UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Concentration of CD3+/CD4+ cells in people with HIV at first presentation and their association with immune reconstitution

Sebastian Noe

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

2023
2024

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Concentration of CD3+/CD4+ cells in people with HIV at first presentation and their association with immune reconstitution

Sebastian Noe

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Contents

List of Figures

List of Tables

Abstract

Background

People living with HIV demonstrate relevant heterogeneity at the time of diagnosis. One major soure of this heterogeneity is imposed by (often hughly) different times between infection with HIV and diagnosis or referral to a specialized center. Since this time is, in turn, related to the risk of a compromised immune system, disease progression, and AIDS-defining condition. Classification into distinct groups might facilitate risk assessment for PWH at first presentation but currently available classification systems (including the definitions of late diagnosis and advanced HIV disease) are partly based on convenience classifications and/or have been derived in times where the course of HIV was markedly different from today, given the availability of hihgly effective antiretroviral treatment options.

Purpose

The purpose of this study was to account for baseline heterogeneity with regard to the CD4 cell concentration at first presentation of people with HIV to a specialised HIV care center and to explore their relevance and meaning in terms of clinical outcomes.

Methods

Retsopective, observational analysis of people living with HIV presenting with documented, viremic (HIV-1 RNA $\geq$ 200 copies/mL) HIV-1 infection at a single outpatient HIV clinic in Munich, Germany, between 2010 throuth 2020. Finite mixture models with component specific variances with both, two and three components, were fitted to the distribution of the first available CD4 cell concentrations. Models were compared using information criteria. Based on the posterior probabilities for each of the sub-populations, each person included into the study was assigned to one of the groups for each of the two models and sub-populations were compared through different models in order to explore potential meaningful differences. Participants were followed up longitudinally and a linear mixed-effects model was used to describe and compare the immune recovery for each of the subgroups.

Results

Overall, 1,452 PWH were included, resulting in a total of 31,334 observations for the longitudinal data analysis. Median age was 38 years (IQR: 30; 46), 300 (20.7 %) were female. 647 (44.6 %) and 305 (21.0 %) of the PWH included in this study had $CD3^+/CD4^+$ 350 cells/μl and 200 cells/μl, respectively. In 454 (31.3 %) of PWH an HIV-1 RNA concentration 100,000 copies/mL was found. A finite mixture models of normals with three components seemed to be best supported by the data, with the distributions being N(43, 27), N(352, 159), and N(644, 241), with component probabilities of p=0.07, p=0.68, and p=0.25. Characteristics of people being assigned to the three groups based on the posterior probabilities of the mixture model differed significantly with regard to age, HIV-RNA, CD4/CD8 ratio, prevalence of AIDS-defining conditions at first presentation, and the number of deaths until the end of the observation, where lower $CD3^+/CD4^+$ cells were associated with less favourable characteristics. Longitudinally, people in the groups of lower $CD3^+/CD4^+$ cells demonstrated a more pronounced, initial increase in $CD3^+/CD4^+$ cell concentratin after ART initiation, but nevertheless plateaued at a lower level when compared to people with higher $CD3^+/CD4^+$ cell concentrations at baseline.

Conclusion

A finite mixture model with three components and component specific-variances seems to identify distinct sub-populations of PWH at different 'stages' of the disease as a source of heterogeneity.

1 Introduction

1.1 Overview

HIV is an acronym for 'human immunodeficiency virus', a human pathogenic virus of the family of retroviruses. While often forgotten, there are two species, namely HIV-1 and HIV-2, with HIV-1 being markedly more prevalent throughout the world[1]. In the remainder of the manuscript, HIV will therefore refer to HIV-1.

A schematic illustration of an HI-virus can be seen in figure 1. Specific proteins are expressed on the surface of the virus' lipoprotein coating. These proteins play an important role in the replicative cycle of HIV by promoting attachment to and interaction with target cells, that are characterized by the presence of the cluster of differentiation 4 (CD4) on their surface[1]. The so-called capsid inside the virus contains two copies of single-strand ribonucleic acid (RNA), representing the virus' genome, as well as viral proteins, including reverse transcriptase (RT) and integrase (IN); together with the protease (PR) these are key enzymes in the viral replication cycle as they allow for the reverse transcription of the viral RNA into desoxy-ribonucleic acid (DNA), the integration of the proviral DNA into the human DNA, as well as the processing of protein precursors following transcription of viral genes[1].

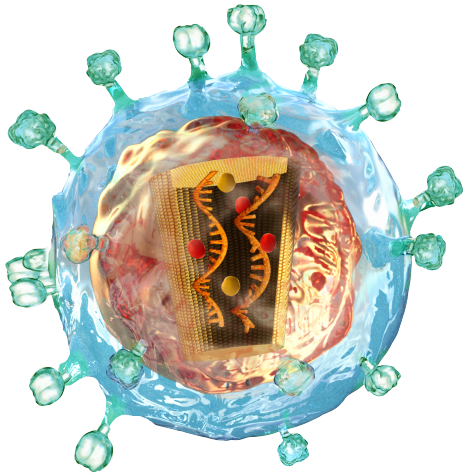


Figure 1: Schematic illustration of an HI-virus.

1.1.1 History of HIV

The beginning of the world-wide HIV pandemic dates back to the early 1980ies, when a clustering of 'unusual' diseases was found in (young) men who have sex with men (MSM), as well as people with hemophilia. These diseases included Pneumocystis jerovecii pneumonia (PCP), Kaposi sarcoma (KS), and chronic ulcerative herpes simplex virus (HSV) infections[1], and their occurrence in seemingly otherwise healthy people was poorly understood at first and virtually all diseased people died, making this 'new disease' a general death sentence. The possibility of a viral genesis was suspected early, but it was only several years later, that a virus which was later to be named HIV-1, was isolated from the blood of diseased people. But even after this important discovery, it took some time to develop effective antiretroviral treatment (ART) strategies. Most importantly, the high rates of errors during viral replication with a consecutive high frequency of mutations relevant to the efficacy of antiretroviral drugs (ARV) made is difficult to have treatments with long-lasting suppressive effects. Only with the combined use of several drugs, targeting several mechanisms in the replication cycle of HIV-1, sustained viral suppression was made possible, making HIV-1 a chronic and non-curable, yet well treatable disease[1].

1.1.2 Epidemiology

Since the beginning of the worldwide HIV epidemic, more than 80 million people have been infected with HIV and more than 40 million have died of HIV-related conditions or AIDS[2]. It was estimated for 2022, that more than 38 million people worldwide were living with HIV with more than 1 million new HIV diagnoses per year[2]. Still today, more than 500,000 people are dying per year from HIV-related causes[2]. While in some areas of the word, including many countries in Western Europe, the number of new HIV diagnoses is declining, globally there is no clear trend; in some parts of the world numbers of new HIV infections are raising. Globally, more women than men are living with HIV[2].

1.1.3 Pathogenesis

After entering the human body, HIV establishes a chronic infection by integrating proviral DNA into the human genome and using the human body's cells mechanisms for its own replication[1]. The virus can be contracted via HIV-containing blood or body fluids of humans. However, it needs an 'intrusion' of HIV-containing fluids with a sufficient number of viruses to serve as a potential source of infection. Everyday body contact, shared use of utensils, or glasses, for example, do not pose a risk even in people with detectable viral load[1]. Condom-less sexual contacts are still one of the most frequent routes of HIV-transmission, in the northern hemisphere particularly in MSM. Shared injection needles, for example among people using intravenous drug, are another potential source of transmission from one person to another. Vertical transmission (e.g. transmission from mother to child during or around birth) has become rare in many countries that have implemented HIV-testing as a routine in care of pregnant women, as early initiation of antiretroviral treatment with viral suppression at the time of delivery or cesarean section for mothers with insufficient or only short-time viral suppression at the time of delivery decrease the risk of transmission massively, being zero for mothers that have been undetectable for a sufficient time prior to delivery. Also, the rigorous testing of donated blood has led to a negligible risk of transmission via blood products[1]. The affinity of HIV to $CD4^+$ explains the particular effect on cells expressing this marker of differentiation on their surfaces. While this is the case for cells including macrophages, dendritic cells, and mikroglia, the characteristic effect of HIV on the immune system is mediated through interaction with a subset of the $CD3^+$ lymphocytes, so called TH cells (helper cells, $CD3^+/CD4^+$ cells). Binding to these cells in the presence of a suitable co-receptor enables the virus to intrude into these cells and via various mechanisms, lead to their diminution over time. Therefore, except for the rare person who can control viral replication on their own (so called "elite controllers"), the peripheral concentration of $CD3^+/CD4^+$ cells is a function of a person's baseline cell count (before infection with HIV), but also of the time living with the uncontrolled infection[1].

1.1.4 Symptoms and clinical presentation

HIV infection typically follows a certain sequence of different 'phases'. At the beginning, short time after contracting the HI-virus, the body's reaction is usually unspecific, including symptoms that can be found in many other acute (viral) infections. These include fever, chills, night sweats, headaches, rashes, and lymphadenopathy. If during this early phase of the disease HIV is not considered as a potential differential diagnosis, the disease often remains undiagnosed for a long time. This is problematic as the phase to follow, which can be several years in duration, most people are a- or only slightly symptomatic which means that

they will not be tested for HIV unless it happens for 'routine' checks, for example in groups of people who know they are or consider themselves to be at increased risk of HIV. As the immune system decreases over time, people might start to develop symptoms of 'opportunistic' diseases, e.g. diseases that develop more easily and/or more frequently in the presence of an altered immune system until in the last 'phase' of this sequence, life threatening opportunistic infections (OI) or malign hematologic or solid neoplasia develop[1]; more a more comprehensive overview will be given in the section on "Late diagnosis and AIDS" (Section 1.2).

1.1.5 Diagnosis

Diagnosis of HIV is straight forward in most cases. Most frequently, antibody tests are used, that can detect the human immune response to an HIV infection by demonstrating the existence of specific antibody in an index person's blood sample. More advanced antibody tests also directly test for HIV's p24 antigen. This approach allows an earlier diagnosis, as it does not require the formation of antibodies, which only occurs some weeks after contraction. Nevertheless, a diagnostic window remains. HIV antibody-tests are highly sensitive. Therefore, particularly when used in people with a low a priori probability of contracting HIV, false-positive results occur. As a consequence, a diagnosis can only be made in the presence of a positive 'confirmation' test. Traditionally, Western blots are used, testing a second blood sample (in order to also exclude a patient mix-up). Today, often direct measurement of viral RNA after polymerase chain reaction (PCR) techniques for amplification are used, which allow not only for rapid testing, but also for a further shortening of the diagnostic window as well as a quantitative determination of the RNA content of a blood sample. However, since primers for HIV RNA measurements are highly specific, an HIV-2 infection will not be detected by running the routine HIV-1 RNA analyses, while antibody tests can not only detect, but also distinguish between both groups of HI viruses. Therefore, as well as for financial reasons, HIV PCR has not become the routine screening test for HIV, except for some distinct situations[1].

1.2 Late diagnosis and AIDS

As the decrease in cellular immunity in PWH is in general a function of time, later diagnosis is usually associated with a more compromised immune system, which is relevant both, clinically, but also epidemiologically, having implications for case detection.

1.2.1 Late diagnosis

Late diagnosis (which was formerly called 'late presentation') is a term used to describe PWH being diagnosed in a stage of an already markedly altered immune system. The definition of late diagnosis was a matter of scientific debate for quite a long time, until a consensus statement defined late diagnosis as the diagnosis of HIV at a $CD3^+/CD4^+$ cell concentration < 350 /µL and/or in the presence of an AIDS-defining disease[3,4]. The motivation for the threshold of 350 $CD3^+/CD4^+$ cells/µL remains, however, poorly understood and might represent a compromise of different observations, and the consensus statement has been criticized with regard to its biomedical significance and relevance[5]. Yet, several studies indicate that persons being diagnosed with HIV 'late' or in whom treatment is initiated late, are at higher risk of adverse events and death and late diagnosis is considered a major problem in care of PWH[6,7]. Clinically and epidemiologically,

distinguishing a group with late compared to early diagnosis therefore seems justified, as they might present with different baseline risk for disease progression and comorbidity, but also represent distinct populations of people.

1.2.2 AIDS

AIDS is an acronym for 'acquired immunodeficiency syndrome' and marks the most advanced stage of an HIV-infection. This stage is defined by the occurrence of certain, so called 'AIDS-defining' diseases, that are mostly opportunistic infections, malign tumors, or hematologic neoplasia[1]. A non-exhaustive overview is given in table 1. The occurrence of AIDS-defining diseases indicates the presence of a severe immune deficiency. While for some of these diseases specific treatments are available, in others only HIV therapy itself is effective; as usual, prevention is in general the best option. With the availability of effective antiretroviral therapies, the progression of HIV to AIDS can be avoided. This implies, that nowadays, most PWH will not experience AIDS, where access to ART is granted, and that HIV and AIDS cannot be used interchangeably.

Table 1: Non-exhaustive list of AIDS-defining diseases (adapted from Hoffmann and Rockstroh).

Candida infection	lower airways, lung, esophagus
CMV infection	with exception of liver, spleen, lymph nodes
CMV retinitis	with loss of visus
Encepaholopathy	HIV-associated
Kaposi Sarkoma	
Lymphoma	Burkitt, immunoblastic, primary cerebral
Mycobacteria infection	M. tuberculosis, M. avium complex, M. kansasii
Pneumocystis jeroveci pneumonia	
Progressive multifocal leukencepaholopathy	
Toxoplasmosis	cerebral
Wasting Syndrome	

1.3 Antiretroviral therapy

The identification of HIV was the basis for the development of antiretroviral therapies. These were and are drugs that interfere with the virus' replication cycle at different stages. Until today, more than one drug with at least two targets to suppress replication is necessary to avoid the development of resistance under the selection pressure and the high rate of mutations occurring during HIV replication. Some of the major targets of modern antiretroviral drugs in the replicative cycle of HIV are depicted in Figure 2.

With the development of effective antiretroviral therapies within years and decades after HIV's discovery, it also became clear, that adequate treatment with sustained virologic suppression can prevent the development of severe comorbidities caused by HIV. Today, life expectancy has improved markedly and is approaching the life expectancy found in a general population[8–11]. HIV has become a chronic but in general well-treatable infectious disease. The paradigm of antiretroviral treatment has undergone considerable change over time: when first available, antiretroviral drugs were often badly tolerated, they had to be taken several times per day, and they did not work for very long: the development of resistance was a major problem. Therefore, PWH were often not treated before their CD4+/CD3+ cell concentration fell under a certain threshold.

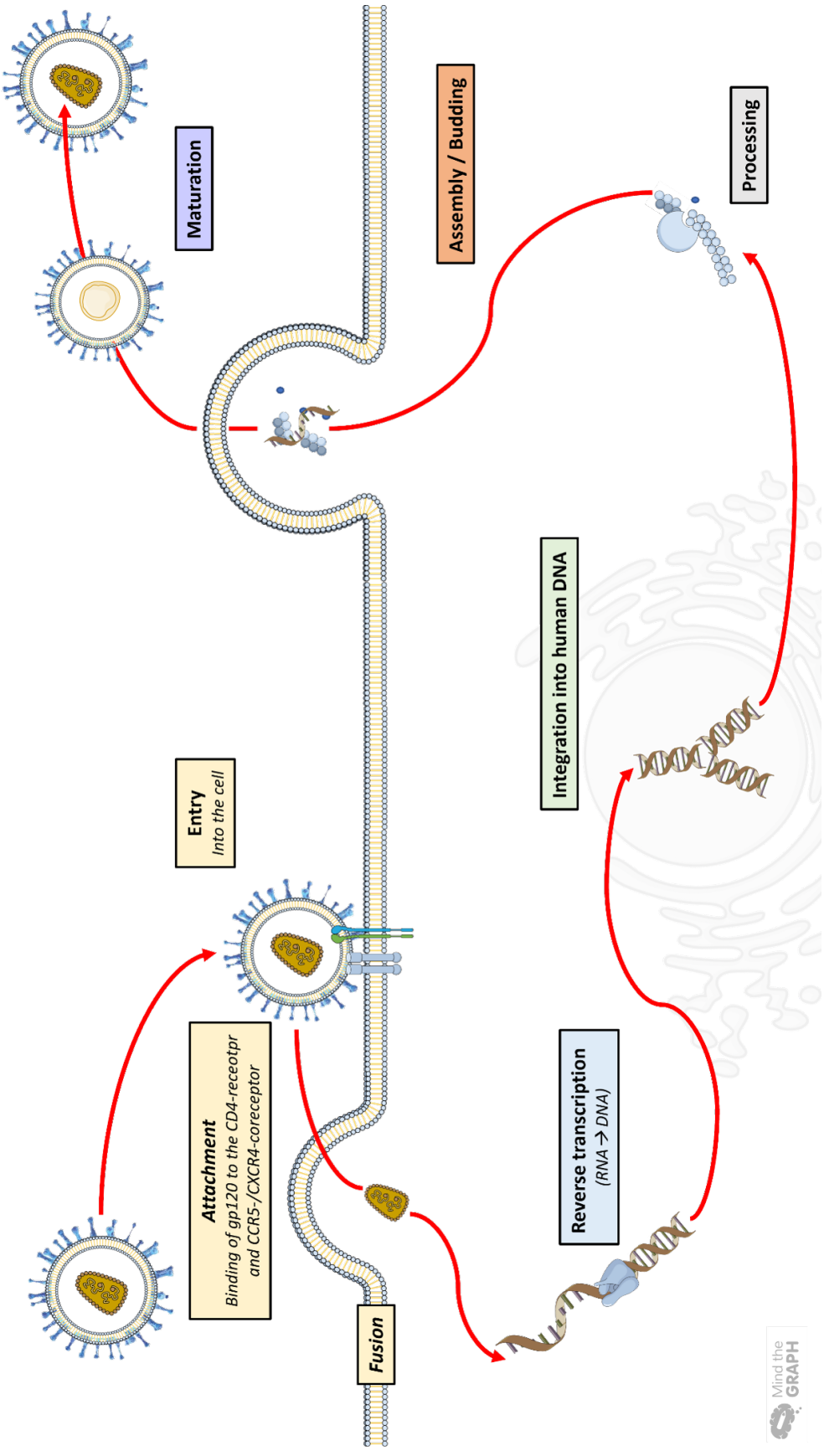


Figure 2: Schematic illustration of the replicative cycle of HIV.

It was only with the publication of the so-called START study[12], that major guidelines changed their recommendation. The START study demonstrated, that later treatment initiation at lower $CD3^+/CD4^+$ cell concentrations was associated with adverse outcomes, including a higher probability of developing AIDS-defining disease[12]. Consequently, in today's guidelines on antiretroviral therapy in PWH do not see a role in deferred treatment initiation anymore, and treatment should be started in every person that wishes to receive ART regardless of $CD3^+/CD4^+$ concentrations[13].

1.4 Purpose of the study

As described before, PWH presenting to HCPs after first diagnosis comprise a heterogeneous group of people, in particular with regard to their immune status at that time. Time from infection to diagnosis or first presentation in a specialized clinic is assumed to highly contribute to this heterogeneity. Given current literature, it seems to be justified to classify PWH into distinct groups according to their $CD4^+/CD3^+$ cells at this time, as lower $CD3^+/CD4^+$ cells, assumedly following longer time of living with HIV, are associated with worse HIV-related and overall outcomes. While both, general classification systems (such as CDC) as well as specific classification criteria for late diagnosis exist, they are sometimes based on convenience cut-offs and developed more than a decade ago, in which marked changes in prevention, diagnosis and treatment of HIV have taken place.

The purpose of this study is to investigate the distribution of $CD3^+/CD4^+$ cell concentrations at the time of first presentation of PWH with viremia ($\geq$ 200 copies/mL) at a specialized HIV clinical care center in Munich, Germany, by means of finite mixture models (FMM) of normal distributions, assuming at least two components. Components will be considered to represent a latent class of time living with HIV. The classification derived from the FMM will be used to assign each PWH in the study to one of the groups and groups will be compared with regard to the presence of AIDS-defining disease at first presentation, but also followed up longitudinally in order to explore differences in immune reconstitution over time between the groups. This second part of the study will therefore explore, whether the latent group classification derived from the FMM is clinically plausible and relevant.

2 Data

2.1 Data collection

Data was collected from electronic patient files of PWH in a single, large outpatient HIV research and clinical care center in Munich, Germany (MVZ München am Goetheplatz). All PWH first presenting to the clinic between January 1st 2010 and December 31st 2020 were considered for inclusion into the analysis; all observations within this time interval were used for longitudinal analyses. Data were extracted from medical records applying a software solution (cvSentinel, Clinovate NET, Munich, Germany) with interfaces adapted to the site-specific electronic patient management system, creating a .csv output file. Presence of AIDS-defining conditions at first presentation and the country of origin were identified by individual electronic patient file review.

2.1.1 Inclusion criteria

- Documented HIV-1 infection
- Age $\geq$ 16 years
- HIV-1 RNA $\geq$ 200 copies/mL at first presentation

2.1.2 Exclusion criteria

- Documented HIV-2 infection
- Age $<$ 16 years
- HIV-1 RNA $<$ 200 copies/mL at first presentation
- No measurement of $CD3^+/CD4^+$ cells available

2.2 Data dictionary

A list of variables that were used for analysis in this study can be found in table 2.

Table 2: Data dictionary for all relevant variables used in this study.

Name	Type	Description	Values
Age	integer	Age of participant at first presentation	16-82
AIDS	integer	Presence of AIDS-defining conditions at first presentation 0=not present; 1=present	0, 1
CD4 cells	integer	Concentration of $CD3^+/CD4^+$ cells [in cells/µL and %]	0-3,031 and 0-65
CD8 cells	integer	Concentration of $CD3^+/CD8^+$ cells [in cells/µL and %]	5-10,379 and 5-92
Country of origin	string	Country of origin / birth	Details in Appendix A.1
Risk of transmission	integer	Suspected risk of HIV transmission 1=MSM, 2-9=others	1-9
Sex	integer	Sex (as per legal status) 0=male; 1=female	0, 1

* MSM: Men who have sex with men

Missing values are represented by blank cell-content in the original data-frame and by NA after data importation.

2.2.1 Exploratory data analysis

Visual data exploration was performed for the distribution of $CD3^+/CD4^+$ cells at baseline as well as the longitudinal follow up using scatter plots. For the longitudinal data, lowess-smoothing was added to the plots in order to explore the marginal average evolution over time. Individual longitudinal profiles were used as an additional component of exploratory data analysis in a small, random subset of PWH.

The distribution of $CD3^+/CD4^+$ cells is depicted in figure 3.

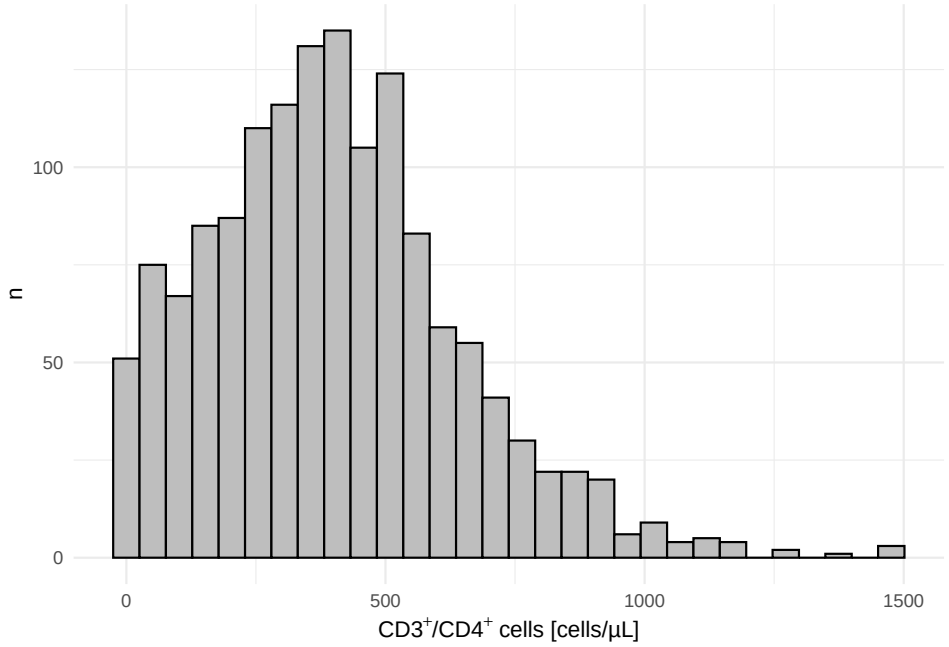


Figure 3: Distribution of $CD3^+/CD4^+$ cells at first presentation

The distribution is right skewed and non-symmetric. It is left-bound by 0 and the maximum observed value was 1478 cells/µL. The mean of the distribution is 400.9 cells/µL with a standard deviation of 241.8 cells/µL.

A scatter plot of $CD3^+/CD4^+$ cell concentration versus the time after baseline for the entire study sample is displayed in figure 4.

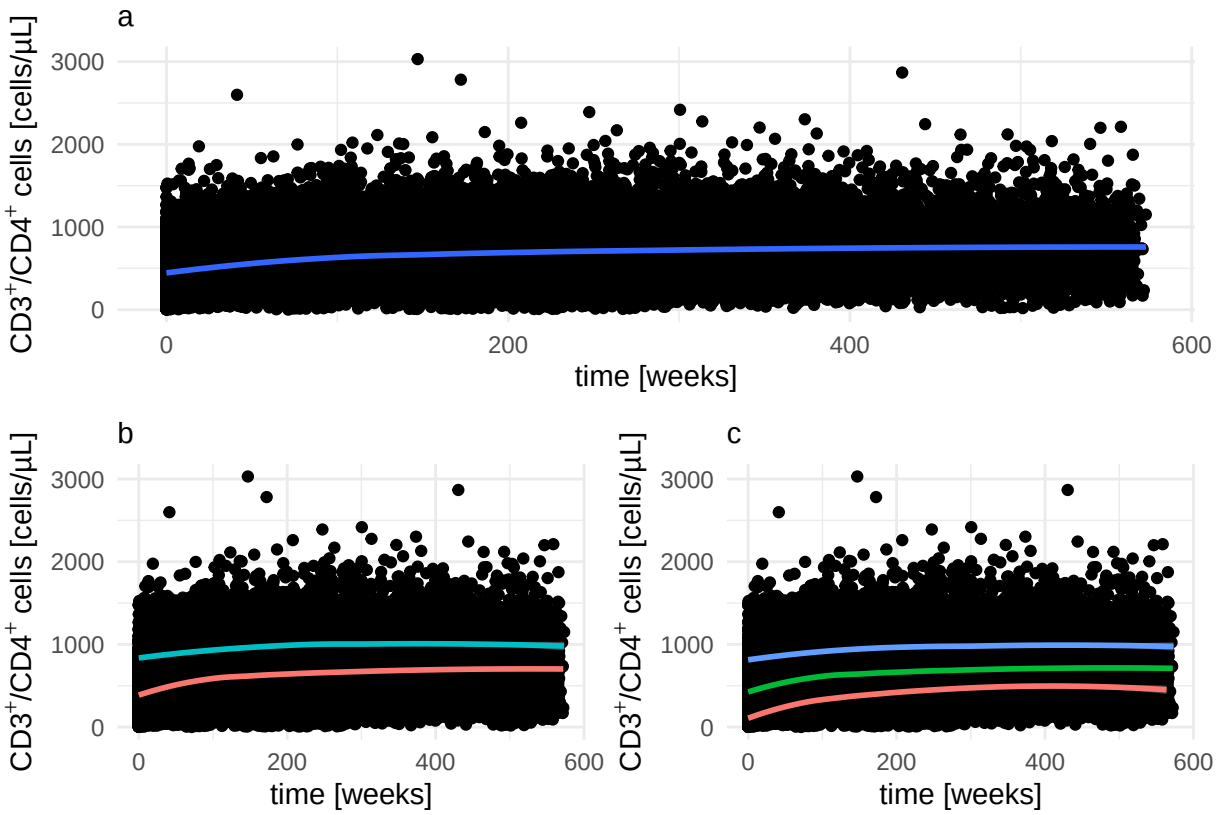


Figure 4: Scatter plots for the pairs of observation of $CD3^+/CD4^+$ and time t together with loess smothing for the entire study sample (a), as well as for the subgroups from the two-(b) and three(c)-component mixture models.

3 Ethics

The use of pseudonymized, monocentric clinical routine data does not require the approval of an ethical committee according to German law. Pseudonymization ensures that patients' identities are protected by replacing direct identifiers with codes, making it difficult to trace back to the original identity. In this study, the identifiers generated by the clinic patient management system were used, so that only people with access to this system were able to potentially trace back identities to indivdual patients. This significantly contributes to the protection of patients' privacy and personal data. Furthermore, all data processing was performed "on site", meaning that no data transfer outside of the clinic was made necessary. The study was conducted in accordance with the Declaration of Helsinki.

In addition to the Declaration of Helsinki, the General Data Protection Regulation (GDPR) of the European Union was adhered to, which imposes strict requirements on the processing of personal data. Compliance with the GDPR ensures that all data protection requirements were met, particularly regarding the rights of the data subjects and the security of the processed data. Adherence to these ethical and legal frameworks ensured that the study was conducted ethically and scientifically sound.

4 Methodology

4.1 Descriptive data analysis

For cross-sectional descriptive data analysis, medians together with the 25th and 75th quantile (IQR) were used to indicated central tendency and spread of the data. For categorical variables, absolute and relative frequencies (%) were calculated. Where applicable, comparison of sub-groups will be performed using Mann-Whitney or Kruskal-Wallis-tests for for continuous variable with two and three groups, respectively. For the comparison of categorical variables between different groups, χ^2 tests were used.

4.2 Finite mixture models

Based on current knowledge, finite mixture models (FMM) with k $\in$ {2,3} components will be fit, allowing for component-specific variances (σ^2):

$$Y|(\mu, \sigma^2) \sim N(\mu, \sigma^2)$$

$$(\mu, \sigma^2) \sim \begin{pmatrix} \mu_1, \sigma_1^2 & \dots & \mu_1, \sigma_k^2 \\ \pi_1 & \dots & \pi_k \end{pmatrix}$$

where π_i is the probability of component k_i.

Assuming that ψ is a vector that contains the parameters of the finite mixture model,

$$\vec{\psi} = (\vec{\pi}, \vec{\theta}),$$

where θ is the vector of parameters describing the densities for the mixture components, in the case of a normal distribution the parameters μ and σ^2.

The expectation-maximization (EM) algorithm will be used to estimate the finite mixture model parameters by maximizing the expectation of the complete data log-likelihood

$$E[\ell(\psi|y, Z)|y]$$

The EM algorithm iteratively estimates a series of $\psi^{(t)}$ in two steps, leading to the convergence to the maximum likelihood estimate of ψ. Starting from a value t, first

$$Q(\psi|\psi^{(t)}) = E[\ell(\psi|y, Z)|y, \psi^{(t)}]$$

is calculated (E-step), which is afterwards maximized with respect to ψ (M-step), leading to an updated estimate $\psi^{(t+1)}$. The procedure is repeated until the difference between ℓ_t and ℓ_{t+1} falls bellow a pre-defined threshold ε (with ε>0).

In order to get an initial impression about whether or not this assumption is compatible with the data, an approach suggested by Schlattman will be used[14]: Non-parametric bootstrap samples will be obtained

from the cross-sectional data set by sampling with replacement. The number of support-points for each of these samples ($\hat{k}$) will be estimated using the vertex exchange method (VEM). The VEM algorithm aims to identify the mixing distribution G that fits the data best by finding $\hat{G}$ for which

$$\ell(\hat{G}) = max(\ell(G))$$

for $G \in \Gamma$, where Γ stands for the class of the distributions. While in general, the estimate for $\hat{G}$ derived from the VEM needs further scrutiny in order to make sure the obtained estimated is a true non-parametric maximum-likelihood estimate (NPMLE) and unique, this was not done for each of the bootstrap replicates for $\hat{G}$ due to the exploratory character of this approach. Instead, the estimated number of components k was registered for each VEM replicate and the distribution of $\hat{k}$ through the replicates was be explored with regard to the frequency of $\hat{k}$=2 and $\hat{k}$=3.

For a direct model comparison for the cases of $k \in 2, 3$, the fit to the data will be determined using the log likelihood obtained under k components, and the two models will be compared using log likelihood but also Akaike Information Criteria (AIC)

$$AIC = 2 \cdot k - 2 \cdot ln(\hat{L})$$

and Bayesian Information Criteria (BIC)

$$BIC = k \cdot ln(n) - 2 \cdot ln(\hat{L})$$

4.3 Longitudinal data analysis

The model for the marginal average evolution was based on the visual exploration of the longitudinal data (overall as well as component-specific) in conjunction with the exploration of a random subset of individual longitudinal profiles. A linear mixed-model fo the following form was be fitted:

$$Y_i = X_i\vec{\beta} + Z_i\vec{b_i} + \epsilon_i$$

$$b_i \sim N(0, D) \quad \epsilon_i \sim N(0, \Sigma_i)$$

where $\vec{\beta}$ and $\vec{b_i}$ describe the fixed- and random-effects, respectively, with D and Σ containing the variance components.
Furthermore, $b_1, ...b_N, \epsilon_1, ..., \epsilon_N$ are assumed to be independent.

Maximum likelihood (ML) will be used for parameter estimation, allowing for likelihood-ratio test inference on model selection. A backward selection will be used to fit the model to the data. The parameter to be excluded will be chosen based on the largest p-value. Models will be compared after each step, using likelihood ratio tests (LRT) of the form

$$\lambda = -2 \cdot (\ell_{p-1} - \ell_p),$$

where ℓ represents the log-likelihood of the models with p and p-1 parameters, respectively. For fixed effects, hypotheses will be tested assuming

$$\lambda \sim \chi^2_1$$

for the reduction of one fixed effect at a time, while for random effects a mixture of χ^2_{p-1} and χ^2_p

$$\lambda \sim \chi^2_{(p-1):p}$$

will be assumed.

Given the unbalanced data set with highly variable times between observations, an unstructured variance-covariance matrix will be assumed without further modification.

Assessment of the model fit will be based on visual exploration of the residuals and, if necessary, be re-fit after Box-Cox transformation of the dependent variable. Box-Cox transformation was performed using transformations

$$Y_t^{(\lambda)} = \begin{cases} \frac{Y_t^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ log_e(Y_t) & \lambda = 0 \end{cases}$$

with

$$\lambda = -2 + 0.5 \cdot k$$

for $k \in Z$ and $0 \leq k \leq 8$.

4.4 Conventions

Where needed, seeds were set using '20232024'.
A level of significance of α=0.05 was used.

5 Results

Overall, 1452 PWH were included, resulting in a total of 31334 observations for the longitudinal data analysis. Characteristics of the study sample can be found in table 3.

Table 3: Characteristics of the study sample at initial presentation at the study site.

Variable	Results		missing
Age [years], median (IQR)	38.0	(30.0; 46.0)	0 (0.0%)
Sex [female], n (%)	300	(20.7)	0 (0.0)
Route of transmission [MSM], n (%)	537	(44.9)	255 (17.6)
$CD3^+/CD4^+$ [cells/µL], median (IQR)	382.0	(229.0; 536.0)	0 (0.0%)
$CD3^+/CD4^+$ [%], median (IQR)	21.0	(14.0; 28.0)	0 (0.0%)
$CD3^+/CD8^+$ [cells/µL], median (IQR)	382.0	(229.0; 536.0)	0 (0.0%)
$CD3^+/CD8^+$ [%], median (IQR)	21.0	(14.0; 28.0)	0 (0.0%)
CD4/CD8 [ratio], median (IQR)	0.4	(0.2; 0.6)	0 (0.0%)
HIV RNA [copies/mL], median (IQR)	38736.5	(8917.2; 150808.2)	0 (0.0%)
AIDS defining condition [present], n (%)	127	(9.0)	36 (2.5)
Status at data cut [dead], n (%)	44	(3.0)	0 (0.0)

At the time of first presentation, 647 (44.6 %) and 305 (21.0 %) of the PWH included in this study had $CD3^+/CD4^+$ cells <350 cells/µL and <200 cells/µL, respectively. In 454 (31.3 %) of PWH an HIV-1 RNA concentration > 100,000 copies/mL was found.

5.1 Distribution of $CD3^+/CD4^+$ cells at baseline

The distribution of the concentration of $CD3^+/CD4^+$ cells at first presentation of PWH in the study sample is displayed in figure 3.

Repeatedly applying the VEM algorithm in order to explore the distribution of the numbers of components identified in non-parametric bootstrap samples as suggested by Schlattmann[14] identified a mode for $\hat{k}$=3 (49%), followed by $\hat{k}$=4 (13%).

Fitting a two-component finite mixture model of normal distributions to the data results in the following mixing distribution:

$$Y \sim \begin{pmatrix} \mu_1 & \mu_2 \\ \pi_1 & \pi_2 \end{pmatrix} = \begin{pmatrix} 328 & 628 \\ 0.76 & 0.24 \end{pmatrix}$$

with $\sigma_1 = 182$ and $\sigma_2 = 264$.

For a three-component finite mixture model of normal distributions, the following mixing distribution is obtained:

$$Y \sim \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ \pi_1 & \pi_2 & pi_3 \end{pmatrix} = \begin{pmatrix} 43 & 352 & 644 \\ 0.07 & 0.68 & 0.25 \end{pmatrix}$$

with with $\sigma_1 = 27$, $\sigma_2 = 159$, and $\sigma_3 = 241$. Information criteria for both finite mixture models can be found in table 4.

Table 4: Information criteria for the two- and three-component finite mixture model

	k=2	k=3	d
Likelihood	-9,981	-9,935	-46
Deviance	19,962	19,870	92
AIC	19,965	19,877	88
BIC	19,976	19,893	83

The distribution of the components in the two- and three-component mixtures can be found in figure 5.

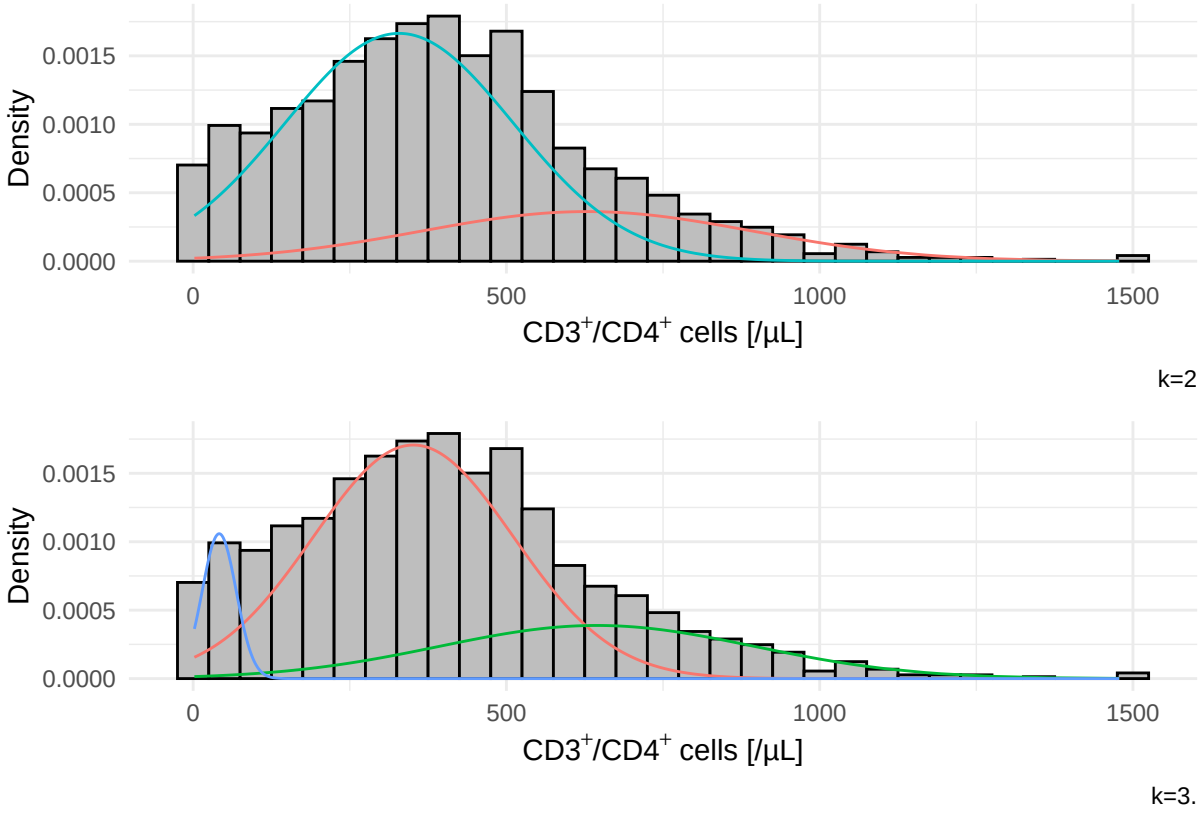


Figure 5: Mixing distribution for the two- and three-component final mixture models.

5.2 Longitudinal data analysis

Exploration of the marginal average evolution over time seems to justify a linear regression with different slopes for the time before and after about 100 weeks; for the sake of practicality, the cut-off for the different slopes will be assumed to be at 104 weeks, corresponding to two years of follow up.
Exploration of the individual profiles of a subset of randomly chosen participants with an observation time of at least 104 weeks (2 years) demonstrates high intra- and inter-individual variability (figure 7), justifying

Table 5: Comparison of characteristics between PWH assigned to each of the components of the two-component finite mixture model.

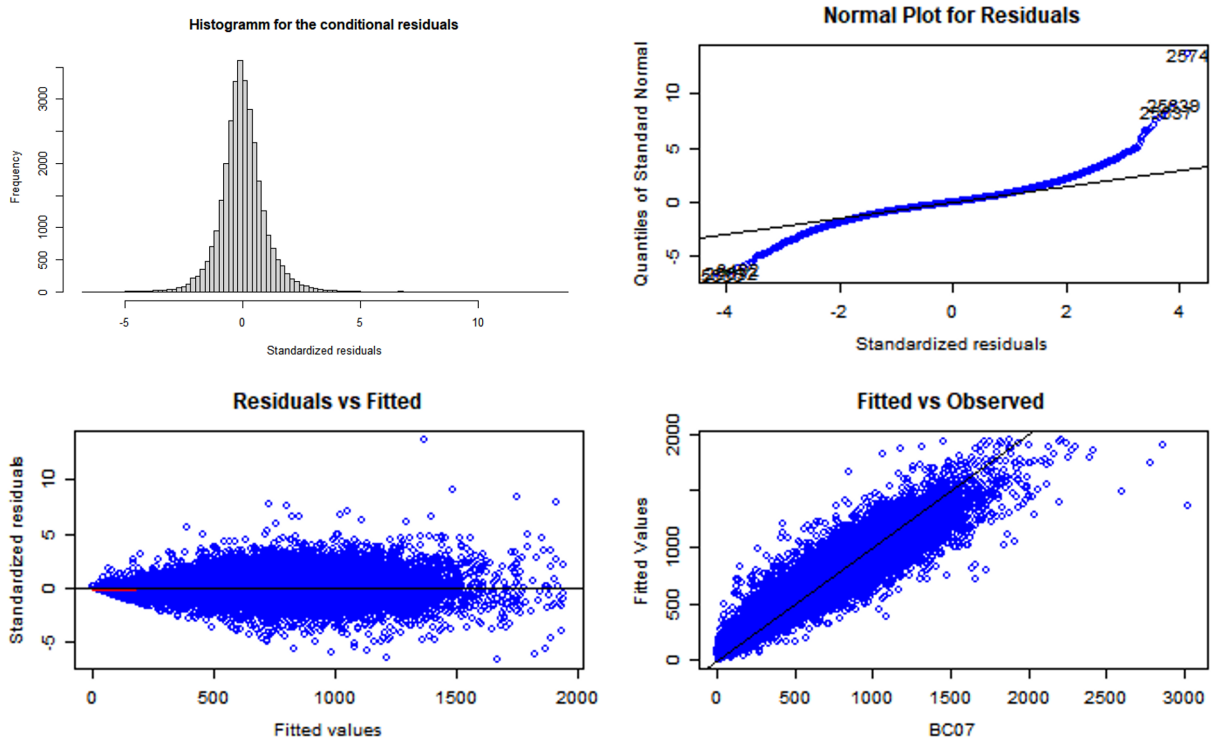
	k=1	k=2	p-value
Age [years], median (IQR)	38.0 (31.0; 46.0)	35.0 (29.0; 45.0)	0.012
missing, n (%)	0 (0.0%)	0 (0.0%)	
Sex [female], n (%)	979 (78.9)	173 (82.0)	0.349
missing, n (%)	0 (0.0)	0 (0.0)	
MSM [yes], n (%)	449 (43.8)	88 (50.9)	0.102
missing, n (%)	217 (17.5)	38 (18.0)	
Region of origin [Subsahran Africa], n (%)	203 (19.0)	25 (13.6)	0.191
missing, n (%)	171 (13.8)	27 (12.8)	
$CD3^+/CD4^+$ [cells/µL], median (IQR)	340.0 (204.0; 465.0)	776.0 (699.0; 894.0)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	
$CD3^+/CD4^+$ [%], median (IQR)	19.0 (13.0; 26.0)	32.0 (26.0; 38.0)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	
CD4/CD8 [ratio], median (IQR)	0.3 (0.2; 0.5)	0.7 (0.5; 1.0)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	
HIV RNA [copies/mL], median (IQR)	47324.0 (11277.0; 173375.0)	11227.0 (2110.0; 42277.5)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	
AIDS defining condition [yes], n (%)	123 (10.1)	4 (2.0)	0.008
missing, n (%)	29 (2.3)	7 (3.3)	
Status at data cut [death], n (%)	42 (3.4)	2 (0.9)	0.091
missing, n (%)	0 (0.0)	0 (0.0)	

random intercepts as well as random time effects. Furthermore, addition of a quadratic time effect might be reasonable.

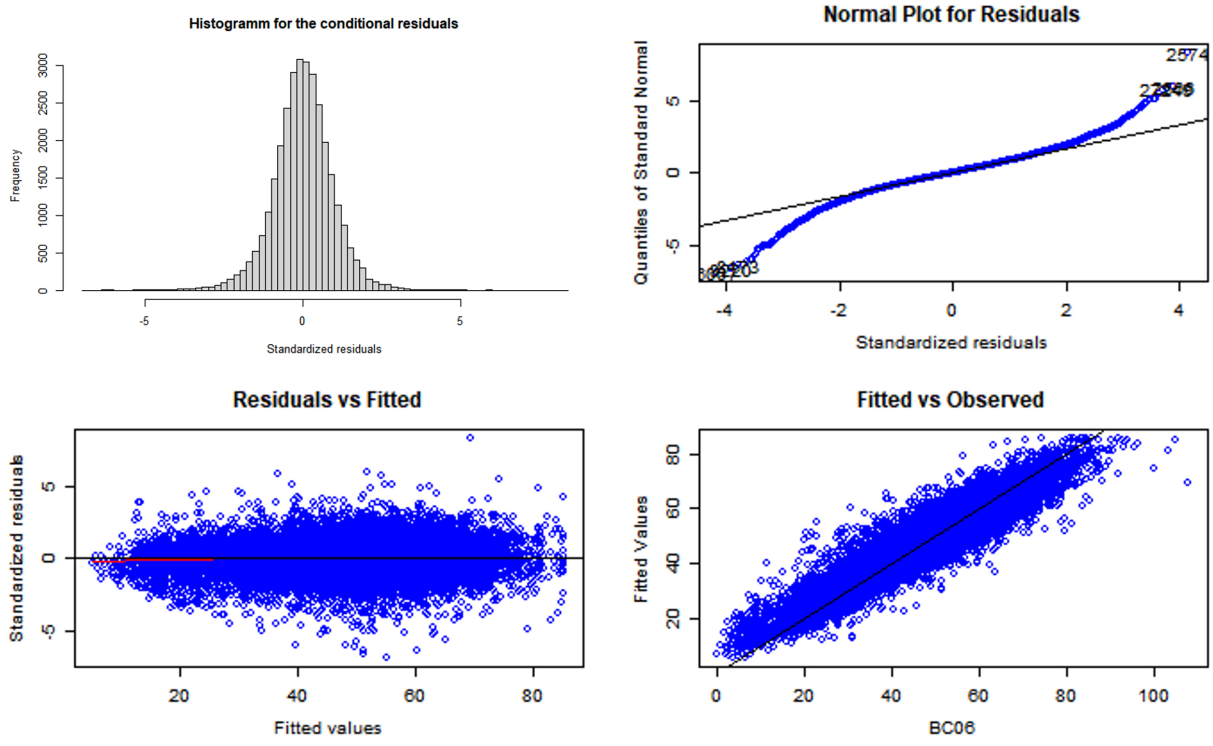
Based on the results of the exploratory data analysis, the most complex model considered for the description of the development of $CD3^+/CD4^+$ cells over time will be:

$$\begin{aligned} Y_{it} = \quad & (\beta_0 + I_2 \cdot \beta_{02} + I_3 \cdot \beta_{03} + b_{0i}) + \\ & (\beta_1 + \beta_{1q} \cdot t_1 + I_2 \cdot (\beta_{12} + \beta_{12q} \cdot t_1) + I_3 \cdot (\beta_{13} + \beta_{13q} \cdot t_1) + b_{1i}) \cdot t_1 + \\ & (\beta_2 + \beta_{2q} \cdot t_2 + I_2 \cdot (\beta_{22} + \beta_{22q} \cdot t_2) + I_3 \cdot (\beta_{23} + \beta_{23q} \cdot t_2) + b_{2i}) \cdot t_2 + \varepsilon_{it} \end{aligned}$$

where β_0,β_1,β_2 indicate common time effects throughout groups, while β_{02},β_{03} are group specific intercepts,



(a) Original scale



(b) Transformed scale

Figure 6: Plots for the model diagnostic on the original (a) as well as the Box-Cox transformed scale (b) of $CD3^+/CD4^+$ cell concentration for the two-component finite mixture model.

Table 6: (#tab:tab:Comparisonk3Model)Comparison of characteristics between PWH assigned to each of the components of the three-component finite mixture model.

	k=1	k=2	k=3	p-value
Age [years], median (IQR)	41.0 (34.0; 48.0)	38.0 (30.0; 46.0)	35.5 (29.0; 45.0)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Sex [female], n (%)	108 (81.8)	854 (78.5)	190 (81.9)	0.387
missing, n (%)	0 (0.0)	0 (0.0)	0 (0.0)	
MSM [yes], n (%)	45 (40.5)	395 (44.1)	97 (50.8)	0.154
missing, n (%)	21 (15.9)	193 (17.7)	41 (17.7)	
Region of origin [Subsahran Africa], n (%)	13 (11.1)	186 (20.0)	29 (14.1)	0.105
missing, n (%)	15 (11.4)	156 (14.3)	27 (11.6)	
CD3$^+$/CD4$^+$ [cells/µL], median (IQR)	37.0 (16.0; 60.2)	361.0 (250.0; 477.2)	757.5 (682.0; 878.2)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
CD3$^+$/CD4$^+$ [%], median (IQR)	4.0 (2.0; 6.0)	20.0 (15.0; 26.0)	32.0 (26.0; 38.0)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
CD4/CD8 [ratio], median (IQR)	0.1 (0.0; 0.1)	0.4 (0.2; 0.6)	0.7 (0.5; 1.0)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
HIV RNA [copies/mL], median (IQR)	271442.5 (85336.8; 705077.5)	40235.5 (10516.8; 124615.0)	11766.0 (2224.2; 45394.5)	<0.001
missing, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
AIDS defining condition [yes], n (%)	56 (42.4)	67 (6.3)	4 (1.8)	<0.001
missing, n (%)	0 (0.0)	27 (2.5)	9 (3.9)	
Status at data cut [death], n (%)	9 (6.8)	33 (3.0)	2 (0.9)	0.006
missing, n (%)	0 (0.0)	0 (0.0)	0 (0.0)	

and β_{12},β_{13},β_{22},β_{23} are group specific slopes with I_2,I_3 being the indicator for the groups k_2,k_3, respectively. For the error term the assumption is

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\varepsilon_i})$$

Index q indicated the coefficients for the quadratic time effects. b_{0i}, b_{1i}, and b_{2i} represent, subject specific random effects, for which we assume:

$$b_{xi} \sim \mathcal{N}(0, \sigma^2_{b_x})$$

For t_1, t_2 the following definitions hold:

$$t_1 = \begin{cases} t & t < 104 \\ 104 & t \geq 104 \end{cases}$$

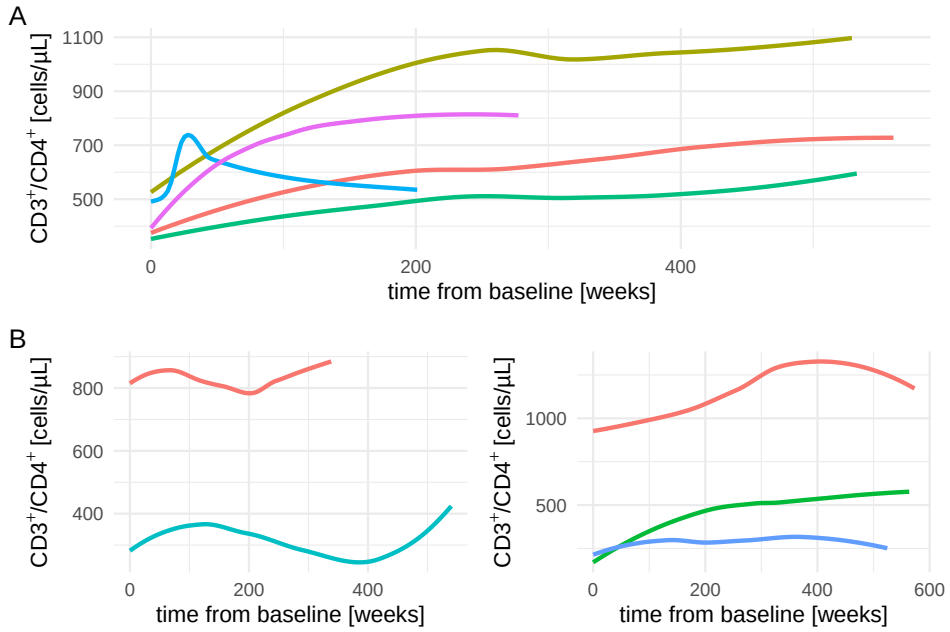


Figure 7: Individual, longitudinal profiles of five randomly chosen study participants (A), and one participant per group for the two-(B) and three-component model (C), respectively, with a minimum observation time of 104 weeks.

$$t_2 = \begin{cases} 0 & t < 104 \\ t - 104 & t \geq 104 \end{cases}$$

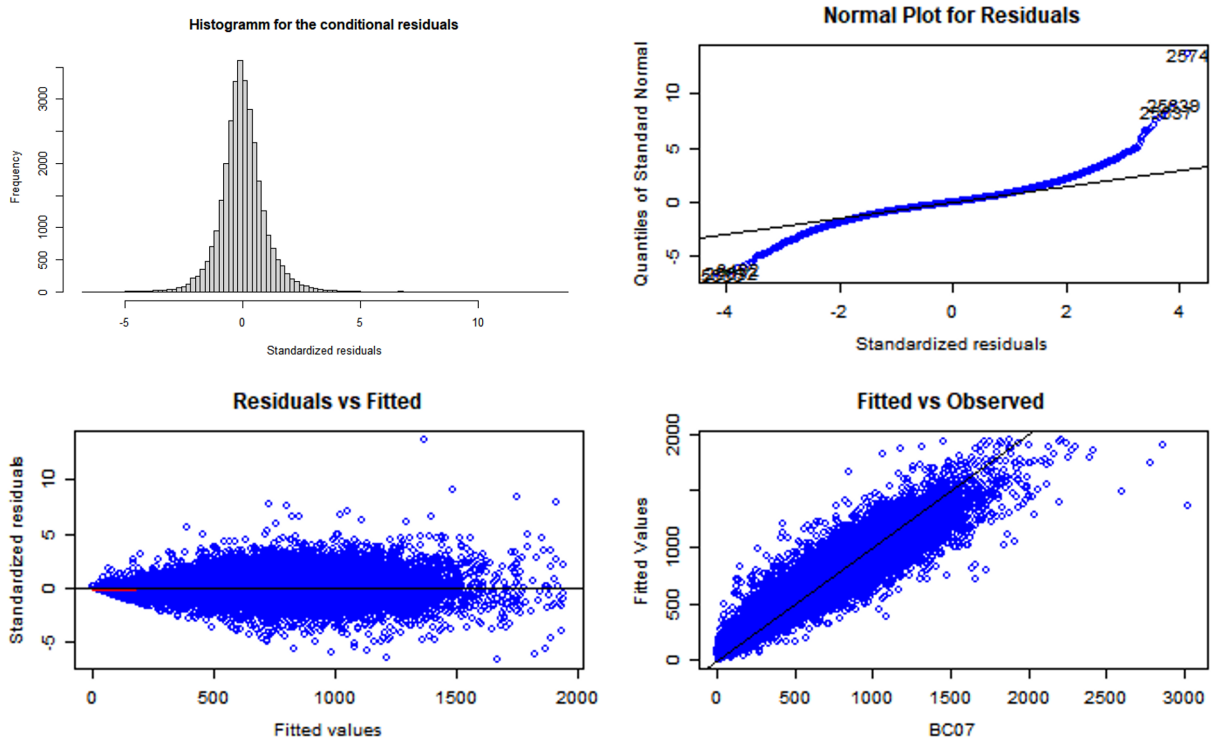
5.2.1 Model-fitting for k=2 components

First, the possibility of reducing the random effects structure of the model was tested by removing one random effect at a time and using likelihood ratio tests for variance components to compare the reduced models against the full random effects model. As for each of the random effects the removal resulted in a significant change of the model fit ($p < 0.001$ for all), the random effects structure remained unchanged. The steps of reduction of the fixed effect structure using a backward selection departing from the initial model can be found in table 7.

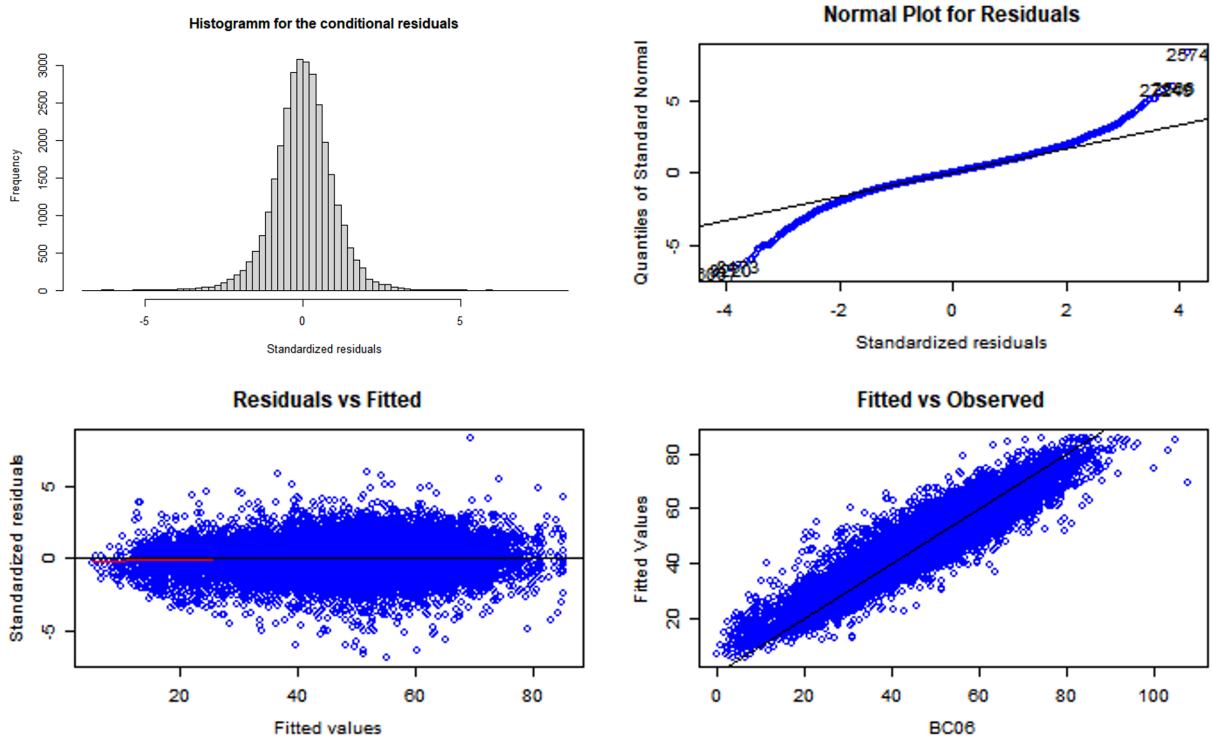
5.2.1.1 Model diagnostics and remedial measures Plots for the diagnostic of the model derived in the previous section can be found in figure 8a.

In the lowest and particularly in the highest ranges of the standardized residuals, a relevant deviation from the standard normal distribution can be found. Also, some heteroscedacicity might be assumed when exploring the standardized residuals over the range of fitted values. Among the series of Box-Cox transformations, a transformation using λ=0.5 seemed to reduce the deviation of the residuals from the normal assumptions best (figure 8).

A comparison between the observed against the predicted mean average evolution of the $CD3^+/CD4^+$ over time is given in figure 9. For both models, before and after transformation, the agreement between the observed and predicted concentrations of $CD3^+/CD4^+$ seem to be comparable.

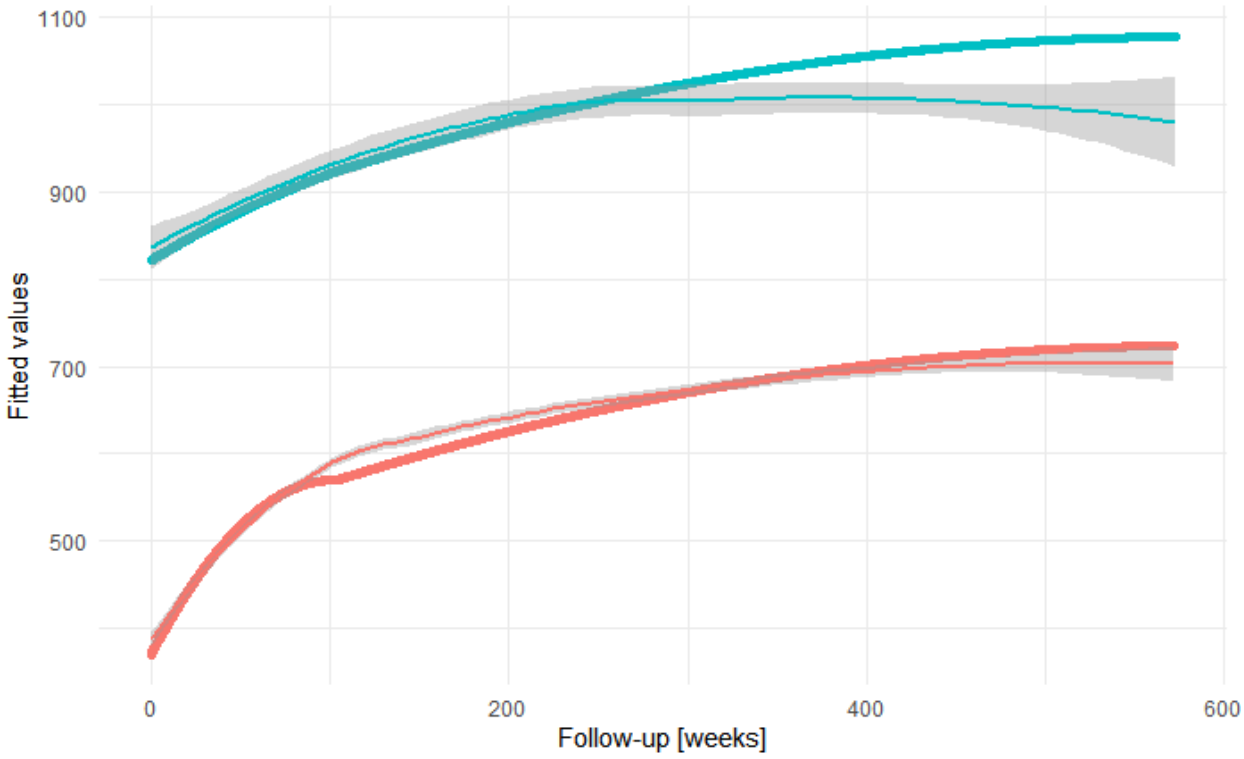


(a) Original scale

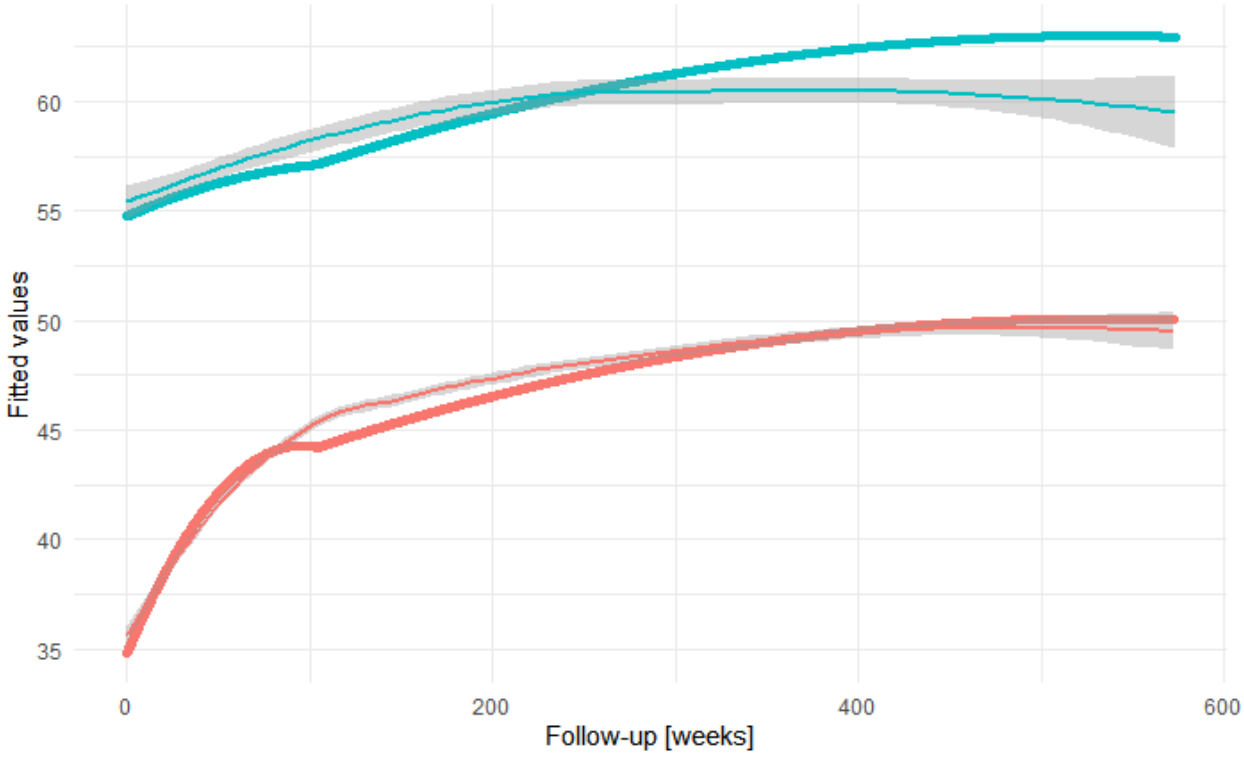


(b) Transformed scale

Figure 8: Plots for the model diagnostic on the original (a) as well as the Box-Cox transformed scale (b) of $CD3^+/CD4^+$ cell concentration for the two-component finite mixture model.



(a) Original scale



(b) Transformed scale

Figure 9: Comparison of observed (loess smoothing with 95% conficdence interval) versus predicted for the two-component model.

Table 7: Backward selection steps for the two-component mixture model.

	Selection step					
	1		2		3	
	estimate	p-value	estimate	p-value	estimate	p-value
β_0	369.4	(p<0.001)	369.4	(p<0.001)	369.6	(p<0.001)
β_{02}	454.0	(p<0.001)	453.4	(p<0.001)	452.4	(p<0.001)
β_1	3.9	(p<0.001)	3.9	(p<0.001)	3.9	(p<0.001)
β_{12}	-2.7	(p<0.001)	-2.6	(p<0.001)	-2.7	(p<0.001)
β_{1q}	0.02	(p<0.001)	0.02	(p<0.001)	0.02	(p<0.001)
β_{12q}	-0.02	(p<0.001)	-0.02	(p<0.001)	-0.02	(p<0.001)
β_2	0.65	(p<0.001)	0.64	(p<0.001)	0.65	(p<0.001)
β_{22}	-0.04	(p=0.701)	—	—	—	—
β_{2q}	0.0007	(p<0.001)	0.0007	(p<0.001)	0.0007	(p<0.001)
β_{22q}	0.0001	(p=0.572)	0.0002	(p=0.263)	—	—
Likelihood and likelihood ratio test						
ℓ	- 199,444.5		-199,444.6	0.701	-199,445.2	0.263

5.2.2 Model-fitting for k=3 components

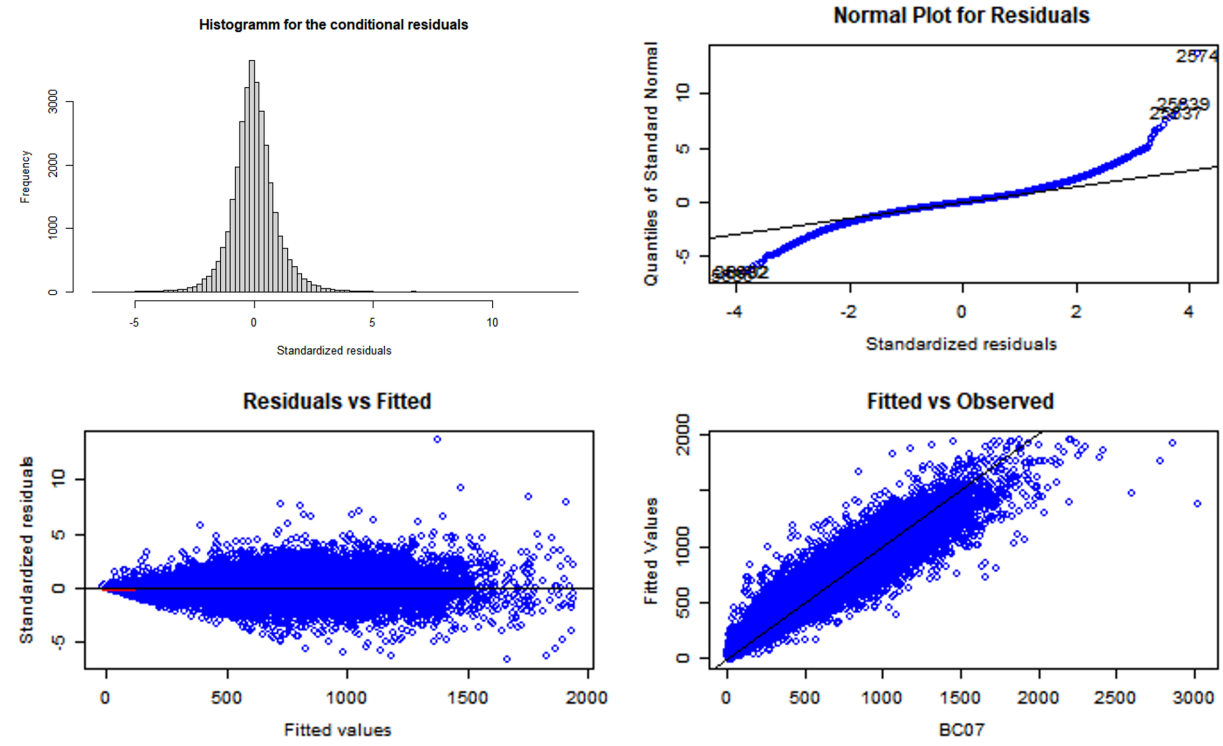
Again, the possibility of reducing the random effects structure of the model was tested by removing one random effect at a time and using likelihood ratio tests for variance components for model comparision. As for each of the random effects, the removal resulted in a significant change of the model fit ($p < 0.001$ for all), the random effects structure remained unchanged. The steps of reduction of the fixed effect structure using a backward selection departing from the initial model can be found in table 8.

5.2.2.1 Model diagnostics and remedial measures Plots for the model diagnostic of the model derived in the previous section can be found in figure 10(a).

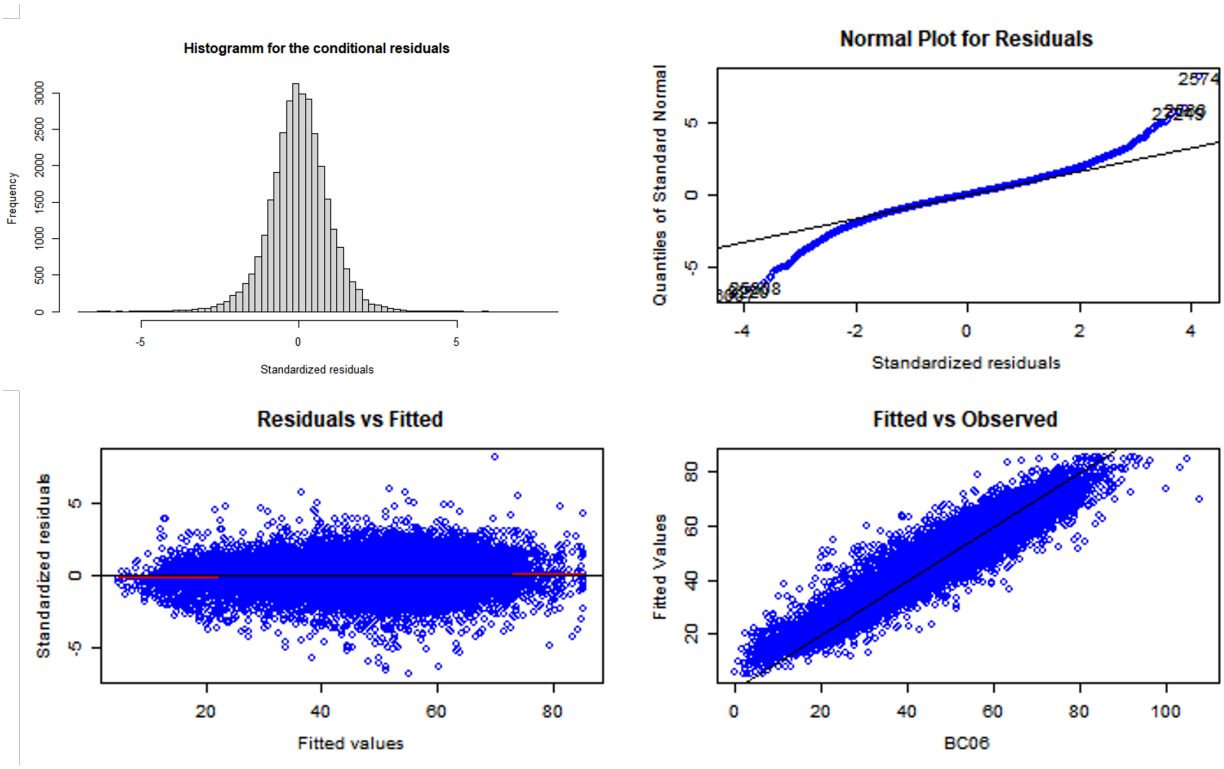
Again, in the lowest and particularly in the highest ranges of the standardized residuals, a relevant deviation from the quantiles of the standard normal distribution can be found and some heteroscedacicity might be assumed when exploring the standardized residuals over the range of fitted values. Among the series of Box-Cox transformation, a transformation using λ=0.5 seemed to reduce the deviation of the residuals from the normal assumptions best (figure 10(b)).

A comparison between the observed against the predicted mean average evolution of the $CD3^+/CD4^+$ over time is given in figure 11.

Of note, the deviation from the predicted versus the observed $CD3^+/CD4^+$ cell concentration increases markedly, particularly in the group with the lowest concentrations at baseline after Box-Cox transformation, seemingly because of a bad fit in the first time interval (up to 104 weeks). This might be related to elimination of the linear time effect for the intermediate group (k=2) from the model, which forces a common coefficient for both groups. Re-introducing the linear time effect into the model on the Box-Cox transformed scale leads to a significantly increased fit of the modified model (p<0.001) and increases the agreement between observed and predicted concentrations of $CD3^+/CD4^+$ cells on the Box-Cox transformed scale (figure 12).

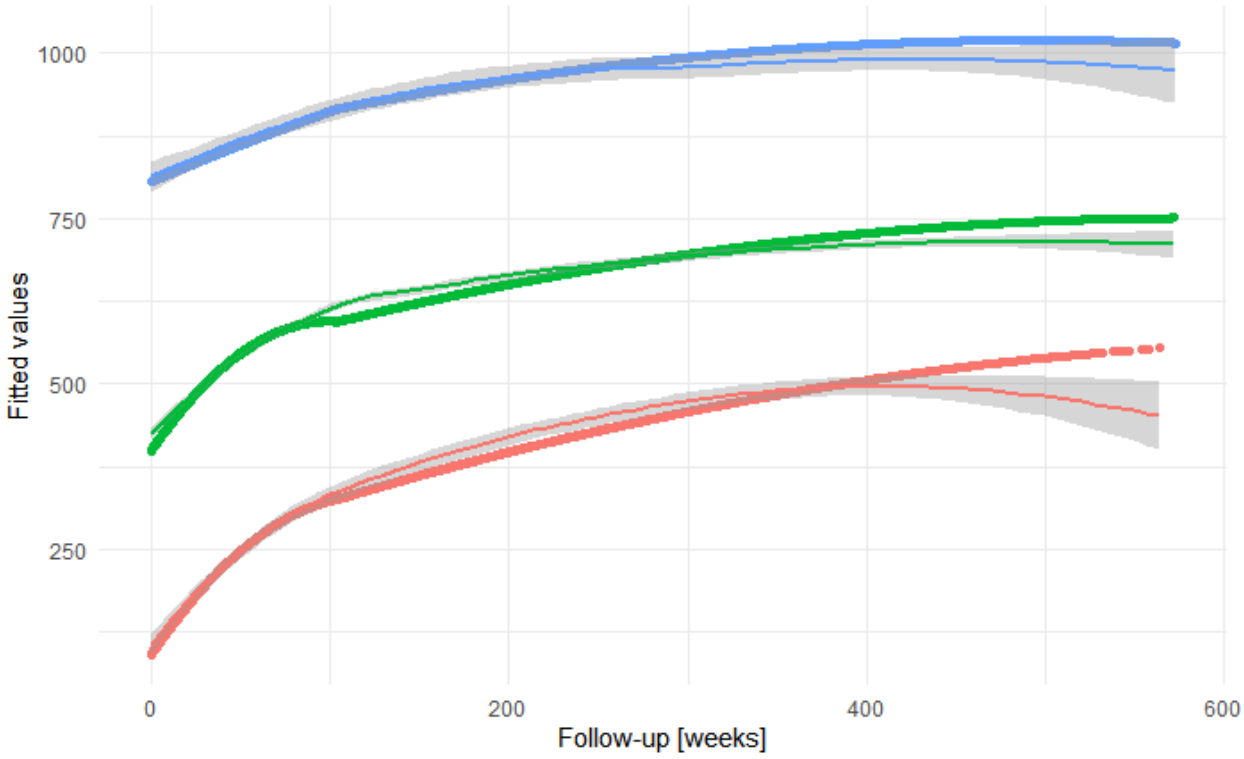


(a) Original scale

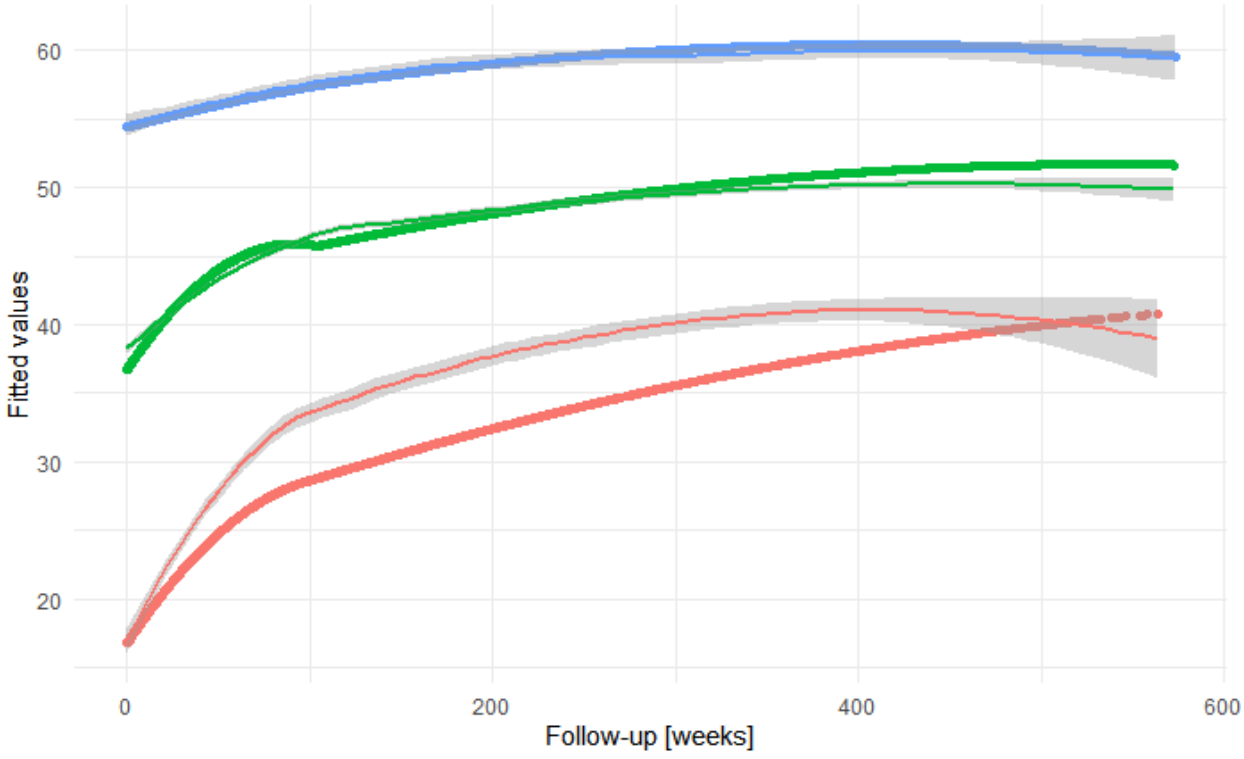


(b) Transformed scale

Figure 10: Plots for the model diagnostic on the original (a) as well as the Box-Cox transformed scale (b) of $CD3^+/CD4^+$ cell concentration for the three-component finite mixture model.

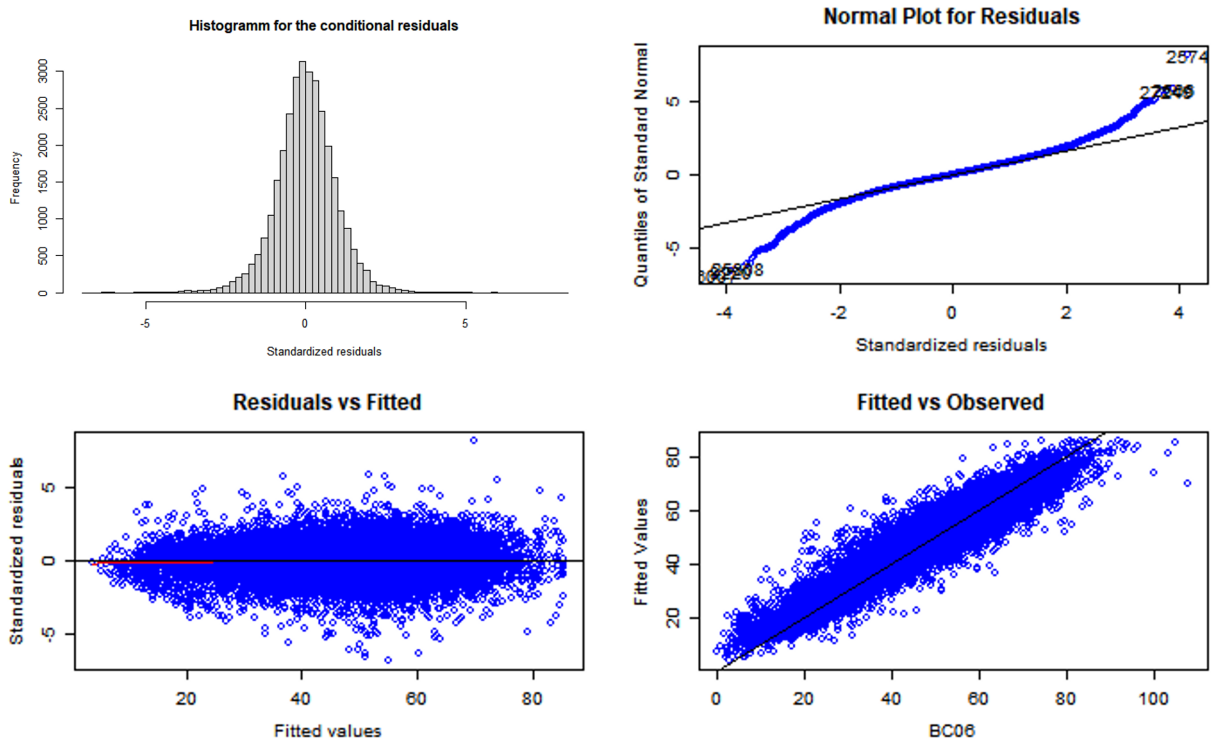


(a) Original scale

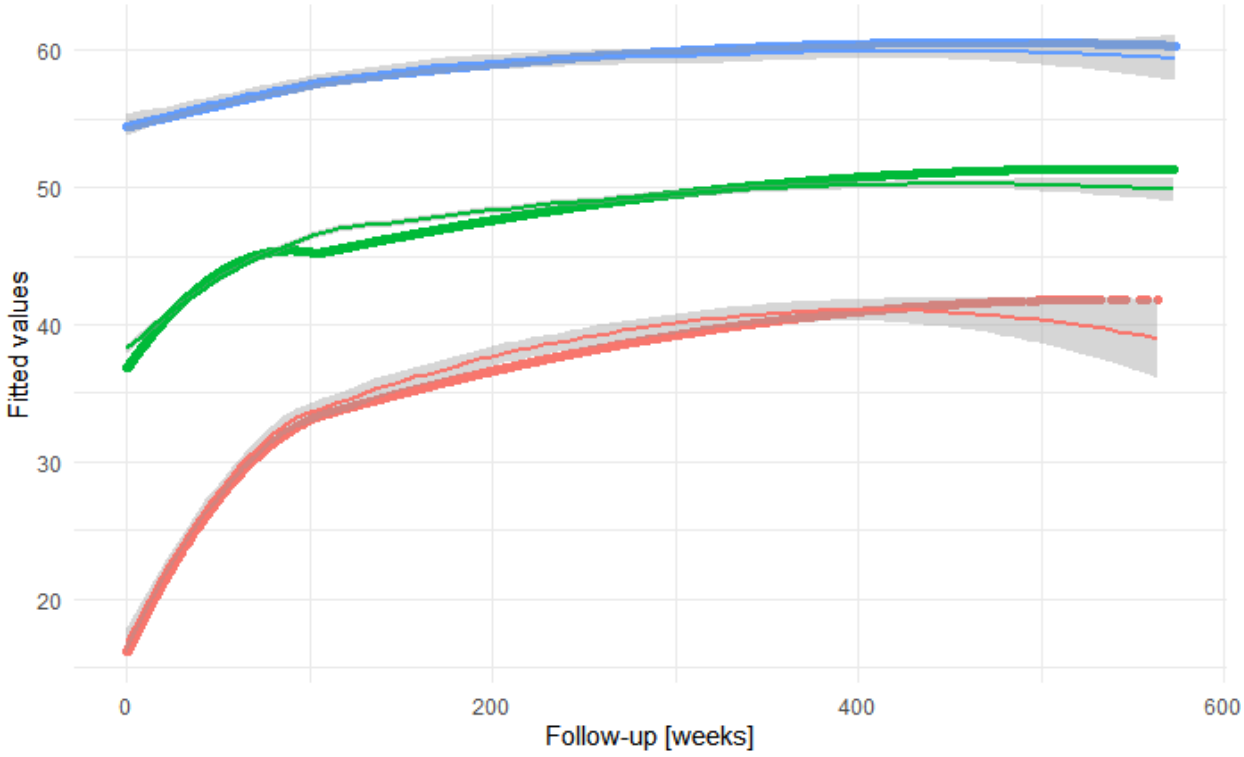


(b) Transformed scale

Figure 11: Comparison of observed (loess smoothing with 95% conficdence interval) versus predicted for the three-component model.



(a) Plot for the model diagnostic for the modified model for the Box-Cox transformed dependent variable



(b) Comparison of observed versus predicted for the modified three-component model

Figure 12: Model diagnostic and observed-vs-fitted plot for the modified three-component model.

Table 8: Backward selection steps for the three-component mixture model.

	Selection step							
	1		2		3		4	
	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value
β_0	88.5	(p<0.001)	91.0	(p<0.001)	91.0	(p<0.001)	90.9	(p<0.001)
β_{02}	311.0	(p<0.001)	308.1	(p<0.001)	308.1	(p<0.001)	308.2	(p<0.001)
β_{03}	718.0	(p<0.001)	715.4	(p<0.001)	715.2	(p<0.001)	715.2	(p<0.001)
β_1	4.3	(p<0.001)	3.9	(p<0.001)	3.9	(p<0.001)	3.9	(p<0.001)
β_{12}	-0.35	(p=0.398)	—	—	—	—	—	—
β_1	-3.1	(p<0.001)	-2.8	(p<0.001)	-2.8	(p<0.001)	-2.8	(p<0.001)
β_{1q}	0.02	(p<0.001)	0.02	(p<0.001)	0.02	(p<0.001)	0.02	(p<0.001)
β_{12q}	0.0003	(p=0.092)	0.003	(p<0.001)	0.004	(p=0.036)	0.004	(p=0.025)
	-0.02	(p<0.001)	-0.02	(p<0.001)	-0.02	(p<0.001)	-0.02	(p<0.001)
β_2	0.95	(p<0.001)	0.94	(p<0.001)	0.83	(p<0.001)	0.81	(p<0.001)
β_{22}	-0.31	(p=0.017)	-0.30	(p=0.022)	-0.19	(p=0.077)	-0.16	(p=0.092)
	-0.44	(p=0.003)	-0.42	(p=0.004)	-0.28	(p=0.010)	-0.28	(p=0.009)
β_{2q}	0.001	(p=0.001)	0.001	(p=0.001)	0.001	(p<0.001)	0.001	(p<0.001)
β_{22q}	-0.0005	(p=0.095)	-0.0005	(p=0.121)	-0.0001	(p=0.548)	—	—
β_{23q}	-0.0005	(p=0.117)	-0.0005	(p=0.144)	—	—	—	—
Likelihood and likelihood ratio test								
ℓ	-199,248.8		-199,249.2	0.398	-199,250.2	0.144	-199,250.4	0.547

5.2.3 Final models

5.2.3.1 Final two-component model The parameter estimates for the final two-component model on the Box-Cox transformed scale for the dependent variable using λ=0.5 is displayed in 9.

5.2.3.2 Final three-component model The parameter estimates for the final three-component model on the Box-Cox transformed scale for the dependent variable using λ=0.5 is displayed in 10.

Table 9: Parameter estimates for the final two-component mixture model on the Box-Cox transformed scale of the dependent variable.

	Value	Std.Error	DF	t-value	p-value
Intercepts					
β_0	34.8	0.3	29885	119.9	<0.001
β_{02}	19.9	0.8	1441	26.2	<0.001
Slopes					
β_1	0.20	0.01	29885	38.4	<0.001
β_{12}	-0.16	0.01	29885	-11.4	<0.001
β_{1q}	0.001	0.000	29885	25.0	<0.001
β_{12q}	-0.0009	0.0001	29885	-8.1	<0.001
β_2	0.03	0.00	29885	21.1	<0.001
β_{2q}	0.00003	0.00000	29885	12.1	<0.001

Table 10: Parameter estimates for the final three-component mixture model on the Box-Cox transformed scale of the dependent variable.

	Value	Std.Error	DF	t-value	p-value
Intercepts					
β_0	16.2	0.70	29880	23.2	<0.001
β_{02}	20.7	0.74	1440	28.1	<0.001
β_{03}	38.3	0.89	1440	43.2	<0.001
Slopes					
β_1	0.28	0.009	29880	30.2	<0.001
β_{12}	-0.08	0.009	29880	-9.9	<0.001
β_{13}	-0.24	0.015	29880	-15.8	<0.001
β_{1q}	0.0011	0.00004	29880	25.5	<0.001
β_{13q}	-0.0011	0.00011	29880	-9.6	<0.001
β_2	0.0391	0.00452	29880	8.6	<0.001
β_{22}	-0.0104	0.00471	29880	-2.2	0.028
β_{23}	-0.02	0.006	29880	-4.1	<0.001
β_{2q}	0.00004	0.000011	29880	4.2	<0.001
β_{22q}	-0.00001	0.000011	29880	-1.0	0.302
β_{23q}	-0.00002	0.000012	29880	-1.8	0.067

6 Discussion

With more than 1,400 PWH for the cross-sectional and over 30,000 observations for the longitudinal analysis, it was possible to identify a considerable number of people to be included into this analysis after applying in- and exclusion criteria. Participants were predominantly male with a median age of 38 years, which is in line with characteristics of the overall population of PWH in Germany[15], which might be of interest when considering generalization of the data presented in this study.

When exploring the distribution of $CD3^+/CD4^+$ cells at first presentation at the study site (figure 3), it is evident that PWH with a wide variety of $CD3^+/CD4^+$ cell concentrations were included into the study (0 - 1478 cells/µL) which seems important for its objective. Of note, 647 (44.6 %) and 305 (21.0 %) of all participants presented with CD4 cells <350 cells/µL and <200 cells/µL, respectively, and therefore represented groups of people with late diagnosis or advanced HIV disease.

This study aimed at accounting for heterogeneity in the distribution of $CD3^+/CD4^+$ cells at the first contact with a specialized HIV clinic by means of finite mixture models. This is not only motivated by the fact that a single normal distribution does not seem to describe the distribution of $CD3^+/CD4^+$ cells at first presentation entirely (figure 3); it is clinically motivated by the knowledge that the time between contracting the virus and the diagnosis of HIV is highly variable but will eventually determine in which 'state' of the disease someone will be diagnosed, be transferred to a specialized HIV-center, or start antiretroviral treatment. Therefore, the distribution as explored in this study can be seen as a mixing distribution for which latent groups are presented by disease 'stage' as a major source of heterogeneity. These 'stages' are highly relevant with regard to the further clinical course of the disease: as described previously, there is a big body of evidence linking late diagnosis with adverse outcomes and a higher risk of disease progression and development of AIDS defining diseases[6,7,16–20]. Therefore, assuming two or three sub-populations in the mixing distribution of $CD3^+/CD4^+$ cells at first presentation might be justified.

Looking at the fit of these two finite mixture models, both seem to describe the general form of the distribution of $CD3^+/CD4^+$ cells well (figure 5). Of interest: adding one more component to the two-component finite mixture model identifies an additional normal distribution at the lower end of the spectrum which seems in good line with the current clinical classification of people with late diagnosis and presentation with advanced HIV disease. While adding a third component might therefore have a good clinical justification, the conclusion is less straightforward from a statistical point of view, as due to the boundary issues of restricted models, the log-likelihood test statistic does not follow a χ^2-distribution and cannot be used for inference in the usual way[21–23]. Information criteria are often used to compare mixture models with different numbers of components, of which AIC and BIC are probably the most well-known ones. With both criteria, the three-component model seemed to fit the model better, even when accounting for the higher number of parameters, which both information criteria do by penalizing a higher number of parameters[21,22]. Preference of the three-component model is also supported by the approach suggested by Schlattmann[14], in which the number of support points k is estimated from VEM replicates on bootstrap samples with replacement from the original data, resulting in a discret distribution of the estimated 'best' number of support points $\hat{k}$. The mode of the distribution of $\hat{k}$ is then considered as best supported by the data.

The sub-populations identified by the finite mixture models should be clincially meaningful in order to have a practical relevance. Having a closer look at the three-component model, it can be seen that the normal distributions in the mixing distribution are centered at about 50, 350, and 650 cells/µL; intersections are

found at approximately 100 and 600 cells/µL. The three subgroups might therefore be seen as representations from populations of PWH with a 'normal' immune status or at an early stage (>600 cells/µL), an intermediate stage at the time of diagnosis (100 – 600 cells/µL), as well as a group with advanced HIV disease or a 'late' stage (< 100 cells/µL). It is noteworthy that in this model the sub-population with the lowest $CD3^+/CD4^+$ cell count is characterized by a much lower threshold than any other routinely used cut-off, of which <200 cells/µL is the lowest. Interestingly, for most opportunistic infections, the median $CD3^+/CD4^+$ cell concentration is considerably lower than 200 at the time of diagnosis, and often around or lower than <100 cells/µL[24]. While the group with the highest $CD3^+/CD4^+$ cell count does not need a lot of explanation, it might be worth giving the intermediate group another though. First of all, this groups contains PWH that are considered 'late diagnosed', and even presenting with advanced HIV disease as per usual classification, harboring both, PWH with < 350 but also < 200 cells/µL, but also people with $CD3^+/CD4^+$ cell counts within the range of people without HIV (> 450 cells/µL). It should be considered that while still demonstrating a concentration of $CD3^+/CD4^+$ cells in the lower-normal range at the time of diagnosis, people in this groups might nevertheless have experienced a decrease in their $CD3^+/CD4^+$ cells, as their 'set-point' prior to being infected with HIV is usually not known. While the meaning of the subgroups in terms of longitudinal evolution will be elaborated on in more detail in a further paragraph of this section, it can already be seen now, that the characteristic of the people assigned to different groups by the posterior probabilities from the finite mixture models, are markedly different (tables 5 and **??**). Regardless of the number of components used for the finite mixture model, higher age seems to be associated with lower $CD3^+/CD4^+$ cell concentrations, most likely as an indicator of later presentation, which is in good line with previous publications[25]. Also, lower $CD3^+/CD4^+$ cell concentrations are associated with higher concentrations of HIV-1 RNA, lower $CD4^+/CD8^+$ ratios, as well as a higher probability to present with AIDS-defining diseases. The difference in the proportion of PWH presenting with AIDS-defining disease between neighbouring sub-populations was most remarkable between the components with low and intermediate $CD3^+/CD4^+$ cell concentrations in the three component model, with 42.4% vs. 6.3%, respectively. This can be seen as an indicator that it *does* make sense to further distinguish PWH that are summarized in the first component of the two-component mixture model. This is further supported by the fact, that in the three component model, there was a higher proportion of PWH that had died at the time of data collection in the groups of lowest $CD3^+/CD4^+$ cells at baseline.

Having a closer look at the meaning of the overall study sample, but also the different sub-populations in terms of longitudinal development of $CD3^+/CD4^+$ cells over time, there seems to be a different evolution in the first two years when compared to the time thereafter. This result of the exploratory data analysis motivated a regression model with different parameters for this initial time, corresponding to the most significant immune reconstitution on average, compared to the time interval afterwards, while the cut-off at 104 weeks was a data-driven choice and therefore a source of over-fitting, other studies have reported similar time spans of most relevant immune recovery following ART initiation[26–28]. It might therefore be justified to assume biologically different situations for the two different time intervals: In the first two years after first presentation, ART initiation might be the dominant factor to determine changes in immune cell concentrations, while in the further course the evolution might be more strongly influenced by other factors, including age, comorbidities, and comedication. While a more 'unique' modelling approach for the entire period of observation might be achieved using a non-linear model, the split regression approach used in the study seems biologically justified, particularly as interest circled mostly around the early phase of the immune reconstitution, which is the main driver of the overall immune recovery. It is worth noting some similarities

for the two- and three-component models: First, the 'plateau' of the $CD3^+/CD4^+$ concentration after the phase of immune reconstitution is different for different subgroups. In particular, lower concentrations at initiation of ART are associated with a lower plateau after restoration of the HIV-caused depletion of $CD3^+/CD4^+$ cells, which is highly supportive of the current recommendation of an early start of ART, as later initiation at lower $CD3^+/CD4^+$ cell concentrations might result in a less complete quantitative immune reconstitution. This is therefore in good line with other results[29–32]. While the magnitude of the plateau is different for different subgroups, all of them are achieved at around two years after first presentation and therefore (in most PWH) ART initiation. This is particularly interesting as it indicated that after two years, a good estimate of the overall immune reconstitution might be obtained, in line with the immune recovery after two years being a good surrogate of long term outcomes in PWH presenting late[33]. Second, both models lost comparably in model fit, when trying to remove random effects from the model. This seems understandable when having a look at the individual profiles of study participants (figure 7), where high inter-individual variability in terms of baseline $CD3^+/CD4^+$ and evolution over time is found. This seems to justify the inclusion of random intercepts and slopes. Third, and more technically, for both models the sub-population belonging to the distribution with the lowest $CD3^+/CD4^+$ mean, was taken as the reference. For the interpretation of the parameters, this parameterization has to be kept in mind for correct inference. Yet, in both models, PWH belonging to the sub-population with the lowest average concentration at baseline are attributed the highest increase of $CD3^+/CD4^+$ cells per time unit.

6.1 Limitations

This study has several limitations. First of all, despite having a considerable sample size, it must be kept in mind that all PWH included in this study were taken care of in a single HIV clinical care center. Therefore, there might be some homogeneity in terms of choice of initial antiretroviral treatment, time to initiation, but also a selection of people attending the clinic in general, which might be relevant in terms of patient characteristics. This is, of course, noteworthy, when thinking about generalizing the results presented to PWH outside this clinic. In particular, the model developed might be very tightly fitted to PWH attending the study site, which highly impairs generalizability. Several factors, that might influence immune recovery after ART initiation, were not taken into account in the model in order to focus. Yet, factors such as sex[30], body-mass index[32,34], and the choice of antiretroviral regimen itself[33,35] might be of interest and contribute to a better fit of the model. However, it must be kept in mind that the primary aim of the study was to explore general differences in immune recovery between different sub-populations in the mixing distribution for $CD3^+/CD4^+$ cells to support the plausibility of the FMM, rather than the optimal modelling of the longitudinal development itself. Another limitation is, that the date of actual initiation of antiretroviral therapy could not be extracted from the data. Therefore, particularly in the sub-groups with higher $CD3^+/CD4^+$ cell concentrations at baseline, antiretroviral therapy might not have been initiated right away. While a higher number of deaths was found in the groups of lower $CD3^+/CD4^+$ cell concentrations, the number of these events was overall low and not adjusted for the possibility of different observation times and age, which does not allow for a robust conclusion. The presence of AIDS-defining conditions first presentation was identified by individual electronic patient file review, which is prone to mistakes.

6.2 Ethics, societal relevance and stakeholder awareness

The results presented in this study imply, that PWH might already be diagnoses 'late' when still having $CD3^+/CD4^+$ cells within the reference range. This is of clinical relevance as this group of people in whom watchful waiting is sometimes advocated even today, might already be at a significantly higher risk of relevant HIV-related comorbidities including AIDS-defining conditions and probably even higher mortality. Interestingly, the classification derived from the finite mixture model differs from the conventionally used classification such as the one from the Centers for Disease Control and Prevention (CDC) or consensus definitions of late diagnosis or diagnosis with advanced HIV disease and might therefore be seen as a complementary measure for individual risk assessment in PWH, particularly in the sub-group of the intermediate group in the three-component model.

7 Conclusion

The distribution of $CD3^+/CD4^+$ cells at the first presentation as analysed in this study seems to be well described by a finite mixture model with three components with component-specific variances. The three groups seem to be relevant with regard to the probability of having AIDS-defining conditions and probably also mortality. Immune recovery seems, on average, to be different for people belonging to different subpopulations of the mixing distribution, where for all groups the most relevant increase in $CD3^+/CD4^+$ cells occurs within the first two years after initial presentation at a specialized center, where most PWH might have received quick ART initiation.

7.1 Further research

First and foremost, further research should try to veri- or falsify what has been described in this study. Before any other steps, it must be made clear if alternative thresholds for classification of PWH according to baseline $CD3^+/CD4^+$ count also seem to be plausibly in independent study samples and if they are comparable to cut-offs identified here (external validation). It seems worth investigating if more elaborate mixture models, including Bayesian mixture models, might contribute to better discrimination of PWH into the different subgroups by taking other factors into account and not only relying on $CD3^+/CD4^+$ cell concentrations.

8 Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Ziv Shkedy, for his expert guidance, patience, and constructive suggestions throughout the completion of this master's thesis. His support has been invaluable in deepening my understanding of the topic.

A big thank you is dedicated to Dr. Eva Wolf for checking the manuscript carefully for concistency, spelling, and grammer.

I would like to thank Mrs. Anna-Maria Balogh for the immense support in data extraction.

9 Literature

1. Hofmann C, Rockstroh JK, editors. HIV Buch. Medizin Fokus Verlag; 2024.

2. Payagala S, Pozniak A. The global burden of HIV. Clinics in Dermatology [Internet] 2024;42(2):119–27. Available from: http://dx.doi.org/10.1016/j.clindermatol.2024.02.001

3. Croxford S, Stengaard AR, Brännström J, et al. Late diagnosis of HIV: An updated consensus definition. HIV Medicine [Internet] 2022;23(11):1202–8. Available from: http://dx.doi.org/10.1111/hiv.13425

4. Antinori A, Coenen T, Costagiola D, et al. Late presentation of HIV infection: a consensus definition. HIV Medicine [Internet] 2010;12(1):61–4. Available from: http://dx.doi.org/10.1111/j.1468-1293.2010.00857.x

5. MacCarthy S, Bangsberg D, Fink G, Reich M, Gruskin S. Late presentation to HIV/AIDS testing, treatment or continued care: clarifying the use of CD4 evaluation in the consensus definition. HIV Medicine [Internet] 2013;15(3):130–4. Available from: http://dx.doi.org/10.1111/hiv.12088

6. Estimating the burden of HIV late presentation and its attributable morbidity and mortality across Europe 2010–2016. BMC Infectious Diseases [Internet] 2020;20(1). Available from: http://dx.doi.org/10.1186/s12879-020-05261-7

7. Mocroft A, Lundgren JD, Sabin ML, et al. Risk Factors and Outcomes for Late Presentation for HIV-Positive Persons in Europe: Results from the Collaboration of Observational HIV Epidemiological Research Europe Study (COHERE). PLoS Medicine [Internet] 2013;10(9):e1001510. Available from: http://dx.doi.org/10.1371/journal.pmed.1001510

8. Trickey A, Sabin CA, Burkholder G, et al. Life expectancy after 2015 of adults with HIV on long-term antiretroviral therapy in Europe and North America: a collaborative analysis of cohort studies. The Lancet HIV [Internet] 2023;10(5):e295–307. Available from: http://dx.doi.org/10.1016/S2352-3018(23)00028-0

9. Teeraananchai S, Kerr S, Amin J, Ruxrungtham K, Law M. Life expectancy of HIV-positive people after starting combination antiretroviral therapy: a meta-analysis. HIV Medicine [Internet] 2016;18(4):256–66. Available from: http://dx.doi.org/10.1111/hiv.12421

10. Marcus JL, Leyden WA, Alexeeff SE, et al. Comparison of Overall and Comorbidity-Free Life Expectancy Between Insured Adults With and Without HIV Infection, 2000-2016. JAMA Network Open [Internet] 2020;3(6):e207954. Available from: http://dx.doi.org/10.1001/jamanetworkopen.2020.7954

11. Trickey A, Sabin CA, Burkholder G, et al. Life expectancy after 2015 of adults with HIV on long-term antiretroviral therapy in Europe and North America: a collaborative analysis of cohort studies. The Lancet HIV [Internet] 2023;10(5):e295–307. Available from: http://dx.doi.org/10.1016/S2352-3018(23)00028-0

12. Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection. New England Journal of Medicine [Internet] 2015;373(9):795–807. Available from: http://dx.doi.org/10.1056/NEJMoa1506816

13. Ryom L, Cotter A, De Miguel R, et al. 2019 update of the European AIDS Clinical Society Guidelines for treatment of people living with HIV version 10.0. Wiley [Internet] 2020;Available from: https://boris.unibe.ch/146424/

14. Schlattmann P. Estimating the number of components in a finite mixture model: the special case of homogeneity. Computational Statistics & Data Analysis [Internet] 2003;41(3-4):441–51. Available from: http://dx.doi.org/10.1016/s0167-9473(02)00173-1

15. Rki. Epidemiologisches bulletin 35/2023. 2023;Available from: www.rki.de/epidbull

16. Domínguez-Domínguez L, Rava M, Bisbal O, et al. Low CD4/CD8 ratio is associated with increased morbidity and mortality in late and non-late presenters: Results from a multicentre cohort study, 2004–2018. BMC Infectious Diseases 2022;22.

17. Mondi A, Cozzi-Lepri A, Tavelli A, et al. Persistent poor clinical outcomes of people living with HIV presenting with AIDS and late HIV diagnosis – results from the ICONA cohort in italy, 2009-2022. International Journal of Infectious Diseases 2024;142.

18. Lee C-Y, Lin Y-P, Wang S-F, Lu P-L. Late cART initiation consistently driven by late HIV presentation: A multicenter retrospective cohort study in taiwan from 2009 to 2019. Infectious diseases and therapy [Internet] 2022;11:1033–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/35301666 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC9124249

19. Severin S, Delforge M, Wit SD. Epidemiology, comorbidities at diagnosis and outcomes associated with HIV late diagnosis from 2010 to 2019 in a belgian reference centre: A retrospective study. HIV Medicine 2022;23:1184–94.

20. Candevir A, Kuscu F, Kurtaran B, et al. Late diagnosis in HIV with new and old definitions; data from a regional hospital in turkey. International Journal of General Medicine 2023;4227–34.

21. Grimm KJ, Houpt R, Rodgers D. Model fit and comparison in finite mixture models: A review and a novel approach. Frontiers in Education [Internet] 2021;6. Available from: http://dx.doi.org/10.3389/feduc.2021.613645

22. Melnykov V, Maitra R. Finite mixture models and model-based clustering. Statistics Surveys [Internet] 2010;4(none). Available from: http://dx.doi.org/10.1214/09-ss053

23. Lo Y. Testing the number of components in a normal mixture. Biometrika [Internet] 2001;88(3):767–78. Available from: http://dx.doi.org/10.1093/biomet/88.3.767

24. Gandhi, Neel, Cescon, Angela, Saag, Michael S., et al. Incidence of AIDS-Defining Opportunistic Infections in a Multicohort Analysis of HIV-infected Persons in the United States and Canada, 2000–2010. The University of North Carolina at Chapel Hill University Libraries [Internet] 2016;Available from: https://cdr.lib.unc.edu/concern/articles/b27741711

25. Darcis G, Lambert I, Sauvage A-S, et al. Factors associated with late presentation for HIV care in a single Belgian reference center: 2006–2017. Scientific Reports [Internet] 2018;8(1). Available from: http://dx.doi.org/10.1038/s41598-018-26852-0

26. De Beaudrap P, Etard J-F, Diouf A, et al. Modeling CD4+ cell count increase over a six-year period in HIV-1-infected patients on highly active antiretroviral therapy in senegal. American Journal of Tropical Medicine and Hygiene 2009;80(6):1047.

27. Asfaw A, Ali D, Eticha T, Alemayehu A, Alemayehu M, Kindeya F. CD4 cell count trends after commencement of antiretroviral therapy among HIV-infected patients in tigray, northern ethiopia: A retrospective cross-sectional study. PloS one 2015;10(3):e0122583.

28. Mugo CW, Shkedy Z, Mwalili S, et al. Modelling trends of CD4 counts for patients on antiretroviral therapy (ART): A comprehensive health care clinic in nairobi, kenya. BMC infectious diseases 2022;22(1):29.

29. Thornhill JP, Fox J, Martin GE, et al. Rapid antiretroviral therapy in primary HIV-1 infection enhances immune recovery. AIDS [Internet] 2023;38(5):679–88. Available from: http://dx.doi.org/10.1097/QAD.0000000000003825

30. Kouamou V, Gundidza P, Ndhlovu CE, Makadzange AT. Factors associated with CD4+ cell count recovery among males and females with advanced HIV disease. AIDS [Internet] 2023;37(15):2311–8. Available from: http://dx.doi.org/10.1097/qad.0000000000003695

31. Chen L, Liu C-H, Kang S, et al. Determinants of suboptimal immune recovery among a Chinese Yi ethnicity population with sustained HIV suppression. BMC Infectious Diseases [Internet] 2022;22(1). Available from: http://dx.doi.org/10.1186/s12879-022-07113-y

32. Han W, Jiamsakul A, Jantarapakde J, et al. Association of body mass index with immune recovery, virological failure and cardiovascular disease risk among people living with HIV. HIV Medicine [Internet] 2020;22(4):294–306. Available from: http://dx.doi.org/10.1111/hiv.13017

33. Martin-Iguacel R, Reyes-Urueña J, Bruguera A, et al. Determinants of long-term survival in late HIV presenters: The prospective PISCIS cohort study. eClinicalMedicine [Internet] 2022;52:101600. Available from: http://dx.doi.org/10.1016/j.eclinm.2022.101600

34. Zhu J, Huang H, Wang M, et al. High baseline body mass index predicts recovery of CD4+ T lymphocytes for HIV/AIDS patients receiving long-term antiviral therapy. PLOS ONE [Internet] 2022;17(12):e0279731. Available from: http://dx.doi.org/10.1371/journal.pone.0279731

35. Tongtong Y, Shenghua H, Yin W, et al. Effectiveness and Safety of Dolutegravir Versus Efavirenz-Based Antiviral Regimen in People Living With HIV-1 in Sichuan Province of China: A Real-World Study. JAIDS Journal of Acquired Immune Deficiency Syndromes [Internet] 2022;91(S1):S1–7. Available from: http://dx.doi.org/10.1097/qai.0000000000003041

A Supplementary Material

A.1 Countries and Regions

Table 11: Overview over the possible countries of origin and their assigment to a geographic region for the purpose of this study.

Asia	Afghanistan, Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei, Cambodia, China, Cyprus, Georgia, India, Indonesia, Japan, Kazakhstan, Kuwait, Kyrgyzstan, Laos, Malaysia, Maldives, Mongolia, Myanmar, Nepal, North Korea, Oman, Pakistan, Palestine, Philippines, Qatar, Saudi Arabia, Singapore, South Korea, Sri Lanka, Syria, Taiwan, Tajikistan, Thailand, Timor-Leste, Turkmenistan, United Arab Emirates, Uzbekistan, Vietnam
Carribean	Anguilla, Antigua and Barbuda, Antigua, Aruba, Bahamas, Barbados, Barbuda, Bermuda, British Virgin Islands, Bonaire, Cayman Islands, Cuba, Curacao, Dominica, Dominican Republic, Grenada, Guadeloupe, Haiti, Jamaica, Martinique, Montserrat, Netherlands Antilles, Puerto Rico, St. Kitts and Nevis, Saint Kitts, Saint Lucia, St. Lucia, St. Vincent and Grenadines, Saint Martin, Sint Maarten, Trinidad and Tobago, Turks and Caicos Islands, US Virgin Islands
Eastern Europe	Ukraine, Belarus, Russia
Europe	Albania, Andorra, Austria, Azores, Belgium, Bosnia and Herzegovina, Bulgaria, Canary Islands, Czech Republic, Croatia, Cyprus, Denmark, Estonia, Faroe Islands, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Isle of Man, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Madeira Islands, Malta, Moldova, Monaco, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, San Marino, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom, Vatican
Latinamerica	Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, Falkland Islands, French Guiana, Guatemala, Guyana, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, South Georgia, Suriname, Uruguay, Venezuela
Middle East	Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates, Iraq, Iran, Israel, Jordan, Lebanon, Palestine, Syria, Turkey, Yemen
North Africa	Algeria, Egypt, Libya, Mauritania, Morocco, Tunisia
North America	Canada, El Salvador, Greenland, USA
Oceania	American Samoa, Australia, Cook Islands, Fiji, French Polynesia, Guam, Kiribati, Marshall Islands, Micronesia, Nauru, New Caledonia, New Zealand, Niue, Norfolk Island, Northern Mariana Islands, Palau, Papua New Guinea, Pitcairn Islands, Samoa, Solomon Islands, Tokelau, Tonga, Tuvalu, Vanuatu, Wallis and Futuna
Subsaharan Africa	Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo - Brazzaville, Congo - Kinshasa, Democratic Republic of the Congo, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mozambique, Namibia, Niger, Nigeria, Republic of Congo, Rwanda, São Tomé and Príncipe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Sudan, Tanzania, Togo, Uganda, Western Sahara, Zambia, Zimbabwe

A.2 Boostrap approach to estimation of destribution of $\hat{k}$

Following a suggestion by Schlattmann[14], bootstrap sampling from the original observations of $CD3^+/CD4^+$ cell concentrations at first presentation was performed with replacement. For each of the $B = 100$ bootstrap samples, the number of components ($\hat{k}$) in the distribution of $CD3^+/CD4^+$ cell concentrations was estimated using the VEM algorithm and the distribution of $\hat{k}$ obtained from the bootstrap replicates was explored visually (figure 13) as well as using a frequency table (table 12).

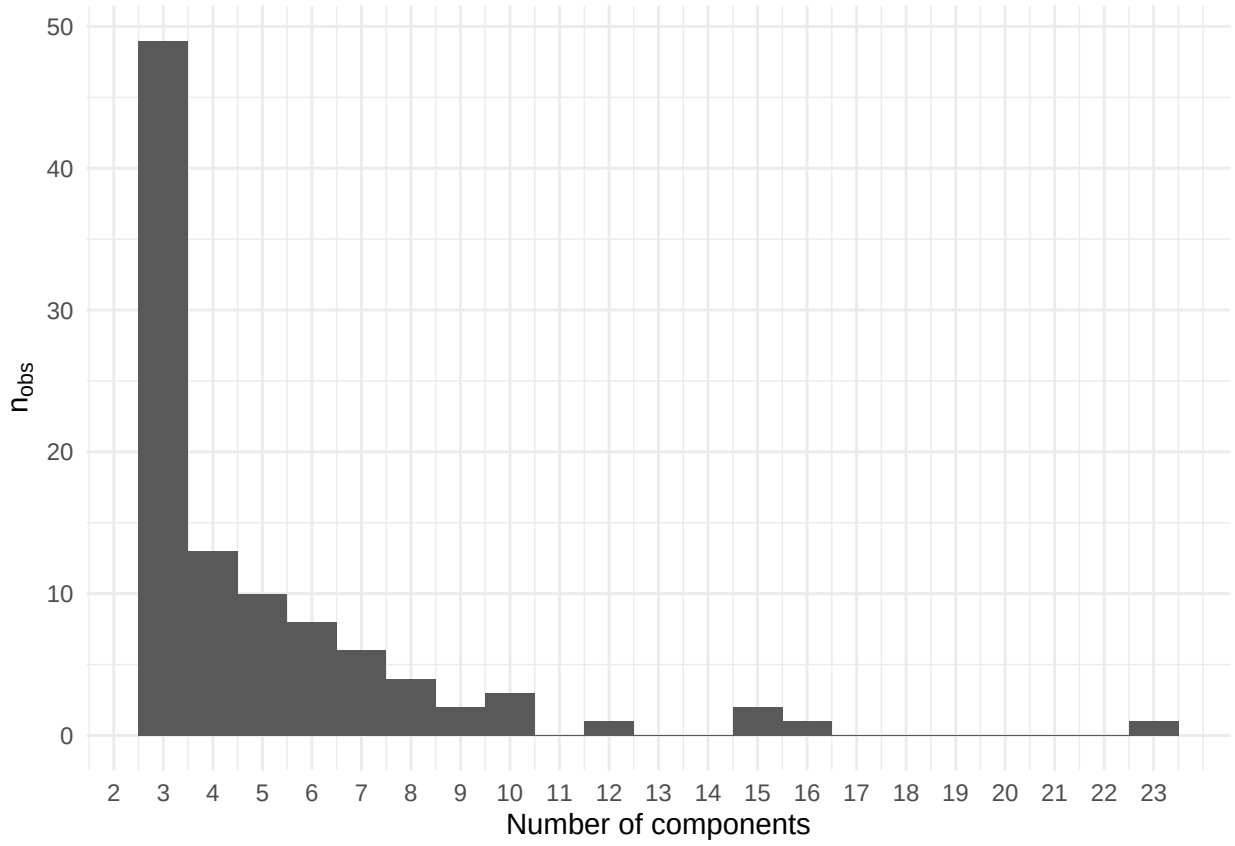


Figure 13: Distribution of the bootstrap replicates for the optimal numbers of components.

Table 12: Number of observation (n) among 100 estimations of the optimal number of components k.

	3	4	5	6	7	8	9	10	12	15	16	23
n	49	13	10	8	6	4	2	3	1	2	1	1

B R Code

B.1 Data preparation

```
## Preparation of demographic data

demographics <- read_excel("Daten/demographics.xlsx")


## Preparation of main data

raw <- read_delim("Daten/Rohdaten.csv", delim = ";", escape_double = FALSE, trim_ws = TRUE)

n_initial = unique(raw$CenterPatId) |> length()

data = subset(raw, is.na(Date_sampletaken) == F)
data = subset(data, is.na(CD4ABS) == F)

data$Date_sampletaken = as.Date(data$Date_sampletaken, format="%Y.%m.%d",
          origin = "1970-01-01")
data$Date_1st_CD4ABS = as.Date(data$Date_1st_CD4ABS, format="%Y.%m.%d",
          origin = "1970-01-01")

data = subset(data, is.na(CenterPatId) == F)

Variablen = c("PID", "YoB", "Alter", "Geschlecht", "InfRisiko", "Erstvorstellung",
          "CD4abs", "CD4rel", "CD8abs", "CD8rel", "Ratio", "HIV_RNA",
          "Letztvorstellung", "Erstvorstellungsjahr", "Land")

n = unique(data$CenterPatId) |> length()

patienten = matrix(NA, nrow = n, ncol = length(Variablen))

patienten[,1] = unique(data$CenterPatId)

for(i in 1:n){
  temp = subset(data, CenterPatId == patienten[i,1])
  temp = temp |> arrange(SampleTakenCD4ABS)
  patienten[i,2] = temp$YearOfBirth[1]
  patienten[i,3] = as.numeric(substr(min(temp$SampleTakenCD4ABS, na.rm=T), 1, 4)) -
          temp$YearOfBirth[1]
  patienten[i,4] = temp$Gender[1]
  patienten[i,5] = temp$InfRisiko[1]
  patienten[i,6] = min(temp$SampleTakenCD4ABS, na.rm=T)
  patienten[i,7] = temp$CD4ABS[1]
```

```
  patienten[i,8] = temp$CD4REL[1]
  patienten[i,9] = temp$CD8ABS[1]
  patienten[i,10] = temp$CD8REL[1]
  patienten[i,11] = patienten[i,7] / patienten[i,9]
  patienten[i,12] = temp$HIV_1RNA[1]
  patienten[i,13] = max(temp$SampleTakenCD4ABS, na.rm=T)
}
patienten = patienten |> as.data.frame() |> "colnames<-"(Variablen)

patienten <- patienten[complete.cases(patienten[ , 'CD4abs']), ]
patienten$Erstvorstellung = as.Date(as.POSIXct(patienten$Erstvorstellung,
        origin = "1970-01-01", tz = "UTC"), format = "Y-m-d")
patienten$Letztvorstellung = as.Date(as.POSIXct(patienten$Letztvorstellung,
        origin = "1970-01-01", tz = "UTC"), format = "Y-m-d")
patienten$Erstvorstellungsjahr = substr(patienten$Erstvorstellung, 1, 4)
patienten$MSM = NA
patienten$MSM[patienten$InfRisiko == 1] = 1
patienten$MSM[patienten$InfRisiko != 1] = 0

patienten = mutate(patienten, FUZeit = difftime(patienten$Letztvorstellung,
        patienten$Erstvorstellung, units = "weeks"))
patienten$FUZeit = as.numeric(patienten$FUZeit)

data = mutate(data, Date_1st_sampletaken = NA)

for(i in 1:nrow(patienten)){
  if(sum(demographics$Nummer == patienten[i,1], na.rm=T) == 0) {patienten[i,15] = NA}
  else {patienten[i,15] = demographics$Land[demographics$Nummer == patienten[i,1]][[1]]}
}

for(i in 1:nrow(data)){
  data[i, 'Date_1st_sampletaken'] <- as.Date(patienten$Erstvorstellung[patienten$PID ==
          data[i, 'CenterPatId']][[1]])
}
data$Date_1st_sampletaken <- as.Date(data$Date_1st_sampletaken)
str(data$Date_1st_sampletaken)


data = mutate(data, followUp = as.numeric(difftime(data$Date_sampletaken,
        data$Date_1st_sampletaken, units = "weeks")))


## Selection of viremic patients
```

```r
viremic = subset(patienten, HIV_RNA >= 200)
viremic_with_FU = subset(viremic, FUZeit > 0)
viremic_long = subset(data, CenterPatId %in% viremic$PID)


viremic = merge(viremic, demographics, by.x="PID", by.y="Nummer", all.x=T)
viremic$verstorben[is.na(viremic$verstorben) == T] = 0
```

B.2 Finite mixture models

```r
nm_v_2 = normalmixEM(viremic$CD4abs, k = 2)
nm_v_3 = normalmixEM(viremic$CD4abs, k = 3)


# Schlattman approach to estimation of k

B = 100
n_pop = NA

set.seed(20232024)

for(i in 1:B){
  Auswahl = sample(1:nrow(viremic), nrow(viremic), replace = T)
  n_pop[i] = mixalg.VEM(obs=Auswahl, family = "gaussian", startk = 100,
          limit = 1)@grid$p |> length()
}


table(n_pop)
prop.table(table(n_pop))

ggplot() +
  geom_histogram(aes(n_pop), binwidth = 1) +
  scale_x_continuous(breaks = seq(1,23,1)) +
  labs(x="Number of components", y=TeX("n$_{obs}$"))


# Assigning the posterior probabilities to the participants

post2 = matrix(NA, ncol=3, nrow=nrow(viremic))
post2[,c(1:2)] = nm_v_2$posterior
```

```
for(i in 1:nrow(post2)){
  if(post2[i,1] > post2[i,2]) {post2[i, 3] = 1}
  else {post2[i,3] = 2}
}

post2 = as.data.frame(post2)
names(post2) = c("k2_1", "k2_2", "post_k2")

post3 = matrix(NA, ncol=4, nrow=nrow(viremic))
post3[,c(1:3)] = nm_v_3$posterior
for(i in 1:nrow(post3)){
  if(post3[i,1] > post3[i,2]) {kat = 1; max = post3[i,1]}
      else {kat = 2; max = post3[i,2]}
  if(max > post3[i, 3]) {kat = kat; max = max} else {kat = 3; max = post3[i,3]}
  post3[i,4] = kat
}

post3 = as.data.frame(post3)
names(post3) = c("k3_1", "k3_2", "k3_3", "post_k3")

viremic = cbind(viremic, post2, post3) |> as.data.frame()
```

B.3 Linear mixed-model

```
# Construction of a long-format data-set

breakpoint = 104

viremic_long = mutate(viremic_long, posteriorGroupCD4_k2 = NA)
viremic_long = mutate(viremic_long, posteriorGroupCD4_k3 = NA)

# Introduction of group variables for both models
viremic_long = mutate(viremic_long, g1_k2 = 0)
viremic_long = mutate(viremic_long, g2_k2 = 0)
viremic_long = mutate(viremic_long, g3_k2 = 0)
viremic_long = mutate(viremic_long, g1_k3 = 0)
viremic_long = mutate(viremic_long, g2_k3 = 0)
viremic_long = mutate(viremic_long, g3_k3 = 0)

# Calculation of t1 and t2 from t
viremic_long = mutate(viremic_long, t1 = viremic_long$followUp)
viremic_long = mutate(viremic_long, t2 = viremic_long$followUp - breakpoint)
```

```
viremic_long$t1[viremic_long$followUp >= breakpoint] = breakpoint
viremic_long$t2[viremic_long$followUp < breakpoint] = 0

# Calculation of the quadratic terms
viremic_long = mutate(viremic_long, t12 = -1*viremic_long$t1^2)
viremic_long = mutate(viremic_long, t22 = -1*viremic_long$t2^2)


for(i in 1:nrow(viremic_long)){
  pid = viremic_long[i, 'CenterPatId'][[1]]
  viremic_long[i, 'posteriorGroupCD4_k2'] = viremic$post_k2[viremic$PID == pid][1]
  viremic_long[i, 'posteriorGroupCD4_k3'] = viremic$post_k3[viremic$PID == pid][1]
}

viremic_long$g1_k2[viremic_long$posteriorGroupCD4_k2 == 1] = 1
viremic_long$g2_k2[viremic_long$posteriorGroupCD4_k2 == 2] = 1

viremic_long$g1_k3[viremic_long$posteriorGroupCD4_k3 == 1] = 1
viremic_long$g2_k3[viremic_long$posteriorGroupCD4_k3 == 2] = 1
viremic_long$g3_k3[viremic_long$posteriorGroupCD4_k3 == 3] = 1


## Box-Cox transformation of CD4 cell concentrations

transformations = seq(-2,2, by=0.5)

BC = matrix(NA, nrow=nrow(viremic_long), ncol=length(transformations))

for(i in 1:nrow(BC)){
  for(j in 1:ncol(BC)){
    if(transformations[j] == 0){
      BC[i,j] = log(viremic_long[i, 'CD4ABS'][[1]])
    } else {
      lambda = transformations[j]
      Y = viremic_long[i, 'CD4ABS'][[1]]
      BC[i,j] = (Y^lambda - 1) /(lambda)
    }
  }
}

colnames(BC) = c("BC01", "BC02","BC03","BC04","BC05","BC06","BC07","BC08","BC09")

bc_viremic_long = cbind(viremic_long, BC)
```

```
rm(i, j, transformations)
```

B.3.1 Model fitting for the two-component model

```
k2_model_1 = lme(fixed = CD4ABS ~ g2_k2 +
                 t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                 t2  + I(g2_k2 * t2) + t22 + I(g2_k2 * t22),
                  random = ~ 1 +t1 +t2 | PatientId,
                 control=lmeControl(returnObject=TRUE),
                  na.action = na.omit,
                  method = "ML",
                  data = viremic_long)
summary(k2_model_1)



k2_model_2 = lme(fixed = CD4ABS ~ g2_k2 +
                 t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                 t2  + t22 + I(g2_k2 * t22),
                  random = ~ 1 +t1 +t2 | PatientId,
                 control=lmeControl(returnObject=TRUE),
                  na.action = na.omit,
                  method = "ML",
                  data = viremic_long)
summary(k2_model_2)
anova(k2_model_1, k2_model_2)



k2_model_3 = lme(fixed = CD4ABS ~ g2_k2 +
                 t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                 t2  + t22,
                  random = ~ 1 +t1 +t2 | PatientId,
                 control=lmeControl(returnObject=TRUE),
                  na.action = na.omit,
                  method = "ML",
                  data = viremic_long)
summary(k2_model_3)
anova(k2_model_2, k2_model_3)
```

```
# Fitting of different models after Box-Cox transformation

k2_BC01 = lme(fixed = BC01 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC02 = lme(fixed = BC02 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC03 = lme(fixed = BC03 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC04 = lme(fixed = BC04 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC05 = lme(fixed = BC05 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
```

```
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC06 = lme(fixed = BC06 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC07 = lme(fixed = BC07 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC08 = lme(fixed = BC08 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k2_BC09 = lme(fixed = BC09 ~ g2_k2 +
                t1  + I(g2_k2 * t1) + t12 + I(g2_k2 * t12)  +
                t2  + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)
```

B.3.2 Model fitting for the two-component model

```
k3_model_1 = lme(fixed = CD4ABS ~ g2_k3 + g3_k3 +
                 t1  + I(g2_k3 * t1) + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                 t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22 + I(g2_k3 * t22) + I(g3_k3 * t22),
                  random = ~ 1 +t1 +t2 | PatientId,
                 control=lmeControl(returnObject=TRUE),
                  na.action = na.omit,
                  method = "ML",
                  data = viremic_long)
summary(k3_model_1)



k3_model_2 = lme(fixed = CD4ABS ~ g2_k3 + g3_k3 +
                 t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                 t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22 + I(g2_k3 * t22) + I(g3_k3 * t22),
                  random = ~ 1 +t1 +t2 | PatientId,
                 control=lmeControl(returnObject=TRUE),
                  na.action = na.omit,
                  method = "ML",
                  data = viremic_long)
summary(k3_model_2)
anova(k3_model_2, k3_model_1)



k3_model_3 = lme(fixed = CD4ABS ~ g2_k3 + g3_k3 +
                 t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                 t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22 + I(g2_k3 * t22),
                  random = ~ 1 +t1 +t2 | PatientId,
                 control=lmeControl(returnObject=TRUE),
                  na.action = na.omit,
                  method = "ML",
                  data = viremic_long)
summary(k3_model_3)
anova(k3_model_3, k3_model_2)



k3_model_4 = lme(fixed = CD4ABS ~ g2_k3 + g3_k3 +
                 t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                 t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
```

```
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = viremic_long)
summary(k3_model_4)
anova(k3_model_4, k3_model_3)



k3_model_5 = lme(fixed = CD4ABS ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = viremic_long)
summary(k3_model_5)
anova(k3_model_5, k3_model_4)


# Fitting of different models after Box-Cox transformation

k3_BC01 = lme(fixed = BC01 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC02 = lme(fixed = BC02 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC03 = lme(fixed = BC03 ~ g2_k3 + g3_k3 +
```

```
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC04 = lme(fixed = BC04 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC05 = lme(fixed = BC05 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC06 = lme(fixed = BC06 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC07 = lme(fixed = BC07 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)
```

```
k3_BC08 = lme(fixed = BC08 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

k3_BC09 = lme(fixed = BC09 ~ g2_k3 + g3_k3 +
                t1  + I(g3_k3 * t1) + t12 + I(g2_k3 * t12) + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22,
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)


# Modification of the model to improve fit

k3_BC06_mod = lme(fixed = BC06 ~ g2_k3 + g3_k3 +
                t1  + I(g2_k3 * t1) + I(g3_k3 * t1) + t12 + I(g3_k3 * t12)  +
                t2  + I(g2_k3 * t2) + I(g3_k3 * t2) + t22 + I(g2_k3 * t22) + I(g3_k3 * t22),
                 random = ~ 1 +t1 +t2 | PatientId,
                control=lmeControl(returnObject=TRUE),
                 na.action = na.omit,
                 method = "ML",
                 data = bc_viremic_long)

anova(k3_BC06_mod , k3_BC06)
```

B.3.3 Generic plots for model diagnostics

```
moi = [Model of interest]

residplot(moi)
hist(resid(moi, type="pearson"), breaks=100,
          main="Histogramm for the conditional residuals",
          xlab="Standardized residuals")
```