1–7 (1989)

[28] Markus, B.H., Dickson, E.R., Grambsch, P.M., Fleming, T.R., Mazzaferro, V., Klintmalm, G.B.G., Wiesner, R.H., Van Thiel, D.H., Starzl, T.E.: Efficacy of liver transplantation in patients with primary biliary cirrhosis. New England Journal of Medicine **320**(26), 1709–1713 (1989)

[29] Tangri, N., Inker, L.A., Hiebert, B., Wong, J., Naimark, D., Kent, D., Levey, A.S.: A dynamic predictive model for progression of ckd. American Journal of Kidney Diseases **69**(4), 514–520 (2017) https://doi.org/10.1053/j.ajkd.2016.07.030

[30] Echouffo-Tcheugui, J.B., Kengne, A.P.: Risk models to predict chronic kidney disease and its progression: A systematic review. PLOS Medicine **9**(11), 1–18 (2012) https://doi.org/10.1371/journal.pmed.1001344

[31] Stroganov, O., Fedarovich, A., Wong, E., Skovpen, Y., Pakhomova, E., Grishagin, I., Fedarovich, D., Khasanova, T., Merberg, D., Szalma, S., Bryant, J.: Mapping of uk biobank clinical codes: Challenges and possible solutions. PLOS ONE **17**(12), 1–15 (2022) https://doi.org/10.1371/journal.pone.0275816

[32] Health Information Technology, O.: National Trends in Hospital and Physician Adoption of Electronic Health Records. Health IT Quick-Stat #61. Accessed: 2024-11-19 (2021)

[33] Lawrence Gould, A., Boye, M.E., Crowther, M.J., Ibrahim, J.G., Quartey, G., Micallef, S., Bois, F.Y.: Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. Statistics in Medicine **34**(14), 2181–2195 (2015) https://doi.org/10.1002/sim.6141 https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6141

[34] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R.: Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine **12**(3), 1–10 (2015) https://doi.org/10.

1371/journal.pmed.1001779

[35] Center, V.U.M.: BioVU: The Vanderbilt University Medical Center Biobank. Accessed: 2024-11-19. victr.vumc.org/biovu-description/

[36] Singapore, P.H.R.: PRECISE. Accessed: 2024-11-19. https://www.npm.sg/

[37] Walser, T., Cui, X., Yanagawa, J., Lee, J., Heinrich, E., Lee, G., Sharma, S., Dubinett, S.: Smoking and lung cancer: the role of inflammation. Proc Am Thorac Soc. 2008;5(8):811-815 (2008) https://doi.org/10.1513/pats.200809-100TH

[38] Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. Journal of Statistical Software **67**(1), 1–48 (2015) https://doi.org/10.18637/jss.v067.i01

[39] Levey, A.S., Stevens, L.A., Schmid, C.H., Zhang, Y., Castro III, A.F., Feldman, H.I., Kusek, J.W., Eggers, P., Van Lente, F., Greene, T., *et al.*: A new equation to estimate glomerular filtration rate. Annals of internal medicine **150**(9), 604–612 (2009)

[40] Vafaei, N., Ribeiro, R.A., Camarinha-Matos, L.M.: Comparison of normalization techniques on data sets with outliers. International Journal of Decision Support System Technology (IJDSST) (14(1)) (2022)

PLOS ONE

RESEARCH ARTICLE

A simulation-based framework for modeling and prediction of personalized blood pressure trajectories in hypertensive patients after antihypertensive treatment

Berit Hunsdieck[1,2*], **Johanna Mielke**[1☯], **Katja Ickstadt**[2,3☯], **Eren Elçi**[1☯]

1 Bayer AG, Research & Early Development, Division Pharmaceuticals, Wuppertal, Germany, **2** Department of Statistics, TU Dortmund University, Dortmund, Germany, **3** Lamarr-Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany

☯ These authors contributed equally to this work.
* berit.hunsdieck@bayer.com

OPEN ACCESS

Citation: Hunsdieck B, Mielke J, Ickstadt K, Elçi E. (2025) A simulation-based framework for modeling and prediction of personalized blood pressure trajectories in hypertensive patients after antihypertensive treatment. PLoS ONE 20(4): e0318549. https://doi.org/10.1371/journal.pone.0318549

Editor: Vinod Kumar Vashistha, University of Lucknow, INDIA

Received: October 07, 2024

Accepted: January 17, 2025

Published: April 10, 2025

Data availability statement: The underlying data used for further simulations is available via the UK biobank (see https://www.ukbiobank.ac.uk/) under Application #28807. The data yielding to the results presented in figure 4 (fig4.tif) are available via https://www.kaggle.com/datasets/berithu/logitudinal-blood-pressure.

Funding: The author(s) received no specific funding for this work.

Abstract

Hypertension, a leading global cause of death, poses challenges in stabilizing blood pressure within target values despite various therapeutic options, often necessitating multiple therapy adjustments and delayed impact assessments. Recently, the first wrist-based wearable blood pressure measurement devices were introduced which allow for a continuous assessment of blood pressure trajectories. This enables the development of statistical methodology for prediction of saturated steady-state of blood pressure under treatment—and thus allowing physicians to adjust the therapy earlier. As a prerequisite for the evaluation of such models and algorithms, it is necessary to simulate reliable and realistic hypothetical patient trajectories under treatment with antihypertensive medication. In this paper, we propose a simulation framework for blood pressure profiles through Pharmacokinetic-Pharmacodynamic modeling, which incorporates individual daily rhythms, patient characteristics, and medication effects. We also propose and evaluate two models for steady-state prediction under antihypertensive therapy, a Gaussian process and a non-linear mixed effect model. When only one day of measurements is available, the Gaussian process is preferred, but in real-world situations with more data, the non-linear mixed effect model is favored. It effectively reduces RMSE and bias in noisy data, outperforming the Gaussian process regardless of sample size.

Introduction

Hypertension is one of the leading causes of death worldwide. While there exist multiple therapeutic options, stabilization the blood pressure under recommended target values is challenging and often requires multiple adaptions of therapy. Traditionally, antihypertensive medicine is prescribed, and the impact of the therapy is only rechecked weeks after treatment

Competing interests: The authors have declared that no competing interests exist.

initiation. Globally, the number of hypertensive patients requiring anti-hypertensive medication increased from 594 million in 1975 to 1.3 billion in 2019, mainly in low- and middle-income countries (relative to growth in population) [1]. Starting at a blood pressure of 115/75 mmHg, the risk of death from a heart attack or stroke at ages 35 to 69 years doubles with every 20-point increase in systolic blood pressure, making it the leading cause of death [2]. High blood pressure accounts for nearly 10% of global health concerns [3]. According to the current state of the art, the efficacy of the anti-hypertensive drug is not checked in the doctor's office until a minimum of four weeks after it is taken for the first time. However, if the patient responds to the drug, an effect can be seen already after a few days [4]. Thus, an incorrect medication is identified only after weeks, although it could already be determined after a few days, whether the corresponding drug responds or not [4].

With the introduction of continuous blood pressure monitoring devices, early identification of ineffective medication becomes possible as this allows the physicians to continuously monitor (remotly) the impact of the therapy on blood pressure measurements. Physicians would be able to stop and adjust therapy early when it becomes clear that it is not leading to a steady-state blood pressure level (i.e., the state in which no additional change in blood pressure occurs as a result of further consistent administration) below the recommended targets. Current cuffless devices on the market are for example a chest patch developed by BioBeat Technologies [5], a wrist-wearable device developed by Aktiia [6], and a CAR-T ring developed by Skylabs [7]. We hypothesise that by making use of longitudinal data and statistical prediction models, sub-optimal blood pressure therapy can be detected even earlier by forecasting the expected saturated steady-state level of blood pressure as early as possible. This may help to identify patients with unsuccessful therapy even earlier.

Although none of these devices were accurately measuring blood pressure, they highlighted the need of re-evaluating the current practice of relying on a single blood pressure measurement in clinical settings [8]. Due to the novelty of wrist-based 24-hour blood pressure monitoring, there neither exists large data bases with relevant patient trajectories under recently initated blood pressure treatment nor simulation frameworks for generating artificial data. Therefore, for developing of forecasting methodology it is crucial to simulate realistic patient trajectories. In this paper, we propose such a simulation framework based on Pharmacokinetic-Pharmacodynamic (PKPD) modeling, which incorporates individual daily rhythms, patient characteristics, and medication effects. In order to integrate realistic patient covariate pattern, we integrate patient data from the UK Biobank [9].

Based on the simulated data, we develop two approaches to predict the steady state of a patient's blood pressure as early as possible. Our approach is, to the best of our knowledge, the first of its kind with a focus on antihypertensive therapy. Previous approaches aimed to forecast the blood pressure-lowering effect of antihypertensive medication on population level by identification of risk factors for unsuccessful drug titration for antihypertensive therapy. However, while these results are relevant for the overall population, none of them exhibits sufficient flexibility to ensure individual patient-dependent modeling. [10] introduced an initial approach to individual modeling based on general pharmacokinetic modeling. However, in this model the trajectory is predetermined by a fixed parametric form, rendering it less flexible. We optimize and tailor this approach to model patient trajectories under anti-hypertensive therapy. In addition, we explore Gaussian Processes (GP) which offer even more flexibility. The models are compared in a simulation study.

Based on simulated data incorporating the circadian rhythm, PKPD effects and measurements errors, the different models are evaluated. When only a single day of measurements is available, the Gaussian process is the preferred choice; however, in real-world scenarios with more data, the non-linear mixed effect model is preferred. It significantly lowers RMSE and

bias in noisy data, surpassing the performance of the Gaussian process regardless of sample size. Integrating the algorithm into wearable devices presents an innovative way to monitor medication adherence in hypertensive patients. It can predict the likelihood of achieving blood pressure reduction goals within the next day, allowing for timely adjustments to dosage and medication if necessary.

A simulation-based framework for patient trajectories under antihypertensive therapy

The proposed framework aims to generate realistic patient data from continuous blood pressure devices under antihypertensive medication, i.e. a trajectory of the change of blood pressure values. We build our model on the following components:

1. Medication effect: The modeling of the medication effect is the core of our model; we use a PKPD model to ensure that the simulated data is realistic
2. Circadian rhythm: Since blood pressure varies during the day, we include the daily rhythm in the model.
3. Inclusion of uncertainty: Since in reality trajectories contain measurement errors, it is important to include within-day as well as day-to-day variability.
4. Realistic covariate patterns: We leverage UK Biobank data for getting realistic distribution of relevant covariates (e.g., age, sex) in the relevant patient population (hypertensive patient without pre-treatment)—it is known that these covariates influence the treatment effect and baseline blood pressure values.

Fig 1 provides an example of the individual components for one subject, illustrating the simulated effects on diastolic blood pressure, taking into account medication effects (a), circadian rhythms (b), and within-day variability. As different medication groups may yield to distinct patterns, the simulation will cover a large amount of variation, allowing for the prediction of diverse patterns independent of the underlying medication group. In the following, we will describe in detail the underlying mechanisms of this figure.

PKPD medication effect

The PKPD modeling of the medication effect is the core of our framework. For each simulated patient, it is assumed that no medication is taken within the baseline period of two days (e.g. $t_{day} \in \{-2, -1\}$). The medication phase begins on day 0 ($t_{day} = 0$), with the medication being administered every morning at 8 a.m. from this point on. The corresponding timeline is illustrated in Fig 2.

The modeling of the medication effect is done by modeling the concentration-effect relationship. In pharmacology, this can usually be achieved by different PKPD modeling approaches. In a PKPD approach, the plasma concentration of the drug as a function of dose has to be modeled first by using a pharmacokinetic model. Given this time-dependent concentration, the effect can be modeled as a function of plasma concentration using a pharmacodynamic model. This scheme is illustrated in the Appendix, S1 Fig. In the following, we first describe the PK model and afterwards the PD model.

Pharmacokinetic modeling

Pharmacokinetic modeling uses compartment models to describe plasma concentration. These models vary in compartment number and absorption rates, such as single or three compartments in the literature. Although there are already complex compartment modeling approaches, a two-compartment model will be applied in the following analysis since

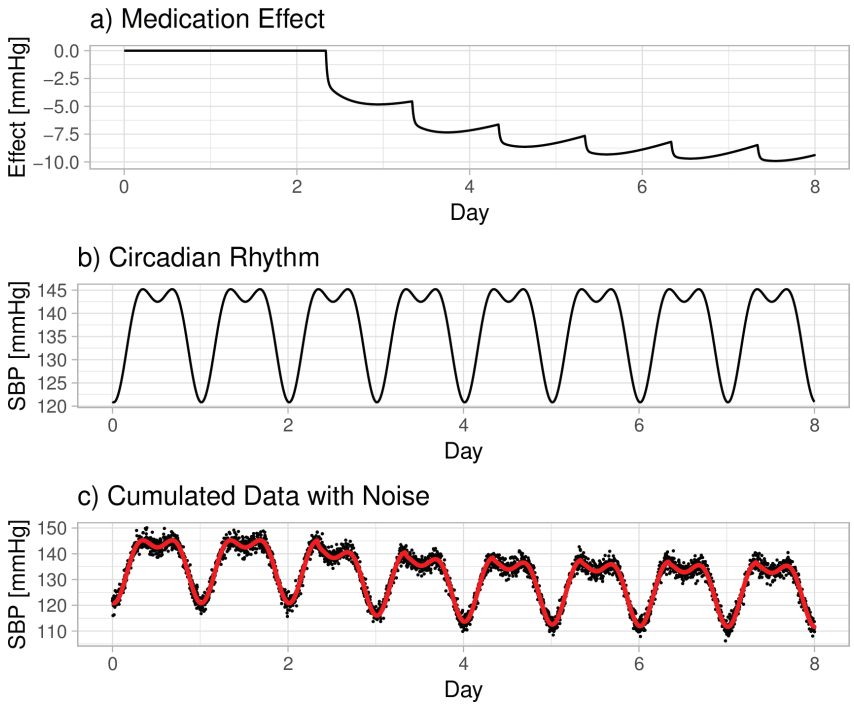

Fig 1. Example: Simulated trajectory of diastolic blood pressure for a hypothetical patient over a four-day period. (a) demonstrates the medication effect as predicted by a pharmacokinetic-pharmacodynamic (PKPD) model. (b) depicts the circadian rhythm's impact across the duration. (c) presents a combined view, with the circadian and medication effects overlaid in red, alongside the original data adjusted for within-day variability depicted in black.

https://doi.org/10.1371/journal.pone.0318549.g001

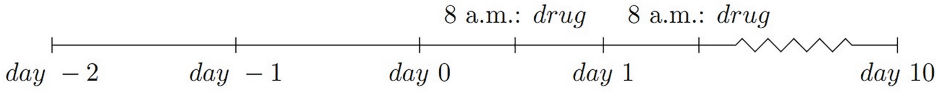

Fig 2. Timeline of simulated drug administration.

https://doi.org/10.1371/journal.pone.0318549.g002

this model is mainly used for modeling medication effects of blood pressure lowering drugs [11–13] . This choice is justified by the fact that a two-compartment PKPD model offers a balanced combination of simplicity, appropriateness, and clinical relevance, making it a preferred choice in many applications [14]. We opt for a widely-used two-compartment effect model, outlined in [15] and [16], with central plasma, gut, and peripheral tissue compartments, as visualized in Fig 3.

The underlying equations for the compartment model illustrated in Fig 3 are given by

$$\begin{aligned}\frac{dC(1)(t)}{dt} &= -K_a \cdot C(1)(t)\\ \frac{dC(2)(t)}{dt} &= \frac{-(K_{10}+K_{21})\cdot C(2)(t) + K_a \cdot C(1)(t) + K_{12}\cdot C(3)(t)}{VC}\\ \frac{dC(3)(t)}{dt} &= K_{21}\cdot C(2)(t) - K_{12}\cdot C(3)(t) \quad . \end{aligned} \tag{1}$$

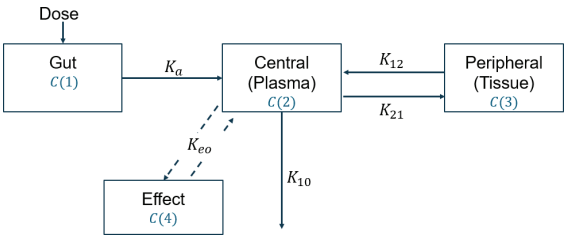

Fig 3. Relationships within the pharmacokinetic model: Compartmental model parameters $C(1)$, $C(2)$, and $C(3)$ represent compartments (Gut, Central, Peripheral), K_a, K_{10}, K_{21}, and K_{12} denote absorption and transport rates; Relationships within the Pharmacodynamic model: $C(4)$ represents effect compartment. K_{e0} denotes rate constant for transfer to the effect compartment.

https://doi.org/10.1371/journal.pone.0318549.g003

The parameters are defined as

$$\begin{aligned} &C(1): \text{Gut} \\ &C(2): \text{Central (Plasma)} \\ &C(3): \text{Peripheral (Tissue)} \\ K_a,\ K_{10} &:= \text{Rate of absorption} \\ &K_{21}: \text{Rate of transport of the drug from the central to the peripheral compartment} \\ &K_{12}: \text{Rate of transport of the drug from the peripheral to the central compartment} \\ &K_{e0}: \text{Rate constant for transfer to effect compartment.} \end{aligned} \tag{2}$$

One important additional parameter is the volume of distribution (or volume capacity, VC), which is the volume of the central compartment. It relates the total amount of drug in the body to the plasma concentration of the drug at a given time. The following equation applies:

$$VC[L] = \frac{\text{Amount of drug in the body [mg]}}{\text{Plasma concentration of drug [mg/L]}} \tag{3}$$

A drug with a high VC tends to leave the plasma and enter the extravascular compartments of the body. Conversely, a drug with a low VC tends to remain in the plasma, meaning that a lower dose of a drug is required to reach a given plasma concentration [17].

An increase of K_{10} corresponds to the situation that the concentration cannot be held at a high level for a long time. In general, the concentration is lower and the steady state is reached faster. An increase in K_{21} behaves similarly to an increase in K_{10}. The general plasma concentration reaches higher values and there are more significant peaks within one day. In addition, the steady state is reached more quickly. In contrast, an increase in K_{12} leads to a longer time to reach the steady state, and the decline within individual days is weaker.The steady state is defined as the plateau where the effect will not increase further. An increase in K_{e0} results in a slight increase in the effective height. An increase in K_a affects the pattern of plasma concentration within a day. It leads to a steeper decrease.

Parameter choice involves considering factors like time to peak concentration (for blood pressure medication: 0.25 to 2 hours [18]) and how age and weight impact clearance and volume capacity [19,20]. A balance between literature values and desired characteristics is essential to avoid unrealistic drug effect curves. The selected corresponding parameters are listed in section "Evaluation of proposed prediction approaches".

Pharmacodynamic modeling

Given the plasma concentration the drug effect can be modeled by common pharmacodynamic models. The specific choice of the concentration-effect relationships needs to consider that, in real life, the effect of anti-hypertensives will reach the steady state after a few days. For this scenario, the model typically used in literature is the sigmoid E_{max} (cf. [21], [22]).

Given the sigmoid E_{max}-model, the blood pressure lowering effect BP_{eff} is defined as

$$BP_{eff} = E_{max} \cdot \frac{CE}{CE + EC_{50}} \quad , \tag{4}$$

where CE is the drug concentration and E_{max} can be interpreted as the maximum effect of a specific drug and EC_{50} as the half-life time, e.g. the time period after which half of the effect has been achieved. These parameters vary depending on the drug and the subject.

In the literature, it is reported that the anti-hypertensive steady state effect is reached in approximately five to seven days based on calcium channel blockers [23,24]). This is used as a reference for simulation. The maximal effect (E_{max}) directly influences a drug's maximal effect, with baseline blood pressure and ethnicity playing roles. Higher baseline blood pressure leads to higher expected absolute drug effects [21]. Ethnicity can also impact drug effectiveness; for instance, the Angiotensin II receptor antagonist Eprosartan has little effect on blood pressure in Black Africans [25]. The selected corresponding parameters are listed in section "Evaluation of proposed prediction approaches".

Circadian rhythm. Often the time of day is not taken into account when blood pressure is assessed [26]. However, for continuous blood pressure trajectories, this is considered a crucial information. We assume a circadian rhythm during the day, which is influenced e.g. by sleeping phases [27]. An example of such a pattern is displayed in Fig 4.

For the simulation, we use a simplified Fourier construction with two cosine functions [28], which yields

$$SBP(t) = BSL + amp_1 \cdot cos(\frac{2\pi \cdot (t + hor)}{24}) + amp_2 \cdot cos(\frac{2\pi \cdot (t + hor)}{12}) \tag{5}$$

with

$$\begin{aligned} nadir &= (BSL - \frac{2}{3} \cdot change) \cdot (1 + exp(\nu)) \\ amp_2 &= \frac{1}{3} \cdot (BSL - nadir) - \frac{4}{9} \cdot change \\ &\quad - \frac{2}{9} \cdot \sqrt{6 \cdot change \cdot (nadir - BSL) + 4 \cdot change^2} \\ amp_1 &= BSL - nadir + amp_2 \quad . \end{aligned} \tag{6}$$

SBP denotes the systolic blood pressure, BSL denotes the baseline (systolic) blood pressure. The parameters $nadir$ and $change$ represent patient characteristics. First, $nadir$ is defined as

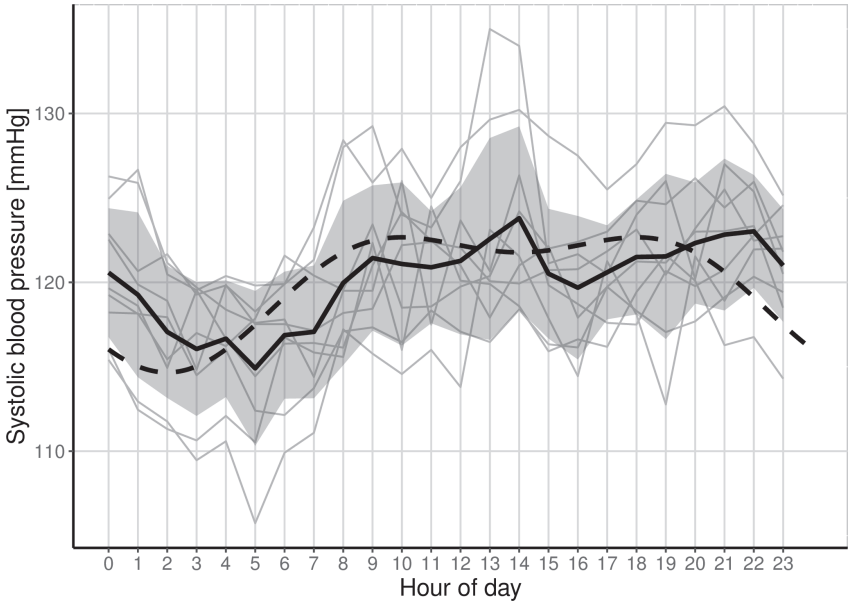

Fig 4. Internal study data with eleven voluntary, healthy employees wearing a wearable device for a time frame of around two weeks. In grey: Individual aggregated circadian rhythms; In black: Mean individual circadian rhythm; In black (dashed): Simulated circadian rhythm.

https://doi.org/10.1371/journal.pone.0318549.g004

the minimum systolic blood pressure during the night. Second, *change* is defined as the difference between the maximum systolic blood pressure during the day and the minimum systolic blood pressure during the night. The corresponding parameters can be derived from data given by an internal study. In the internal study, eleven voluntary, healthy employees tested a wearable device for a time frame of around two weeks.

The individual trajectories as well as the derived simulated trajectory based on internal study data are given in Fig 4. The specific parameter selection, which results from this, can be found in the subsequent section.

Variations from the ideal curve: Within-day and day-to-day variability. Noise on the data should be added for capturing the measurement error present in real-world data. This noise can fluctuate within a day (intra-daily, within-day) as well as between different days (inter-daily, day-to-day noise).

To account for additional noise within a day, such as coffee consumption, individual lifestyle, or environmental impacts, we can introduce variability by adding random noise to each measurement. For this intra-daily variability, we rely on data from current marketed wearable devices, such as Aktiia (see https://aktiia.com/de/, accessed on May 11, 2024). It was estimated that the within-day variability of individual measurements is approximately 5 mmHg. This is consistent with the variability measured in an internal study as well the variability given in [29].

We assume that the intra-daily variability is independently normally distributed around zero. When assuming that 5 mmHg corresponds to the 95%-quantile of this distribution, the final distribution is given by

$$noise_{intra} \sim \mathcal{N}(0, (5/z_{1.99/2})^2) = \mathcal{N}(0, 1.94^2)$$

for the intra-daily variability.

In addition, we assume inter-daily variability given by about ±8 mmHg. This can be validated by the internal study as well. Inter-daily variability in the measurements can be caused by unusual activities like doing sports or being sick. As before, assuming that the inter-daily variability is normally distributed around 0, the corresponding standard deviation can be derived by assuming 8 mmHg as the 99% quantile of it's distribution, which then yields to

$$noise_{inter} \sim \mathcal{N}(0, (8/z_{1.99/2})^2) = \mathcal{N}(0, 3.11^2) \quad .$$

Covariate-specific impact on blood pressure trajectories

We have already mentioned that covariates like the body weight can have a huge impact on the blood pressure as well as on the blood pressure lowering effect of anti-hypertensives [20] and, therefore, some of the parameters introduced above for the PKPD-model are dependent on patient specific covariates. A summary of the hypothesized relationships between (clinical) covariates and the individual submodels is presented in Fig 5: e.g, we assume that weight, age and sex both influences the *VC* and baseline blood pressure. Ethnicity influences the maximum effect of the antihypertensive drug. Indirectly, all components contribute to the individual blood pressure profiles. The individual literature sources can be found in the Appendix, S2 table.

Assuming that we have simulated a dataset with realistic patient characteristics, we now need to simulate a blood pressure trajectory per patient. For that, we need to consider that it is known that some parameters from the PKPD model need to be adjusted to account, for example, for different sex and age. More concretely, according to Fig 5, the above introduced parameters from the PKPD model, E_{max}, *VC*, and *BSL*, need to be adjusted by modeling the impact of the patient-specific covariates (relative to the population mean) by

$$\begin{aligned}
\mu_{BSL_i} &= BSL_0 + \alpha_{age} \cdot (age_i - \overline{age}) + \alpha_{sex} \cdot sex_i + \alpha_{weight} \cdot (weight_i - \overline{weight}) \\
&\Rightarrow BSL_i \sim \mathcal{N}(\mu_{BSL,i}, sd^2_{BSL}) \\
\mu_{VC,i} &= VC_0 + \gamma_{age} \cdot (age_i - \overline{age}) + \gamma_{weight} \cdot (weight_i - \overline{weight}) \\
&\Rightarrow VC_i \sim \mathcal{N}(\mu_{VC,i}, sd^2_{VC}) \\
\mu_{E_{max},i} &= E_{max,0} + \beta_{BSL} \cdot (BSL_i - \overline{BSL}) + \beta_{ethn} \cdot (ethn_i - \overline{ethn}) \\
&\Rightarrow E_{max,i} \sim \mathcal{N}(\mu_{E_{max,i}}, sd^2_{Emax})
\end{aligned} \tag{7}$$

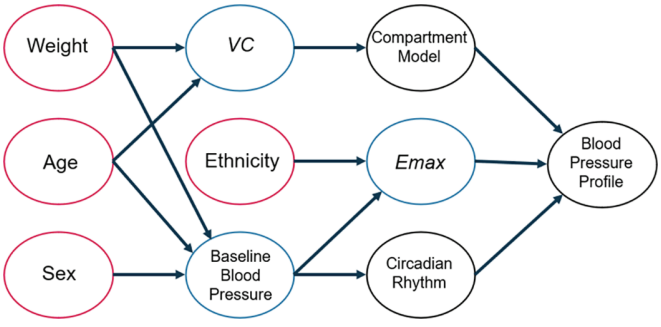

Fig 5. Impact of covariates on simulated blood pressure profiles; red: Clinical covariates, blue: Variables affected by covariates essential for PKPD modeling and circadian rhythm, black: Models encompassing circadian rhythm and medication effects.

https://doi.org/10.1371/journal.pone.0318549.g005

The above introduced parameters from the PKPD model, BSL_0, VC_0, and $E_{max,0}$, are the baseline parameters without any influence of additional covariates. The factors α_{age}, α_{sex}, α_{weight}, γ_{age}, γ_{weight}, β_{BSL}, and β_{ethn}, are derived by literature review (e.g., see [21,30,31]) as well as explorative analyses so that the effects given in literature are well expressed. The final values are given in the Appendix, S5 table. The corresponding reference values, i.e. for ethnicity, are given by the population means, i.e. denoted by $\overline{ethn}$. For the categorical variable ethnicity, $\overline{ethn} = 0$ denotes people of white ethnicity.

Models for forecasting individual patient trajectories

Based on the simulated data, the aim is to model the drug effect on blood pressure and predict the effect course of anti-hypertensive medication in the early stages of titration phases. Since the individual courses within a day are irrelevant for this goal, the daily average values are considered. In this section, we develop a non-parametric (GP) and a parametric (nlme) which can be used for forecasting individual patient trajectories. Let the time courses be aggregated by the daily means of the systolic blood pressure for N individuals given in the present data set. In addition, the covariates height, weight, age, sex, and ethnicity are assumed to be known for all individuals.

In the following, $i,\ i \in \{1, \dots, N\}$ denotes the individual patient ID and $t,\ t \in \{0, \dots, n_o\}$ the day of observation. $t_{max},\ t_{max} \in \{1, \dots, n_m\}$ is the day of last known observation in test data, which is constant over all test individuals. Further, $y_i(t)$ is the blood pressure value of individual i at day t. The corresponding predicted values of individual i at day $t,\ t \in \{1, \dots, n_o\}$, are denoted as $\hat{y}^{(GP)}_{i,t_{max}}(t)$ and $\hat{y}^{(nlme)}_{i,t_{max}}(t)$ based on the GP model and nlme model given t_{max} data points, respectively.

Parametric approach: Non-linear mixed effects model

In cases where additional information concerning the shape of a curve is available such information can be harnessed by a parametric model to optimize the parameters of interest. A widely adopted model for this purpose is the mixed effects model, which enables the estimation of both population-wide influences and subject-specific effects. In our use case, the model formula is motivated by [10], where the blood pressure reduction $BP_{Reduction}(t)$ at time t is modeled by an asymptotic approach to a plateau given by

$$BP_{Reduction}(t) = R_{max} \cdot \left(1 - exp\left(-\frac{t \cdot ln(2)}{dt_{1/2}}\right)\right) \tag{8}$$

with R_{max} as the maximal reduction in blood pressure, $dt_{1/2}$ as dynamic half-life time, i.e., time to reach 50% of maximal reduction, and t is the time since the beginning of therapy or taking medication.

To account for uncertainty in the analysis and enable more robust inference and predictive accuracy despite limited data, we will adopt a Bayesian modeling framework.

For our use case, we need to adjust the model given by formula (8) as our objective is to model individual patient trajectories. This specifically means that we need to allow for patient-specific saturation levels R_{max} and patients-specific half lives $dt_{1/2}$. We also introduce an additional population-wide parameter ω, that allows for additional variations in formula (8). To avoid high complexity and computational intensity of the model, an individual effect on ω is not considered, as it exerts no direct influence on effect magnitude or time to full effect. In other words, we assume that the general curve shape depends on ω, with variations

limited to half-life time and effect magnitude among individuals. To ensure that the estimators remain in the positive range, we log-transform the parameters, resulting in the following equation for the likelihood:

$$\hat{y}^{(nlme)}_{i,t_{max}}(t)|_{\eta_i} = exp(lR_{max,i,t_{max}}) \cdot \left(1 - exp\left(-\left(\frac{t \cdot ln(2)}{exp(lt_{1/2,i,t_{max}})}\right)^{exp(l\omega)}\right)\right) \tag{9}$$

$$lR_{\max,i,t_{max}} = lR_{\max,\text{pop}} + \eta_{i,1,t_{max}}, \quad lt_{1/2,i,t_{max}} = lt_{1/2,\text{pop}} + \eta_{i,2,t_{max}}$$

with $\boldsymbol{\eta}_{i,t_{max}} = (\eta_{i,1}, \eta_{i,2})^T$. The priors for the Bayesian framework are given by

$$lR_{\max,\text{pop}} \sim \mathcal{N}(1.2, 0.1), \quad lt_{1/2,\text{pop}} \sim \mathcal{N}(0.5, 0.2), \quad l\omega \sim \mathcal{N}(1, 0.1) \quad . \tag{10}$$

resulting from analysing the shape with regard to the known properties regarding mean half-life time and mean effect size. The model is fitted using Bayesian techniques using the *brms* R package [32]. For this model, we assume homoscedasticity, normal-distributed residuals and independence of individuals.

We assume that data is split in a training dataset (which is used for fitting the model) and then applied to unseen observations. In practice, the first step needs to be completed prior to the first application of this approach whereas the application to unseen observations represents the realistic scenario for a use in practice (novel patients need to get a forecast).

Non-parametric approach: Advanced Gaussian process

GPs provide a non-parametric method for estimating functions, allowing for the fitting of flexible models which capture the data while quantifying uncertainties in predictions. In hierarchical GPs, the data follows a nested structure given by a population-based GP and an individual GP. The population-based GP captures broad trends, while the individual GPs link observations to these trends. Prior knowledge enhances predictions [33,34].

To address challenges with convergence of Bayesian estimation as well as extrapolation problems using GPs, we propose an alternative approach: The proposed step-wise procedure, here called advanced GP, combines classical likelihood estimation methods and Bayesian methods in GP modeling. The analytical process consists of fitting a mean population GP to capture global trends. Residuals are then calculated. A Bayesian approach refines the model by estimating individual effects using training data. This completes the model building. For an estimation of effects on previously unseen data, we match the unseen observations to the subjects from the training data, identify trajectory-like data and make use of those for an accurate forecasting of the blood pressure saturation levels. More concretely, we follow these steps:

1. *Fit of Mean Population Gaussian Process:* In the initial step of our analysis, we fit a mean GP $\hat{y}_{mean,GP}(t)$ given by

$$\hat{y}_{mean,GP}(t)|(y_i(t), t) \sim \mathcal{N}(\mu_{mean}(t), \Sigma_{mean}(t)) \tag{11}$$

given the squared exponential kernel

$$k(x, x_\star) = \sigma^2 \cdot exp\left(-\frac{(x - x_\star)^2}{2l^2}\right) \quad . \tag{12}$$

Therefore, first the hyperparameters $\theta = (l, \sigma)$ are fitted by optimizing the log marginal likelihood

$$ln\, p(\boldsymbol{y}|\boldsymbol{X}, \theta) = -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{1}{2}ln\,|\boldsymbol{K} + \sigma_n\boldsymbol{I}| - \frac{n}{2}ln(2\pi) \quad , \tag{13}$$

given the training data $\{(t, y_i(t)) : t \in \{0, \ldots, n_o\}$, individual i included in training subset$\}$ and $\boldsymbol{K} = (k(t_i, t_j))_{i,j,\in\{0,\ldots,n_o\}}$. To achieve this, we employ the GPy Python package [35], which provides the necessary tools and functionalities for GP modeling. Given the fitted kernel $\hat{k}$ for the fitted hyperparameters $\hat{\theta} = (\hat{l}, \hat{\sigma})$, the fitted GP is used to estimate $f_\star$. $f_\star \in \mathbb{R}^{n_o+1}$ denotes the the estimated mean population values, which are be derived by

$$\begin{aligned} f_\star|(\boldsymbol{X}_\star, \boldsymbol{X}, f) \sim \mathcal{N}(&\hat{\boldsymbol{K}}(\boldsymbol{X}_\star, \boldsymbol{X}) \cdot \hat{\boldsymbol{K}}(\boldsymbol{X}, \boldsymbol{X})^{-1} f, \\ &\hat{\boldsymbol{K}}(\boldsymbol{X}_\star, \boldsymbol{X}_\star) - (\hat{\boldsymbol{K}}(\boldsymbol{X}_\star, \boldsymbol{X}) \cdot \hat{\boldsymbol{K}}(\boldsymbol{X}, \boldsymbol{X})^{-1}(\hat{\boldsymbol{K}}(\boldsymbol{X}, \boldsymbol{X}_\star)), \end{aligned} \tag{14}$$

with $(\boldsymbol{X}, f) = \{(t, y_i(t)) : t \in \{0, \ldots, n_o\}$, individual i included in training subset$\}$ given $\boldsymbol{X}_\star = (0, \ldots, n_o)$.

2. *Calculation of Remaining Residuals:* After fitting the mean GP model, we calculate the remaining residuals

$$\epsilon_i(t) = y_i(t) - \hat{y}_{mean,GP}(t) \quad . \tag{15}$$

These residuals represent the differences between the observed data points and the values predicted by the GP model.

3. *Bayesian Fit of GP for Training Data:* Moving forward, we transition into a Bayesian framework for model fitting. Specifically, we apply Bayesian methods to fit GP models on the remaining residuals $\epsilon_i(t)$, so that the corresponding estimates $\hat{\epsilon}_i(t)$ are given by

$$\hat{\epsilon}_i(t) \sim \mathcal{N}\left(\mu_{indiv}(t), \Sigma_{indiv}(t)\right) \tag{16}$$

at the individual level within the training data set. This Bayesian approach allows us to quantify uncertainties, and make probabilistic predictions based on the observed data.

4. *Bayesian Gaussian Process Modeling for Test Data:* When dealing with the test dataset, we employ a slightly different strategy. We identify a pre-selected number of individuals from the training data that are closest to each test case. To determine this proximity, we use a distance measure, specifically the smallest Euclidean distance

$$dist_{t_{max}}(\epsilon_i, \epsilon_j) = \sqrt{\sum_{t=1}^{t_{max}} (\epsilon_i(t) - \epsilon_j(t))^2} \tag{17}$$

computed from the residuals of the GP models calculated in step 2. The data of these five nearest individuals are selected to inform the subsequent modeling step and to serve as the basis for our Bayesian fitting of a GP. The selection process is driven by the need to have a representative and informative subset for modeling the test data, allowing us to leverage the information contained within the training data set. The resulting distribution is given by

$$\hat{\epsilon}_i(t)|\boldsymbol{S}_{nearest,i,t_{max}} \sim \mathcal{N}\left(\mu_{indiv}(t), \boldsymbol{\Sigma}_{indiv}(t)\right) \quad , \tag{18}$$

where $\boldsymbol{S}_{nearest,i,t_{max}}$ contains all data of the nearest training IDs and the data up to day t_{max} of individual i.

Combining the results of step 1 to step 4, the individual effect estimate $\hat{y}^{(GP)}_{i,t_{max}}(t)$ for individual i at time t is given by

$$\hat{y}^{(GP)}_{i,t_{max}}(t) = \hat{y}_{mean,GP}(t) + \hat{\epsilon}_i(t)\ . \tag{19}$$

Evaluation of simulation framework and comparison of statistical models

In this section, we describe how we use the proposed simulation framework to generate realistic patient trajectories. Afterwards, we explain how we test the two proposed forecasting approaches, namely the non-linear mixed-effects model and the advanced GP. We aim to compare the performance of the parametric and the non-parametric model for different simulation settings.

Simulation of realistic patient characteristics

In a simulation study, it is necessary to simulate a realistic set of covariates of each patient. In order to generate realistic patient-dependent datasets, we need to ensure that the underlying baseline characteristics are representative for a patient population (with high untreated blood pressure). Modeling realistic distributions of the covariates (e.g., generating a dataset with artificial patients with specific sex, age, height) can be achieved by mimicking a given data set of hypertensive patients and their covariates. One way to obtain a large data set of hypertensive patients without medication is given by the UK Biobank [9]. The UK Biobank gives access to detailed phenotype information as well as real world information regarding health status, genetics, clinical tests and many other parameters for around 500.000 adults across the UK. For the retrieval of realistic distributions of covariates, we focus on on patients reported not to take any antihypertensive medication, but still having a diastolic blood pressure of at least 90 mmHg or a systolic blood pressure of at least 140 mmHg which corresponds to so-called hypertension stage 2 (see Appendix, S1 table).

Evaluation of proposed prediction approaches

We aim to compare the proposed approaches in a simulation study. For that, we conducted an analysis based on 50 bootstrap test data sets. We employed a train-test split ratio of 2/3 to 1/3 to evaluate the model's performance. For the patients from the training dataset, all observations from day 0 to day 10 are included. For patients from the test dataset, we aim to mimic the situation in practice and therefore only include data for the one individual of interest with data spanning from day 0 to day t_{max} (where $1 \leq t_{max} \leq 5$). We aim to predict the blood pressure level at day $t = 10$.

We vary the number of days to be included for the test data (t_{max}) from 1 to 5 to identify the minimal number of observations for solid predictions. We vary the number of subjects ($N \in \{60, 120, 200\}$) to estimate a minimum sample size for the different approaches. Lastly, the influence of different numbers of neighbours for the GP are compared given two cases, namely 5 and 10 nearest neighbours.

The PKPD parameter choices for simulating the corresponding patient trajectories can be found in S3 table.

For the evaluation of the proposed methodlogy, we focus on two different strategies: Firstly, the model-based predicted blood pressure curve can be compared with the true "raw"

value (without noise). However, in a real-world setting, this value is always influenced by noise. Therefore, it is sensible to consider a second scenario, in which the estimated value is compared with the underlying measured noisy value, i.e. the observed value. We assess the models' quality of fit by calculating the bias and the root mean squared error (RMSE) per day.

Results

As outlined in previous chapters, the covariates must be simulated first in order to simulate the trajectories. This will be done based on the UK Biobank. A normal distribution is assumed, whereby the mean value and the standard deviation are estimated based on the UK Biobank cohort. Therefore, the UK Biobank patients with a systolic blood pressure of at least 140 mmHg or a diastolic blood pressure of at least 90 mmHg and no ongoing medication are obtained. The distribution of age and weight, separated by sex and ethnicity, is given by Table 1.

The resulting simulated dataset is based on these simulated covariates. Three concrete examples of simulated patient trajectories are given in Fig 6. The simulation incorporates both individual covariate-specific effects as well as within-day and day-to-day variability. It is noted that if specific characteristics of a mediciation class would be known (e.g., a specific half life), these could be easily integrated into the simulation framework. First, we compare the red lines (without noise) and note that the individual patient characteristics (gender, age, height, size, ethnicity) highly influence the observed trajectories. For example, Patient 2 observes a higher reduction of blood pressure. The introduced variability leads to further variation which are visible in the observed data (black dots).

Before systematically comparing the properties of the proposed forecasting approaches in the described simulation study, we aim to illustrate the characteristics in a single-patient example. In Fig 7 an example of individual predictions based on the different model types are given. The cohort size is given by $N = 200$. The forecasting tasks changes based on the available information from left to right: In the figure on the left, blood pressure measurement has only been measured for a single day after initiation of treatment. From left to right, more and more days of observations are given so that on the right, we have five days of blood pressure measurements given, which can be used for predicting the further course of blood pressure. In yellow, the corresponding predicted curve of using the GP with 5 nearest neighbours is given. In blue the non-linear mixed effects model is used for prediction. The black curve marks the underlying real trajectory without noise. It has been observed that, with limited information, the GP estimates the true curve more accurately than the nlme model. However, as more days of measurements are provided, the nlme model begins to outperform the advanced GP.

Table 1. Summary statistics of covariates age (in years), weight (in kg), and sex in patients with hypertension stage 2, i.e., diastolic blood pressure greater than 90 mmHg or systolic blood pressure greater than 140 mmHg (see Appendix, S1 table), in the UK Biobank data.

Sex	Ethnicity	Covariate	Mean	Standard deviation	Cohort size
Female	Black	Weight	82.3	16.8	2.228
		Age	54.2	7.94	
Female	White	Weight	73.3	14.7	109.327
		Age	59	7.19	
Male	Black	Weight	86.5	14.4	1.025
		Age	51.6	7.95	
Male	White	Weight	86.3	13.8	79.705
		Age	57.2	7.92	

https://doi.org/10.1371/journal.pone.0318549.t001

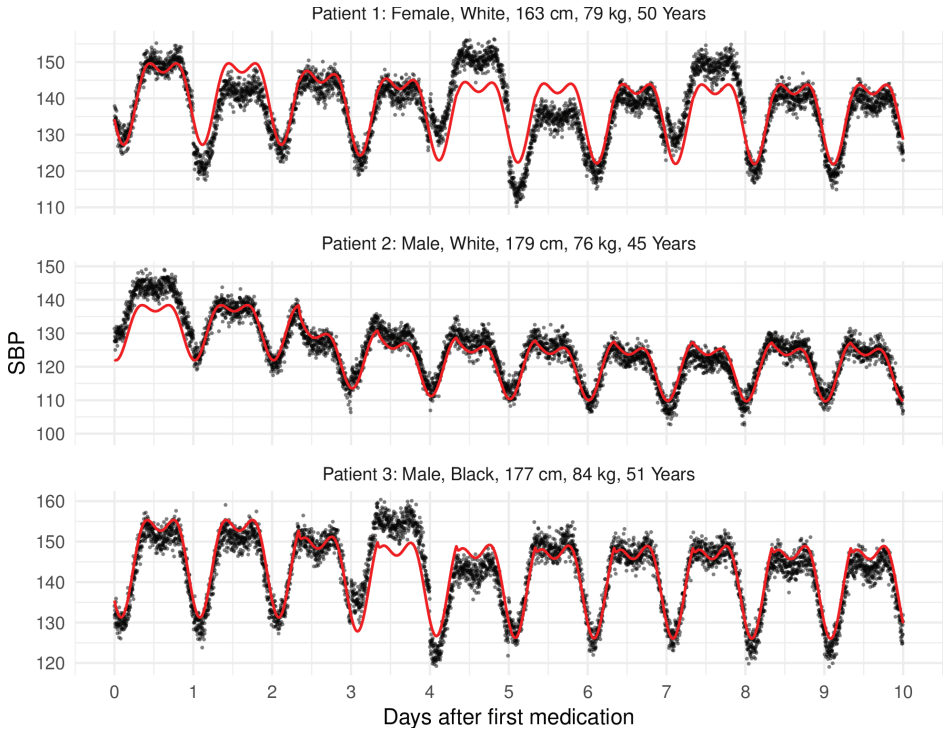

Fig 6. Simulated trajectories of systolic blood pressure after taking anti-hypertensive medication of three hypothetical patients. The red line corresponds to the raw blood pressure curve without any noise.

https://doi.org/10.1371/journal.pone.0318549.g006

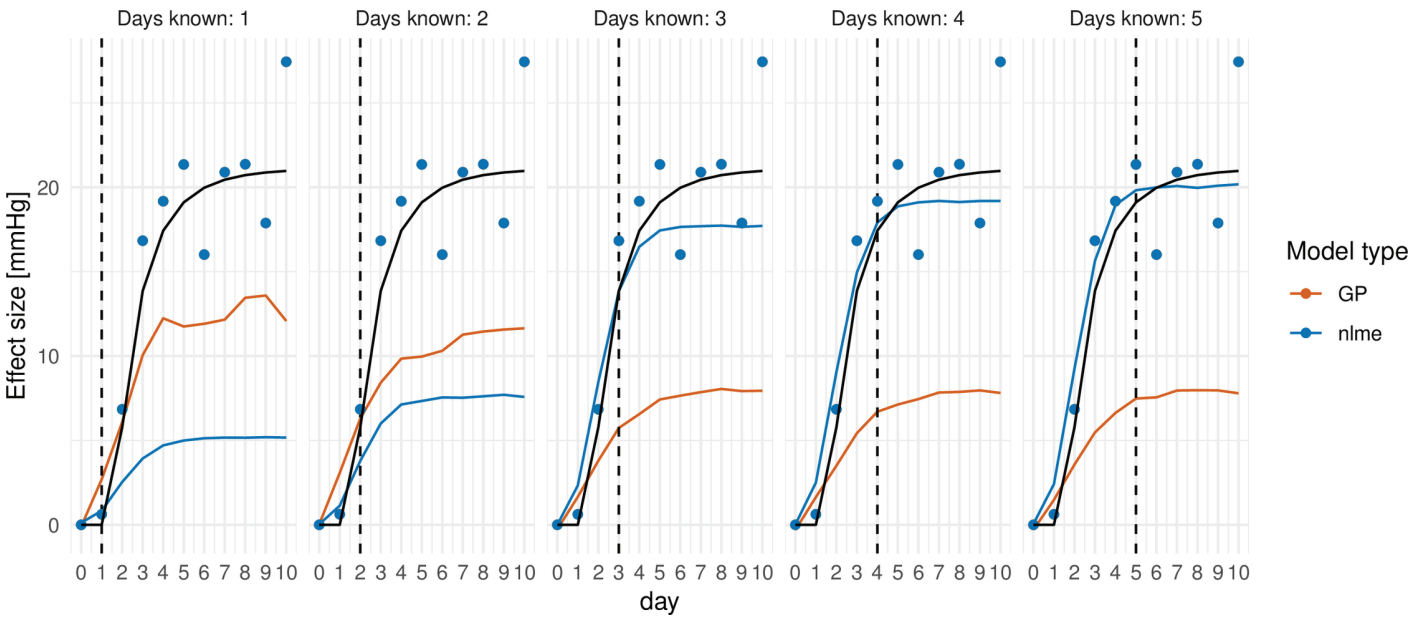

Fig 7. Prediction of individual effect curve using a non-linear mixed-effects model (nlme) and an advanced Gaussian process model (GP) with five nearest neighbours, given t_{max}, $t_{max} \in \{1, \ldots, 5\}$, days of measurements. In black: true curve of medication effect. The dots represent the observed values.

https://doi.org/10.1371/journal.pone.0318549.g007

Specifically, after five days of measurements, the estimated trajectory closely aligns with the true (unobserved) trajectory.

Generalising this to the full cohort with $N = 200$ patients, we will further focus on the comparing the predicted blood pressure values to the observed (noisy) values since in a real-world data setting, the true underlying values without any noise are not accessible. To ensure robustness, 50 datasets are simulated and the performance is compared.

First, the bias and rmse results are evaluated on a daily basis as illustrated in Fig 8. With a maximum absolute bias of approximately 1 mmHg all models are almost unbiased. While for the nlme model, we observe that adding in more information, i.e., increasing the number of available days with measurement, improves the performance by lowering the RMSE, the same is not true for the GP models: there, the RMSE is not reduced if more data is added. Looking at the RMSE and bias based on the noise-free values (see Appendix, S2 Fig), similar observations can be made.

One possible reason for this could be that the non-linear mixed-effects model implicitly has access to all training data during fitting. When NLME models are fitted to the data, they integrate information from both the fixed and random effects. The random effects are capturing the variation between different entities or subjects in the data set. In the GP models, only the population mean of these data points influences the model, while the individual fits of the residuals have access to only a fraction of the training data. Therefore, the NLME model

Fig 8. Root mean squared error (RMSE) and bias across days ($t = 0 \dots, 10$) compared to actually observed values, models (non-linear mixed-effects model (nlme) in blue, Gaussian process (GP) with five resp. ten nearest neighbours in yellow resp. red) ($t_{max} = 1, \dots, 5$, vertical line).

https://doi.org/10.1371/journal.pone.0318549.g008

may have the capability to fit the data more accurately, leading to a lower RMSE. This could be especially true when the data contain a significant amount of noise or individual variability.

Looking at the performance of both models, the advanced GP performs slightly better for a single day in terms of RMSE, while the nlme approach demonstrates improved performance over time based on the bias as well as the RMSE. Therefore, the GP is the preferred method only if very limited data is available.

These findings appear also independent across a wide range of realistic sample sizes (N = 60 to N = 200). We also note that there is no major gain in performance if more subjects are added (see Appendix, S3 Fig) indicating that the saturation level in terms of performance is reached early.

Conclusion and outlook

This paper presents a novel simulation framework for generating blood pressure profiles, which integrates Pharmacokinetic-Pharmacodynamic modeling with individual daily rhythms, patient characteristics, and medication effects. This model is very transparent and highly flexible and therefore can be adjusted to different medication classes. Also, the underlying patient characteristics are taken into account—allowing for tailoring to a target patient population, for example if inclusion/exclusion criteria for a clinical trial are already known.

Additionally, two models for predicting steady-state responses under anti-hypertensive therapy are developed and applied, namely an advanced Gaussian process and a non-linear mixed effect model. In scenarios where only one day of measurements is available, the Gaussian process is the preferred choice. However, in the real-world setting including more data is preferred. When multiple days of data are available, the nlme model becomes the preferred choice. It significantly reduces the RMSE as well as the bias for data with noise compared to the Gaussian process and demonstrates superior performance independent of the given sample size.

The algorithm's potential integration into wearable devices offers a innovative solution for monitoring medication adherence in hypertensive patients. The algorithm has the capability to provide a current probability or forecast of whether the patient will achieve the desired blood pressure reduction goal within the next day. This information can be used to make early adjustments to the dosage and medication if needed.

A growing interest has emerged in forecasting the individual effects of anti-hypertensive medications. This trend is supported by recent studies, including [36], which explored the use of machine learning algorithms to predict responses to antihypertensive therapy.

This innovative approach, leveraging real-time assessments through wearable technology, streamlines monitoring, enhances treatment effectiveness, and contributes to better blood pressure control. This integration signifies a promising step towards personalized healthcare for hypertension management. In further steps, the different algorithms have to be tested on real-world data to re-evaluate its performance and robustness given real-world data.

Appendix

Supporting information

Simulation

S1 Table. Overview of systolic and diastolic blood pressure (SBP and DBP) ranges for different hypertension stages [37].
(JPG)

S1 Fig. Pharmacokinetic-Pharmacodynamic scheme. top left: Relationship between time and plasma concentration given a single dose, top right: Relationship between plasma concentration and effect given a single dose, bottom: Relationship between time and effect given a single dose (based on [38]).
(TIF)

S2 Table. Literature sources for the influence of covariates. Based on Fig 5.
(JPG)

S3 Table. Parameter choices for simulating the circadian rhythm, based on internal study c.f. [39]. The parameters for the average circadian rhythm are determined based on the parameter estimators provided by [39] and are further modified by the parameters derived from the average circadian rhythm observed in the internal study (see Fig 4).
(JPG)

S4 Table. Parameter choices for simulating the medication effect using Pharmacokinetic-Pharmacodynamic (PKPD) modeling assuming normal distribution; For E_{max} covariate effects need to be added (see S5 tab).
(JPG)

S5 Table. Parameter choices for simulating the influence of the covariates Age, Weight, Ethnicity, and Sex, on baseline blood pressure (BSL), maximum effect (E_{max}), and volume capacity (VC).
(JPG)

Modeling approaches

Derivation of model formula for non-linear mixed effects model. The equation can be motivated by the cumulative distribution function of the exponential distribution

$$f(x) = 1 - e^{-\lambda x}, \qquad \lambda \in \mathbb{R} \tag{20}$$

which is generalized by

$$f(x) = a - e^{-b \cdot x + c}, \qquad \lambda, a, b, c \in \mathbb{R} \tag{21}$$

so that the plateau value denoted by a and the factor b determining the time to reach the plateau can vary. Additionally, in our application zero must be mapped to zero since there is no medication effect prior to the first dose, which leads to

$$f(x) = a - e^{-bx + ln(a)} \ , \tag{22}$$

which finally leads to formula (9).

Results

S2 Fig. Comparing predictive accuracy based on noise-free values. Root mean squared error (RMSE) and bias across days ($t = 0 \ldots, 10$) based on values without noise, models (non-linear mixed-effects model (nlme) in blue, advanced Gaussian process (GP) with five resp. ten nearest neighbours in yellow resp. red), and known measurements ($t_{max} = 1, \ldots, 5$).
(TIF)

S3 Fig. Comparison of sample size effect on root mean squared error (RMSE) and bias for observed values and noise-free values. Given five days of measurements: (Non-linear in

yellow) mixed-effects model (nlme) in blue, advanced Gaussian process (GP) with five nearest neighbours. We assess the impact of different sample sizes on the model performance using the five nearest neighbours for the GP model. Three different scenarios are analysed, given by sample sizes of $N = 60$, $N = 120$, and $N = 200$. Again, for robustness 50 datasets are simulated and assessed for each case.

The RMSE as well as the bias are evaluated for noise-free and noisy values given 5 days of measurements. It can be seen that increasing the sample size will not change the models performance in terms of the RMSE, neither looking at the noisy nor the noise-free values. Looking at the bias, a higher sample size does not change the bias of the nlme model. Using the GP model, a higher sample size can reduce the bias.

This suggests that, given the data assumptions (such as noise, rhythm, effect size, etc.), a sample size of approximately 120 patients would be sufficient to make an adequate prediction based on 5 days of provided measurements.
(TIF)

Acknowledgments

Katja Ickstadt acknowledges the support of BMBF and MKW.NRW within the Lamarr-Institute for Machine Learning and Artificial Intelligence. This research has been conducted using the UK Biobank Resource (Application 28807).

Author contributions

Conceptualization: Berit Hunsdieck, Johanna Mielke, Katja Ickstadt, Eren Elçi.

Formal analysis: Berit Hunsdieck.

Methodology: Berit Hunsdieck, Johanna Mielke, Eren Elçi.

Software: Berit Hunsdieck.

Supervision: Katja Ickstadt.

Visualization: Berit Hunsdieck.

Writing – original draft: Berit Hunsdieck.

Writing – review & editing: Johanna Mielke, Katja Ickstadt, Eren Elçi.

References

1. WHO. Hypertension; 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/hypertension
2. WHO. Global report on hypertension: the race against a silent killer. 2023.
3. Gaziano TA, Bitton A, Anand S, Weinstein MC. The global cost of nonoptimal blood pressure. J Hypertens. 2009;27(7):1472–7. https://doi.org/10.1097/hjh.0b013e32832a9ba3
4. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J. 2018;39(33):3021–104. https://doi.org/10.1093/eurheartj/ehy339 PMID: 30165516
5. Eisenkraft A, Goldstein N, Merin R, Fons M, Ishay AB, Nachman D, et al. Developing a real-time detection tool and an early warning score using a continuous wearable multi-parameter monitor. Front Physiol. 2023;14:1138647. https://doi.org/10.3389/fphys.2023.1138647
6. Sola J, Vybornova A, Fallet S, Polychronopoulou E, Wurzner-Ghajarzadeh A, Wuerzner G. Validation of the optical Aktiia bracelet in different body positions for the persistent monitoring of blood pressure. Sci Rep. 2021;11(1):20644. https://doi.org/10.1038/s41598-021-99294-w PMID: 34667230

7. Kim J, Chang S-A, Park SW. First-in-human study for evaluating the accuracy of smart ring based cuffless blood pressure measurement. J Korean Med Sci. 2024;39(2):e18. https://doi.org/10.3346/jkms.2024.39.e18 PMID: 38225785
8. Schutte AE. Wearable cuffless blood pressure tracking: when will they be good enough? J Hum Hypertens. 2024;38(9):669–72. https://doi.org/10.1038/s41371-024-00932-3 PMID: 38997475
9. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379
10. Lasserson DS, Buclin T, Glasziou P. How quickly should we titrate antihypertensive medication? Systematic review modelling blood pressure response from trial data. Heart. 2011;97(21):1771–5. https://doi.org/10.1136/hrt.2010.221473 PMID: 21586424
11. Baek I, Yun M, Yun H, Kwon K. Pharmacokinetic/pharmacodynamic modeling of the cardiovascular effects of beta blockers in humans. Arch Pharm Res. 2008;31(6):814–21. https://doi.org/10.1007/s12272-001-1231-4 PMID: 18563366
12. Heo YA, Holford N, Kim Y, Son M, Park K. Quantitative model for the blood pressure-lowering interaction of valsartan and amlodipine. Br. J. Clin. Pharmacol. 2016;82(6):1557–1567. https://doi.org/10.1111/bcp.13082
13. van Rijn-Bikker PC, Ackaert O, Snelder N, van Hest RM, Ploeger BA, Koopmans RP, et al. Pharmacokinetic-pharmacodynamic modeling of the antihypertensive effect of eprosartan in Black and White hypertensive patients. Clin Pharmacokinet. 2013;52(9):793–803. https://doi.org/10.1007/s40262-013-0073-6 PMID: 23696281
14. Rowland M, Tozer T. Clinical pharmacokinetics and pharmacodynamics. 2019.
15. Upton RN. The two-compartment recirculatory pharmacokinetic model–an introduction to recirculatory pharmacokinetic concepts. Br J Anaesth. 2004;92(4):475–84. https://doi.org/10.1093/bja/aeh089 PMID: 14766714
16. Baek I, Yun M, Yun H, Kwon K. Pharmacokinetic/pharmacodynamic modeling of the cardiovascular effects of beta blockers in humans. Arch Pharm Res. 2008;31(6):814–21. https://doi.org/10.1007/s12272-001-1231-4 PMID: 18563366
17. Oie S. Drug distribution and binding. J Clin Pharmacol. 1986;26(8):583–6. https://doi.org/10.1002/j.1552-4604.1986.tb02953.x PMID: 3793947
18. Weber C, Birnböck H, Leube J, Kobrin I, Kleinbloesem CH, Brummelen P. Multiple dose pharmacokinetics and concentration effect relationship of the orally active renin inhibitor remikiren (RO 42-5892) in hypertensive patients. Br J Clin Pharmacol. 1993;36(6):547–554. https://doi.org/10.1111/j.1365-2125.1993.tb00413.x
19. Tsai MC, Wu J, Kupfer S, Vakilynejad M. Population pharmacokinetics and exposure-response of a fixed-dose combination of azilsartan medoxomil and chlorthalidone in patients with stage 2 hypertension. J Clin Pharmacol. 2016;56(8):988–98. https://doi.org/10.1002/jcph.684 PMID: 26632101
20. Lee H, Jang I-J, Yu K-S, Choi J, Oh B-H. A population pharmacokinetic analysis of fimasartan, a selective angiotensin II receptor antagonist, in healthy caucasian subjects and korean patients with hypertension. Clin Pharmacol Drug Dev. 2013;2(2):162–72. https://doi.org/10.1002/cpdd.10 PMID: 27121670
21. Otani Y, Kasai H, Tanigawara Y. Pharmacodynamic analysis of hypertension caused by lenvatinib using real-world postmarketing surveillance data. CPT Pharmacomet Syst Pharmacol. 2021;10(3):188–98. https://doi.org/10.1002/psp4.12587 PMID: 33471960
22. Larsson R, Karlberg BE, Gelin A, Aberg J, Regårdh CG. Acute and steady-state pharmacokinetics and antihypertensive effects of felodipine in patients with normal and impaired renal function. J Clin Pharmacol. 1990;30(11):1020–30. https://doi.org/10.1002/j.1552-4604.1990.tb03589.x PMID: 2243149
23. Terakawa M, Tokuma Y, Kuwahara N, Shishido A, Noguchi H. Multiple-dose pharmacokinetics of nilvadipine in healthy volunteers. J Clin Pharmacol. 1988;28(4):350–5. https://doi.org/10.1002/j.1552-4604.1988.tb03157.x PMID: 3392233
24. Jackson R, Bellamy M. Antihypertensive drugs. BJA Education. 2015;15(6):280–5. https://doi.org/10.1093/bjaceaccp/mku061
25. van Rijn-Bikker PC, Ackaert O, Snelder N, van Hest RM, Ploeger BA, Koopmans RP, et al. Pharmacokinetic-pharmacodynamic modeling of the antihypertensive effect of eprosartan in Black and White hypertensive patients. Clin Pharmacokinet. 2013;52(9):793–803. https://doi.org/10.1007/s40262-013-0073-6 PMID: 23696281
26. Costello HM, Gumz ML. Circadian rhythm, clock genes, and hypertension: recent advances in hypertension. Hypertension. 2021;78(5):1185–96. https://doi.org/10.1161/HYPERTENSIONAHA.121.14519 PMID: 34601963

27. Douma LG, Gumz ML. Circadian clock-mediated regulation of blood pressure. Free Radic Biol Med. 2018;119:108–14. https://doi.org/10.1016/j.freeradbiomed.2017.11.024 PMID: 29198725
28. van Rijn-Bikker PC, Snelder N, Ackaert O, van Hest RM, Ploeger BA, van Montfrans GA, et al. Nonlinear mixed effects modeling of the diurnal blood pressure profile in a multiracial population. Am J Hypertens. 2013;26(9):1103–13. https://doi.org/10.1093/ajh/hpt088
29. Chia Y, Kario K, Tomitani N, Park S, Shin J, Turana Y, et al. Comparison of day-to-day blood pressure variability in hypertensive patients with type 2 diabetes mellitus to those without diabetes: Asia BP@Home study. J Clin Hypertens. 2019;22(3):407–414. https://doi.org/10.1111/jch.13731
30. Trocóniz IF, de Alwis DP, Tillmann C, Callies S, Mitchell M, Schaefer HG. Comparison of manual versus ambulatory blood pressure measurements with pharmacokinetic-pharmacodynamic modeling of antihypertensive compounds: application to moxonidine. Clin Pharmacol Ther. 2000;68(1):18–27. https://doi.org/10.1067/mcp.2000.106907 PMID: 10945312
31. Heo Y, Holford N, Kim Y, Son M, Park K. Quantitative model for the blood pressure-lowering interaction of valsartan and amlodipine. Brit J Clinical Pharma. 2016;82(6):1557–67. https://doi.org/10.1111/bcp.13082
32. Bürkner P-C. brms: An R package for Bayesian multilevel models using stan. J Stat Soft. 2017;80(1). https://doi.org/10.18637/jss.v080.i01
33. Quadrianto N, Kersting K, Xu Z. In: Sammut C, Webb GI, editors. Gaussian process. Boston, MA: Springer; 2010. p. 428–439. https://doi.org/10.1007/978-0-387-30164-8_324
34. Rasmussen C, Williams C. Gaussian processes for machine learning. MIT Press; 2008.
35. GPy. GPy: A Gaussian process framework in python. 2012.
36. Mroz T, Griffin M, Cartabuke R, Laffin L, Russo-Alvarez G, Thomas G, et al. Predicting hypertension control using machine learning. PLoS One. 2024;19(3):e0299932. https://doi.org/10.1371/journal.pone.0299932 PMID: 38507433
37. Whelton PK, Carey RM, Aronow WS, Casey DE Jr, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American college of cardiology/American heart association task force on clinical practice guidelines. Hypertension. 2018;71(6):e13–115. https://doi.org/10.1161/HYP.0000000000000065 PMID: 29133356
38. Derendorf H, Meibohm B. Modeling of pharmacokinetic/pharmacodynamic (PK/PD) relationships: concepts and perspectives. Pharmaceutical research. 1999.
39. Hempel G, Karlsson MO, de Alwis DP, Toublanc N, McNay J, Schaefer HG. Population pharmacokinetic-pharmacodynamic modeling of moxonidine using 24-hour ambulatory blood pressure measurements. Clin Pharmacol Ther. 1998;64(6):622–35. https://doi.org/10.1016/s0009-9236(98)90053-4