



**HAL**  
open science

## Between group comparison of AUC in clinical trials with censored follow-up: Application to HIV therapeutic vaccines

Marie Alexandre, Mélanie Prague, Rodolphe Thiébaud

### ► To cite this version:

Marie Alexandre, Mélanie Prague, Rodolphe Thiébaud. Between group comparison of AUC in clinical trials with censored follow-up: Application to HIV therapeutic vaccines. *Statistical Methods in Medical Research*, 2021, 10.1177/09622802211023963 . hal-03241637v2

**HAL Id: hal-03241637**

**<https://hal.science/hal-03241637v2>**

Submitted on 15 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Between-group comparison of area under the curve in clinical trials with censored follow-up: Application to HIV therapeutic vaccines

Statistical Methods in Medical Research

0(0) 1–18

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802211023963

journals.sagepub.com/home/smm



Marie Alexandre<sup>1,2</sup> , Mélanie Prague<sup>1,2</sup> and Rodolphe Thiébaud<sup>1,2</sup>

## Abstract

In clinical trials, longitudinal data are commonly analyzed and compared between groups using a single summary statistic such as area under the outcome versus time curve (AUC). However, incomplete data, arising from censoring due to a limit of detection or missing data, can bias these analyses. In this article, we present a statistical test based on splines-based mixed-model accounting for both the censoring and missingness mechanisms in the AUC estimation. Inferential properties of the proposed method were evaluated and compared to ad hoc approaches and to a non-parametric method through a simulation study based on two-armed trial where trajectories and the proportion of missing data were varied. Simulation results highlight that our approach has significant advantages over the other methods. A real working example from two HIV therapeutic vaccine trials is presented to illustrate the applicability of our approach.

## Keywords

Area under the curve, longitudinal data, statistical test, mixed-effects model, study drop out, left-censoring

## 1 Introduction

The area under the curve (AUC) is a summary measure commonly used in various applications when the outcome of interest is based on a quantitative variable such as a biomarker concentration. In pharmacokinetics, the AUC of the drug concentration versus time is typically analyzed to account for drug exposure and clearance from the body<sup>1</sup> or to evaluate the bioequivalence of vaccines,<sup>2</sup> or the quality of life by summarizing individual scores.<sup>3–6</sup> In preclinical cancer drug screening tumor xenograft experiments, the ratio or the difference of AUC can be used to replace the commonly used treatment-to-control ratio<sup>7,8</sup> or summarize symptoms<sup>9</sup> to evaluate therapy effectiveness. In infectious diseases, the AUC can summarize the exposure to the HIV virus<sup>10</sup> or influenza.<sup>11,12</sup> When AUC is an outcome to be compared between arms in a clinical trial, estimates can be biased because of incomplete data. Two frequent sources for the lack of completeness can arise: censoring due to a limit of detection (LOD) of assay and study drop out.

In this context, various methods for the calculation of AUC have been proposed. Allisson et al.<sup>13</sup> and Venter et al.<sup>14</sup> compared different approaches based on incremental AUC. Incremental AUC consists in computing the AUC only for observations that are above a threshold, which can be viewed as particularly compelling when there

<sup>1</sup>University of Bordeaux, Inria Bordeaux Sud-Ouest, Inserm, Bordeaux Population Health Research Center, SISTM Team, France

<sup>2</sup>Data Science Division, Vaccine Research Institute (VRI), Créteil, France

### Corresponding author:

Mélanie Prague, University of Bordeaux, Inria Bordeaux Sud-Ouest, Inserm, Bordeaux Population Health Research Center, SISTM Team, UMR 1219, F-33000 Bordeaux, France.

Email: melanie.prague@inria.fr

is left-censored observations. However, Potteiger et al.<sup>15</sup> pointed out the potential bias in resulting conclusions when using incremental AUC even in presence of complete data. Wilding et al.<sup>16</sup> have developed an approach to evaluate treatment effect by comparing longitudinal data from two groups of patients through AUC calculation when data are subject to missing completely at random (MCAR) missingness process. Bell et al.<sup>17</sup> extended this method to missing at random (MAR) data and incorporated the within-subject variability through random effects using linear mixed effects models (LMEMs). In both cases, the comparison of the mean AUC using maximum likelihood (ML) between groups was more robust than the comparison of the average individuals' AUC with standard two-sample  $t$ -tests. Furthermore, the estimation of the mean AUC using LMEM can be adapted to outcomes subject to left-censoring.<sup>18</sup>

In this paper, we propose a statistical parametric test for AUC based on splines-based MEMs which is extending the previously described approaches by adding flexibility in the modeling, accounting for left-censored data and dealing with MAR monotonic censored follow-up. Estimation of parameters in LMEMs model is possible using ML-based approach leading to robust inference in presence of right-censored<sup>19</sup> and left-censored outcome.<sup>20,21</sup> To do so, we use an expectation-maximization EM algorithm for computing the maximum likelihood in nonlinear mixed effects models with censored response as describe in Vaida et al.<sup>22</sup>

Multiple other non-parametric approaches have been developed to solve this type of problem. Schisterman and Rotnizky<sup>23</sup> developed a semi-parametric estimator of a K-sample U-statistic when data are missing at random combining information from both outcomes and auxiliary variables. Thereafter, Spritzler et al.<sup>24</sup> extended these results by proposing a valid semi-parametric two-sample test of equal AUC when observations are MAR monotonic and/or missing completely at random (MCAR). Both works are based on weighting approaches and thus require strong assumptions on the missing data process. Alternative non parametric tests have been developed by Vardi et al.<sup>25</sup> based on permutation tests. However, parametric approaches may help in the situation of incomplete data.

This work was motivated by the evaluation of HIV therapeutic vaccine in clinical trials where high rate of censoring can occur. The goal of the vaccines in HIV-infected patients is to boost the immune system to control the viral replication when antiretroviral treatments (ART) are interrupted. Hence, analytical treatment interruption (ATI) is the ultimate way to assess the ability of new vaccine strategies to control viral replication after ART discontinuation.<sup>26</sup> However, HIV-infected patients undergoing ATIs are subject to high risks of immune damage with expansion of the existing reservoir, clinical symptoms, resistance emergence, increased risk of HIV transmission as well as loss of therapeutic benefits from ART.<sup>27,28</sup> Therefore, ATI periods are short and patients are followed carefully. Specification of criteria determining ART resumption may vary from one study to another: development of Grade-3 adverse events or AIDS-related events, the CD4 cell count fell below 350 cells/mm<sup>3</sup>, or a HIV RNA load exceeding a given virologic threshold.<sup>29-34</sup> Following these criteria, ART resumption may occur before the end of the planned ATI period leading to missing data comparable to study drop out. Also, HIV RNA viral load is subject to left censoring due to LOD usually around 50 copies/mL.<sup>20</sup> Therefore, the comparison of AUC in HIV therapeutic vaccine trials constitutes a particularly relevant context for the application of the method described in the paper.

The article is structured as follows. In section 2, we briefly describe two HIV therapeutic vaccine studies which motivated the development of our ML based-model proposed approach to estimate the difference of mean AUCs between two groups of patients when observations are left-censored and subject to follow-up censoring presented in section 3. In section 4, we investigate the inferential properties of this method and compare them with both traditional methods and a non-parametric test through simulation studies. To illustrate the applicability of the approach, we provide a real working example from the two motivating examples in section 5. To conclude, we summarize the paper and propose future research in section 6.

## 2 Motivating examples

In this paper, we focus on two HIV therapeutic vaccine trials testing the efficacy of vaccines through ART interruption in HIV-1-infected patients. The first one is the HIV therapeutic vaccine trial VRI02 ANRS 149 LIGHT.<sup>35</sup> This study is a randomized double-blind, two-arm placebo-controlled Phase-II trial. Its primary objective was to evaluate the virological efficacy after ART interruption of a therapeutic immunization compared to a placebo. The therapeutic immunization is based on a recombinant DNA vaccine (GTU-MultiHIV B) and a lipopeptide vaccine (LIPO-5). This study enrolled 105 patients (35 in the placebo control group vs. 70 in the

vaccinated group) whose 91 of them (32 placebo and 59 vaccinated) experienced ATI. HIV RNA load was repeatedly measured at times 0, 2, 4, 6, 8 and 12 weeks after ATI. The second study is the HIV therapeutic vaccine trial ANRS 093 Vac-IL2 (Vac-IL2).<sup>36</sup> This study is a randomized two-arm placebo-controlled Phase-II trial enrolling 71 patients (37 in the control group and 34 in the vaccinated group). Its primary objective was to evaluate the immunogenicity of a therapeutic immunization strategy combining two different vaccines, recombinant ALVAC-HIV (vCP1433) and Lipo-6T (HIV-1 lipopeptides), followed by the administration of subcutaneous interleukin-2 (IL-2). Therapeutic immunization was followed by 12 weeks of ATI with repeated measures of HIV RNA load at times 0, 1, 2, 3, 4, 6, 8, 10, 12 weeks after ATI.

### 3 Method

#### 3.1 Definition of the AUC by interpolation method

We consider  $N$  subjects divided into  $G$  vaccine arms, with  $N = \sum_{g=1}^G n_g$ , with  $n_g$  being the number of patient in group  $g$ . Let  $Y_{ij,g}$  be the response measured for the subject  $i$  belonging to group  $g$  at its  $j$ th time point,  $t_{ij,g}$ , with  $i \in \{1, \dots, N\}$ ,  $j \in \{1, \dots, m_i\}$  and  $g \in \{1, \dots, G\}$ . Moreover, we define  $\{t_{ij,g}\}$  as the set of time points at which data are observed for the patient  $i$  and  $m_i = |\{t_{ij,g}\}|$  the cardinal of this set. At group level, we equivalently note  $\{t_{j,g}\} = \cup_{i \in g} (\{t_{ij,g}\})$  the set of time points at which outcome of interest is measured for at least one patient in  $g$ , whose  $m_g$  is the cardinal. As defined, this framework allows the consideration of unbalanced group design and group-specific time points. The area under the response of interest curve can be calculated by the trapezoid interpolation method. The AUC summary measure for the  $i$ th subject belonging to the group  $g$  and summary statistics for the entire group  $g$  can then be approximated by the following equations. Without loss of generality, we define the lower limit of the integration interval as well as the first time point in each group as zero

$$\text{AUC}_i = \int_0^{T_i} Y_{i,g}(t) dt \simeq \sum_{j=2}^{m_i} \frac{(t_{ij,g} - t_{ij-1,g})}{2} (Y_{ij,g} + Y_{ij-1,g})$$

$$\text{AUC}_g = \int_0^{T_g} \bar{Y}_g(t) dt \simeq \sum_{j=2}^{m_g} \frac{(t_{j,g} - t_{j-1,g})}{2} (\bar{Y}_{j,g} + \bar{Y}_{j-1,g})$$

where  $\bar{Y}_{j,g}$  is defined as the mean value of the outcome  $Y$  in the  $g$ th group at its  $j$ th time point,  $\bar{Y}_{j,g} = \frac{1}{n_g} \sum_{i \in g} Y_{ij,g}$ ,  $T_i = \max_j(\{t_{ij,g}\})$  and  $T_g = \max_j(\{t_{j,g}\})$  the individual and group time of follow-up. Whereas the trapezoid method is known as the cumulative area over  $m - 1$  time period in which the value of interest  $Y$  is approximated by a straight line between two adjacent points  $(t_{j-1}, y_{j-1})$  and  $(t_j, y_j)$ , two other interpolation methods have been studied in this work to approximate AUC using either global or piecewise cubic polynomials instead of linear function: (1) the Lagrange method and (2) the Spline method (see Online Appendices A and B for more details, respectively). These methods are not described in the main body of the article as they provide similar results to the described trapezoid interpolation method.

When calculating individual's AUC, it is usual to divide the AUC by the delay of follow-up to take into account the variability in follow-up due to early drop-out for example.<sup>37-40</sup> Although we propose in this article a method based on modeling that would allow to work directly on the raw AUC, we will use a normalized AUC (nAUC), that is the AUC divided by the number of days/weeks of follow-up, for the sake of comparison with individual level methods. The nAUC are given by equations (1) and (2)

$$\text{nAUC}_i = \frac{1}{T_i} \int_0^{T_i} Y_{i,g}(t) dt \simeq \frac{1}{T_i} \sum_{j=2}^{m_i} \frac{(t_{ij,g} - t_{ij-1,g})}{2} (Y_{ij,g} + Y_{ij-1,g}) \quad (1)$$

$$\text{nAUC}_g = \frac{1}{T_g} \int_0^{T_g} \bar{Y}_g(t) dt \simeq \frac{1}{T_g} \sum_{j=2}^{m_g} \frac{(t_{j,g} - t_{j-1,g})}{2} (\bar{Y}_{j,g} + \bar{Y}_{j-1,g}) \quad (2)$$

### 3.2 Estimation of nAUC by mixed effects model

We assume the MEM given by equation (3) to describe the outcome  $Y_{ij,g}$  of the subject  $i$  in the group  $g_i$  at the  $j$ th time point

$$Y_{ij,g_i} = f_0(t_{ij,g_i}) + \sum_{g=1}^G \mathbb{1}_{[g_i=g]} \times F_g(t_{ij,g}) + h_i(t_{ij,g_i}) + \varepsilon_{ij} \quad (3)$$

where the function  $f_0$  gathers all non-group-specific terms, e.g. an intercept, the functions  $F_g$  are non-linear smooth functions of time describing the fixed effect specific to each group and  $h_i$  are polynomial time-dependent random effects modeling the inter-individual variability. In the following, the functions  $F_g$  are set to linear combinations such as  $F_g(t_{ij,g}) = \sum_{k=1}^{K_g} \beta_k^g f_k^g(t_{ij,g})$  where  $K_g$  is the number of time-dependent components describing the group-specific dynamics, e.g. spline basis, and  $\beta_k^g$  are the regression coefficients.

For generalization purpose, the LMEM given in equation (3) can be re-expressed with matrix formulation as follow

$$\mathbf{Y} = \mathbf{X}_0 \boldsymbol{\gamma} + \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{b} + \boldsymbol{\varepsilon}$$

where  $\mathbf{Y}$  is the vector of the outcome of interest,  $\mathbf{X}_0$ ,  $\mathbf{X}$ , and  $\mathbf{Z}$  are respectively the design matrices for the non-group- and group-specific fixed effects and random effects. Because vaccine or randomized controlled trials involve often adjustment of treatment effects on covariates, such as baseline covariates, the use of MEM allows it through the definition of the design matrices, whether at population, group or individual level. The vectors  $\boldsymbol{\gamma}$ ,  $\boldsymbol{\beta}$  and  $\mathbf{b}$  are the unknown non group- and group-specific fixed parameters and the random parameters respectively, while  $\boldsymbol{\varepsilon}$  is the vector of error terms supposedly normally distributed such as  $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$  and  $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$ . Moreover, we assume that  $\mathbb{E}(\mathbf{b}) = \mathbf{0}$  and  $\text{Var}(\mathbf{b}) = \boldsymbol{\Omega}$ , with  $\mathbf{b} \perp \boldsymbol{\varepsilon}$ . By construction, the matrix  $\mathbf{X}$  is defined as a diagonal block matrix such as  $\mathbf{X} = \text{diag}(\mathbf{X}_1, \dots, \mathbf{X}_G)$ , where each sub-matrix  $\mathbf{X}_g$  is group-specific. Similarly, the vector  $\boldsymbol{\beta}$  can be written as  $\boldsymbol{\beta}^T = (\boldsymbol{\beta}^{1T}, \dots, \boldsymbol{\beta}^{GT})$ , each vector  $\boldsymbol{\beta}^g$  being only specific to the group  $g$ . It can be demonstrated that the estimate of the nAUC in group  $g$  (2) can be re-expressed as a linear combination of the responses at each time, as

$$\text{nAUC}_g = \frac{1}{T_g} \sum_{j=1}^{m_g} w_{j,g} \bar{Y}_{j,g} = \frac{1}{T_g} \mathbf{w}_g^T \bar{\mathbf{Y}}_g \quad (4)$$

where  $\mathbf{w}_g = (w_{1,g}, \dots, w_{m_g,g})^T$ ,  $\bar{\mathbf{Y}}_g = (\bar{Y}_{1,g}, \dots, \bar{Y}_{m_g,g})^T$ , with

$$w_{j,g} = \begin{cases} \frac{t_{j+1,g} - t_{j,g}}{2}, & j = 1 \\ \frac{t_{j,g} - t_{j-1,g}}{2}, & j = m_g \\ \frac{t_{j+1,g} - t_{j-1,g}}{2}, & \text{otherwise} \end{cases} \quad (5)$$

In our method, the approximation of the summary statistics nAUC is obtained post-estimation of the MEM parameters. To this end, we denote  $\hat{\boldsymbol{\mu}}_g = \mathbb{E}(\hat{\mathbf{Y}}_g)$  being the expected value of the estimation of  $\mathbf{Y}$  in the  $g$ th group, where  $\hat{\boldsymbol{\mu}}_g = (\hat{\mu}_{1,g}, \dots, \hat{\mu}_{m_g,g})^T$  with  $\hat{\mu}_{j,g} = \mathbb{E}(\hat{Y}_{j,g})$  and  $\hat{\mathbf{Y}}_g = (\hat{Y}_{1,g}, \dots, \hat{Y}_{m_g,g})^T$ . It follows that  $\hat{\mu}_{j,g}$  is expressed as a linear combination of the fixed parameter estimates denoted  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\gamma}}$  for the group- and non-group-specific. Indeed, by noting  $\mathbf{X}_0^{[g]}$  the sub-matrix of  $\mathbf{X}_0$  corresponding to the group  $g$ , we obtain  $\hat{\boldsymbol{\mu}}_g = \mathbf{X}_0^{[g]} \hat{\boldsymbol{\gamma}} + \mathbf{X}_g \hat{\boldsymbol{\beta}}^g$  leading to

$$\hat{\mu}_{j,g} = \sum_{v=1}^{\dim(\hat{\boldsymbol{\gamma}})} X_{0jv}^{[g]} \cdot \hat{\gamma}_v + \sum_{v=1}^{K_g} X_{g jv} \cdot \hat{\beta}_v^g$$

Replacing  $\bar{\mathbf{Y}}_g$  by  $\hat{\boldsymbol{\mu}}_g$  in equation (4), the approximation of nAUC in the group  $g$ ,  $\widehat{\text{nAUC}}_g$ , can be written as

$$\widehat{\text{nAUC}}_g = \frac{1}{T_g} \mathbf{w}_g^T \hat{\boldsymbol{\mu}}_g \quad (6)$$

### 3.3 Statistical testing of difference between groups

We want to identify whether or not two groups of treatment can be differentiated by their mean value of the area under the response curve. Consequently, we defined the hypotheses of interest for the two compared groups  $g$  and  $\tilde{g}$  as the equality and the difference of their nAUC for the null hypothesis,  $H_0$  and the alternative one,  $H_1$ , respectively.

While the mechanism of follow-up censoring and the resulting missing data have no direct impact on the method of the MEM estimation, the statistical test must be written to take it into account. The presence of informative censoring impacting directly the time of follow-up and thus the time interval of AUC calculation for each group,  $[0, T_g]$ , the statistical test is build to compare the mean value of AUC on the same time interval. To do this, we define the upper integration limit for nAUC calculation as  $T = \min(T_g, T_{\tilde{g}})$  given the time restricted nAUC for each group calculated as

$$n\widehat{\text{AUC}}_g^{\text{rest}} = \frac{1}{T} \int_0^T \hat{\mu}_g(t) dt \simeq \frac{1}{T} \mathring{\omega}_g^T \hat{\mu}_g^{\text{rest}} \quad (7)$$

where  $\mathring{\omega}_g = (\omega_{1,g}, \dots, \omega_{m_g,g})^T$  and  $\hat{\mu}_g^{\text{rest}} = (\hat{\mu}_{1,g}, \dots, \hat{\mu}_{m_g,g})^T$  with  $m_g = |\{t_{j,g} | t_{j,g} \leq T\}|$ .

Based on equation (7) of the approximation of nAUC in the group  $g$ , the test hypotheses may be re-expressed in terms of model fixed parameters such as

$$H_0 : n\widehat{\text{AUC}}_g^{\text{rest}} = n\widehat{\text{AUC}}_{\tilde{g}}^{\text{rest}} \iff \frac{1}{T} \mathring{\omega}_g^T (\mathring{X}_0^{[g]} \hat{\gamma} + \mathring{X}_g \hat{\beta}^g) = \frac{1}{T} \mathring{\omega}_{\tilde{g}}^T (\mathring{X}_0^{[\tilde{g}]} \hat{\gamma} + \mathring{X}_{\tilde{g}} \hat{\beta}^{\tilde{g}}) \quad (8)$$

$$H_1 : n\widehat{\text{AUC}}_g^{\text{rest}} \neq n\widehat{\text{AUC}}_{\tilde{g}}^{\text{rest}} \iff \frac{1}{T} \mathring{\omega}_g^T (\mathring{X}_0^{[g]} \hat{\gamma} + \mathring{X}_g \hat{\beta}^g) \neq \frac{1}{T} \mathring{\omega}_{\tilde{g}}^T (\mathring{X}_0^{[\tilde{g}]} \hat{\gamma} + \mathring{X}_{\tilde{g}} \hat{\beta}^{\tilde{g}})$$

where  $(g, \tilde{g}) \in (1, \dots, G)^2, g \neq \tilde{g}$  and  $\mathring{X}_0^{[g]}$  and  $\mathring{X}_g$ , respectively, defined as  $X_0^{[g]}$  and  $X_g$  but restricted to the time interval  $[0, T]$ . Because  $\beta$  and  $\gamma$  are the parameters of a mixed model and assuming normality hypothesis, it follows that their respective maximum likelihood estimates are approximately normally distributed following the laws  $\mathcal{N}(\hat{\beta}, \widehat{\text{Var}}(\hat{\beta}))$  and  $\mathcal{N}(\hat{\gamma}, \widehat{\text{Var}}(\hat{\gamma}))$  and implies that both  $\hat{\mu}_g^{\text{rest}}$  and  $n\widehat{\text{AUC}}_g^{\text{rest}}$  are normally distributed. Let note  $\hat{\Sigma}$  the variance-covariance matrix of the estimated fixed parameters given by the inverse of the Fisher information matrix and  $\hat{\Sigma}^g$  the sub-variance covariance matrix of  $(\hat{\gamma}^T, \hat{\beta}^{gT})^T \in \mathcal{M}_{\dim(\hat{\gamma}) + K_{g,1}}(\mathbb{R})$ . By construction we obtain,  $\mathbb{E}(\hat{\mu}_g^{\text{rest}}) = \mathring{X}_0^{[g]} \hat{\gamma} + \mathring{X}_g \hat{\beta}^g$ ,  $\text{Var}(\hat{\mu}_g^{\text{rest}}) = (\mathring{X}_0^{[g]} \mathring{X}_g) \hat{\Sigma}^g (\mathring{X}_0^{[g]} \mathring{X}_g)^T$  and  $\mathbb{E}(n\widehat{\text{AUC}}_g^{\text{rest}}) = \frac{1}{T} \mathring{\omega}_g^T \mathbb{E}(\hat{\mu}_g^{\text{rest}})$ ,  $\text{Var}(n\widehat{\text{AUC}}_g^{\text{rest}}) = \frac{1}{T^2} \mathring{\omega}_g^T (\mathring{X}_0^{[g]} \mathring{X}_g) \hat{\Sigma}^g (\mathring{X}_0^{[g]} \mathring{X}_g)^T \mathring{\omega}_g$ . Consequently, the asymptotic normal distribution of the estimated difference of the restricted nAUC between the two groups can be inferred with

$$\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}} \sim \mathcal{N}\left(\mathbb{E}\left(\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}}\right), \text{Var}\left(\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}}\right)\right)$$

with  $\mathbb{E}(\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}}) = \frac{1}{T} \mathring{\omega}_{\tilde{g}}^T \mathbb{E}(\hat{\mu}_{\tilde{g}}^{\text{rest}}) - \frac{1}{T} \mathring{\omega}_g^T \mathbb{E}(\hat{\mu}_g^{\text{rest}})$  and  $\text{Var}(\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}}) = \mathring{\omega}^T (\mathring{X}_0 \mathring{X}) \hat{\Sigma} (\mathring{X}_0 \mathring{X}) \mathring{\omega}$ ,  $\mathring{\omega} \in \mathcal{M}_{m_g + m_{\tilde{g}}, 1}(\mathbb{R})$  being defined as  $\frac{1}{T} (\mathring{\theta}^T, \mathring{\omega}_g^T)^T - \frac{1}{T} (\mathring{\omega}_g^T, \mathring{\theta}^T)^T$ . For a test of the null hypothesis defined in equation (8), we can build the standard normally distributed Z-statistic given by

$$Z = \frac{\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}}}{\sqrt{\text{Var}\left(\Delta n\widehat{\text{AUC}}_{g-\tilde{g}}^{\text{rest}}\right)}}$$

Under the null hypothesis, the Z-statistics follows a  $\mathcal{N}(0, 1)$ . By weighted averaging incomplete measures, the impact of potential heteroscedasticity is reduced due to the AUC-based approach. If still variance heterogeneity

between the group occur, the Z-statistics can be modified into a Student's  $t$ -test like statistics with degree of freedom  $\tau$  (equals to  $\infty$  in case of Z-statistic). As matter of fact, in case of remaining heterogeneity, data specific to each group should be fitted with specific and independent mixed effects model. The T-statistic resulting from this procedure will differ from our Z-statistic by its standard deviation simply defined as the squared root of the sum of the variances of the group-specific nAUC, and with a degree of freedom defined by the Satterthwaite approximation<sup>41,42</sup>

$$\tau = \frac{\left( \text{Var}\left(n\widehat{\text{AUC}}_g^{\text{rest}}\right) + \text{Var}\left(n\widehat{\text{AUC}}_{\bar{g}}^{\text{rest}}\right) \right)^2}{\frac{\text{Var}\left(n\widehat{\text{AUC}}_g^{\text{rest}}\right)}{n_g-1} + \frac{\text{Var}\left(n\widehat{\text{AUC}}_{\bar{g}}^{\text{rest}}\right)}{n_{\bar{g}}-1}}$$

Similarly, in case of small sample size, our Z-test can be modified into Student's  $t$ -test with degree of freedom defined by the Kenward-Roger approximation.<sup>43</sup> Similarly to Bailer,<sup>44</sup> a  $100(1-\alpha)\%$  confidence interval for  $\Delta n\widehat{\text{AUC}}_{g-\bar{g}}^{\text{rest}}$  can be derived from the statistic, as

$$\Delta n\widehat{\text{AUC}}_{g-\bar{g}}^{\text{rest}} \pm z_{\tau, \alpha/2} \sqrt{\text{Var}\left(\Delta n\widehat{\text{AUC}}_{g-\bar{g}}^{\text{rest}}\right)}$$

where  $z_{\tau, \alpha/2}$  is the  $(1-\alpha/2)100^{\text{th}}$  percentile of the distribution.

An extension to k-sample design is straightforward deriving a one-way ANOVA testing the equality of normalized AUCs. Similarly to our Z-statistics, nAUCs are compared on the same interval of calculation  $[0, T]$  with  $T = \min_{g \in \{1, \dots, G\}}(T_g)$ .

$$\begin{cases} H_0 : n\widehat{\text{AUC}}_1^{\text{rest}} = n\widehat{\text{AUC}}_2^{\text{rest}} = \dots = n\widehat{\text{AUC}}_K^{\text{rest}} , \\ H_1 : \exists(i, j) | n\widehat{\text{AUC}}_i^{\text{rest}} \neq n\widehat{\text{AUC}}_j^{\text{rest}} \end{cases}$$

where  $K$  is the number of groups compared by the k-sample test,  $K \leq G$ . Similarly to classic one-way ANOVA, we define the statistic  $F$  following Fisher law as

$$F = \frac{\frac{SS_{\text{between}}}{K-1}}{\frac{SS_{\text{within}}}{N_K - K}} \sim F(K-1, N_K - K)$$

where  $N_K = \sum_{g=1}^K n_g$  and  $SS_{\text{between}}$  and  $SS_{\text{within}}$  define respectively the inter- and intra-group variability and are calculated as

$$SS_{\text{between}} = \sum_{g=1}^K n_g \left( n\widehat{\text{AUC}}_g^{\text{rest}} - \frac{1}{K} \sum_{k=1}^K n\widehat{\text{AUC}}_k^{\text{rest}} \right)^2$$

$$SS_{\text{within}} = \sum_{g=1}^K n_g^2 \text{Var}\left(n\widehat{\text{AUC}}_g^{\text{rest}}\right)$$

## 4 Simulation study

In this section, we conduct a simulation study to analyze the statistical properties of our approach. The simulation setting is driven by the motivating examples described in section 2.

#### 4.1 Generation of simulated data

We simulate longitudinal data mimicking a randomized HIV therapeutic vaccine trial involving two groups of treatment in which the outcome of interest is the HIV RNA load measurement. We simulated data using a LMEM as described by (9)

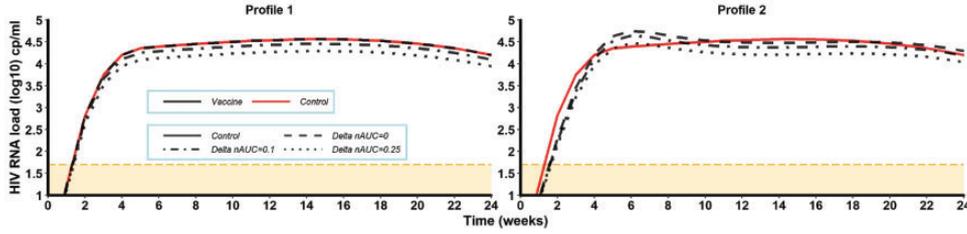
$$\begin{aligned}
 Y_{ij,g} = & \gamma_0 + \mathbb{1}_{[g=1]} \sum_{k=1}^{K_1} \beta_k^1 \phi_k^1(t_{ij,1}) + \mathbb{1}_{[g=2]} \sum_{k=1}^{K_2} \beta_k^2 \phi_k^2(t_{ij,2}) \\
 & + b_{0i} + \sum_{k=1}^{K_i} b_{ki} \Psi_k^i(t_{ij,g}) + \varepsilon_{ij}
 \end{aligned} \tag{9}$$

where  $Y_{ij,g}$  is the outcome of the  $i$ th subject belonging to the  $g$ th group at the  $j$ th time point where  $i \in \{1, \dots, n_g\}$ ,  $j \in \{1, \dots, m_g\}$  and  $g \in \{1, 2\}$ . In this model, the non-group-specific function  $f_0$  is a global intercept labeled  $\gamma_0$ , while random effects are described by individual smooth cubic B-splines curves defined as linear combination of the cubic B-spline basis  $\Psi^i = (\Psi_1^i, \dots, \Psi_{K_i}^i)^T$  with  $\mathbf{b}_i = (b_{1i}, \dots, b_{K_i i})^T$  as regression coefficients,  $\forall i \in \{1, \dots, N\}$ ,  $N = n_1 + n_2$ . Similarly, the group-specific fixed effects are modeled by cubic B-spline curves with  $\phi^g = (\phi_1^g, \dots, \phi_{K_g}^g)^T$  and  $\beta^g = (\beta_1^g, \dots, \beta_{K_g}^g)^T$  as spline basis and regression coefficients, respectively. Random effects describing the inter-individual variability are assumed to be normally distributed  $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Omega)$  as well as the error terms  $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ . Based on the HIV RNA load data from the Vac-IL2 trial (see section 2, *Motivating Examples*), we evaluated the regression coefficient estimates  $\gamma_0$ ,  $\beta^1$ ,  $\beta^2$  and  $\mathbf{b}$  as well as the parameters  $K_g$  and  $K_i$  being respectively the number of spline basis involved in the group-specific and individual spline curves. The model involving a global intercept  $\gamma_0$ , the splines basis have been built without including intercept terms making  $K_g$  and  $K_i$  equal to the sum of the number of internal knots and the degree (fixed at 3 in our case) of the respective spline curves.

For the purpose of examining the properties of the proposed approach developed to test the equality of nAUCs, we generate numerous vaccine trials. As illustrated in Figure 1, we simulated two types of mean trajectory profiles: one in which the timing of viral rebound is similar in control and treatment group but the magnitude of the rebound may differ, and one in which the timing of viral rebound is expected to be longer in the treatment group compared to the control group. Finally, outcomes are measured at a constant time interval such as  $t = (0, 1, 2, \dots, 24)^T$  weeks and the number of patients by group  $n = n_1 = n_2$  varied amongst 20, 50 and 100. They reproduce the trajectories found in the Vac-IL2 and LIGHT trials (see section 2, *Motivating Examples*). Based on the Vac-IL2 data, we set the values of  $\sigma_\varepsilon^2 = 0.2$ , the fixed intercept  $\gamma_0 = -0.44$  and the fixed parameters of the first group of treatment ( $g = 1$ ) seen as the control group,  $\beta^1$  (see Table 1). The five fixed parameters of the treatment group in both profiles  $\beta^2$  have been chosen such as given values of  $\Delta \text{nAUC}_{1-2}$  are targeted to specific values. To test the properties of the method, we simulated data with  $\Delta \text{nAUC}_{1-2}$  taking values of 0,  $-0.1$  and  $-0.25 \log_{10}$  cp/ml. We defined the number of fixed splines basis as  $K_1 = K_2 = 5$  for both profiles with the two internal knots fixed at (0.25, 5.62) weeks for both groups in profile 1 and (0.25, 5.62) and (3.23, 7.63) weeks in profile 2 for control and vaccine groups, respectively. Similarly, we fixed the number of random spline basis  $K_i = 5$  with (2.0, 4.5) weeks as internal knots in profile 1 and (2.0, 4.5) and (5.0, 8.0) weeks in Profile 2 for control and vaccine groups, respectively. Number and positions of internal knots have been optimally chosen on Vac-IL2 data by applying the R-package *freenknotspline*<sup>45</sup> using AIC as optimization criterion.

The covariance matrix of the random effects  $\Omega$  is defined as diagonal such as  $\Omega = \sigma_b^2 \mathbf{I}_{K_i+1}$  where the value of  $\sigma_b^2$  has been chosen according to the targeted values of  $\text{Var}(\text{nAUC}_g)$ . The estimated variances of nAUC were 0.027 and 0.021, respectively, in the control and the treatment group in Vac-IL2 trial. Hence, in simulations, we tested the impact of the intra-group variability when  $\text{Var}(\text{nAUC}_g)$  was equal to 0.02 and 0.1, in both groups.

We generated MAR monotonic missing data as follows. For each subject  $i$  at each time point  $j$ , the outcome  $Y_{ij,g}$  was labeled as missing if  $Y_{ij,g} \in \{Y_{ij,g} | \exists j' \leq j, \{Y_{ij',g} \geq \alpha\} \cap \{Y_{ij'-1,g} \geq \alpha\}\}$ , with  $\alpha$  being a fixed threshold. A patient dropped out from the trial if his/her HIV RNA load exceeded the threshold  $\alpha$  at two consecutive time points. The subsequent measurements were considered as missing. We investigated the impact of the missing data on the robustness of the method by considering three values for the threshold  $\alpha$ : 100,000 ( $5 \log_{10}$ ), 50,000 ( $\sim 4.7 \log_{10}$ ) and 10,000 ( $4 \log_{10}$ ) cp/ml. As illustrated in Figure 2 for the profile 1, the percentage of drop-out in each trial was inversely linked to the value of  $\alpha$ . Due to the difference of nAUC between the two groups, each

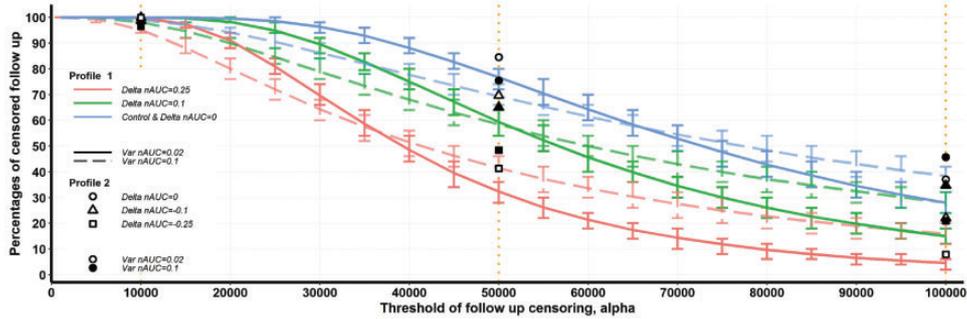


**Figure 1.** Simulated mean trajectories of HIV RNA load over time for both profiles 1 and 2. Note: Red solid line represents Group 1 (Control), dashed, dot dashed and dotted lines represent Group 2 (treatment) when  $\Delta nAUC$  with Group 1 is equal to 0,  $-0.1$  and  $0.25$ , respectively. Orange dashed line and area delimit the  $LOD = \log_{10}(50)$ . LOD: limit of detection.

**Table 1.** Fixed parameter values used to simulate control and vaccine groups for both profiles, according to  $\Delta nAUC$  values.

Treatment group	Profile 1	Profile 2
Control group, $\beta^1$	( $-0.55, 4.72, 4.96, 5.18, 4.64$ )	( $-0.55, 4.72, 4.96, 5.18, 4.64$ )
$\Delta nAUC = 0, \beta^2$	( $-0.55, 4.72, 4.96, 5.18, 4.64$ )	( $1.38, 5.57, 4.53, 5.20, 4.74$ )
$\Delta nAUC = 0.1, \beta^2$	( $-0.54, 4.61, 4.85, 5.07, 4.54$ )	( $1.35, 5.44, 4.43, 5.09, 4.63$ )
$\Delta nAUC = 0.25, \beta^2$	( $-0.52, 4.46, 4.69, 4.90, 4.39$ )	( $1.31, 5.26, 4.28, 4.92, 4.48$ )

Note: The value of the global intercept was fixed at  $\gamma_0 = -0.44$ .



**Figure 2.** Percentages of censored follow-up when data simulated by both profiles are impacted by the threshold of lost of follow-up  $\alpha$ . Note: Lines display percentages obtained for the profile 1 with solid and dashed lines representing data simulated with  $Var(nAUC) = 0.02$  and  $0.1$ , respectively. Blue lines describe both Group 1 (Control) and Group 2 (treatment) when  $\Delta nAUC$  with Group 1 is equal to 0, green and pink lines represent Group 2 when  $\Delta nAUC = 0.1$  and  $0.25$ , respectively. Marks display percentages obtained for the Profile 2 with empty and full marks representing data simulated with  $Var(nAUC) = 0.02$  and  $0.1$ , respectively. The squares, triangles and circles describe Group 2 when  $\Delta nAUC = 0, 0.1$  and  $0.25$  with the control group in blue, respectively. Vertical dotted lines highlight the positions of  $\alpha = 100,000, 50,000$  and  $10,000$  cp/ml.

value of  $\alpha$  generated both equal ( $\Delta nAUC = 0$ , blue curves) and unequal ( $\Delta nAUC \neq 0$ , blue curve for control and green/pink curves for treatment group) drop-out rates. While  $\alpha = 100,000$  cp/ml led to approximately 30% of drop-out in control group and respectively 30%, 15% and 5% in treatment group when  $\Delta nAUC = 0, 0.1$  and  $0.25$ , for  $Var(nAUC) = 0.02$ , these percentages increased respectively until 75%, 75%, 60% and 35% for  $\alpha = 50,000$ . Finally, the choice of  $\alpha = 10,000$  allowed to test the method with extremely high percentages of drop-out which were in the neighborhood of 100%. The consideration of the second profile of data simulation led to a slight increase of these percentages of approximately 7% when the variance of nAUC was equal to 0.1 and 10% for 0.02.

We also generated left-censored outcomes using the limit of detection for viral load at  $50 \sim 1.7 \log_{10}$  cp/ml, which has been chosen in accordance with values typically encountered in our motivating examples. This choice of LOD generated mean percentages of undetectable data in each group ranging from 7.30% to 7.70% for profile 1 and from 7.30% to 8.70% for profile 2, representing approximately two time points with undetectable outcome over 25.

## 4.2 Analysis of simulated data

We analyzed the simulated data using a well-specified model. Formulas for nAUC are derived from equation (9). MEM estimations took into account left-censored outcomes using an hybrid EM-algorithm implemented in the R-package *lmecc*.<sup>46</sup> Let note  $(\widehat{\gamma}_0, \widehat{\beta}^1, \widehat{\beta}^2)^T$  the vector of the estimated fixed parameters where  $\widehat{\beta}^g = (\widehat{\beta}_1^g, \dots, \widehat{\beta}_{K_g}^g)^T$ , for  $g \in \{1, 2\}$ . Using the model in equation (9), the expected value of  $Y$  in the  $g$ th group at any time  $t_{j,g}$  is  $\widehat{\mu}_{j,g} = \widehat{\gamma}_0 + \sum_{k=1}^{K_g} \widehat{\beta}_k^g \phi_k^g(t_{j,g})$ , which allows to approximate the nAUC in each group, its variance and the difference in nAUC as follows

$$\begin{aligned} n\widehat{\text{AUC}}_g &= K_{\gamma g} \widehat{\gamma}_0 + \sum_{k=1}^{K_g} \widehat{\beta}_k^g C_{kg} \\ \Delta n\widehat{\text{AUC}}_{1-2} &= \widehat{\gamma}_0 (K_{\gamma 2} - K_{\gamma 1}) + \sum_{k=1}^{K_2} \widehat{\beta}_k^2 C_{k2} - \sum_{k=1}^{K_1} \widehat{\beta}_k^1 C_{k1} \\ \text{Var}(n\widehat{\text{AUC}}_g) &= (K_{\gamma g})^2 \text{Var}(\widehat{\gamma}_0) + \sum_{k=1}^{K_g} (C_{kg})^2 \text{Var}(\widehat{\beta}_k^g) + 2 \sum_{k=1}^{K_g-1} \sum_{\bar{k}=k+1}^{K_g} C_{kg} C_{\bar{k}g} \text{Cov}(\widehat{\beta}_k^g, \widehat{\beta}_{\bar{k}}^g) \\ &\quad + 2 \sum_{k=1}^{K_g} K_{\gamma g} C_{kg} \text{Cov}(\widehat{\gamma}_0, \widehat{\beta}_k^g) \end{aligned}$$

where  $C_{kg}$  and  $K_{\gamma g}$  are defined by  $C_{kg} = \frac{1}{T_g} \sum_{j=2}^{m_g} \frac{(t_{j,g} - t_{j-1,g})}{2} [\phi_k^g(t_{j,g}) + \phi_k^g(t_{j-1,g})]$  and  $K_{\gamma g} = \frac{2}{T_g} \sum_{j=2}^{m_g} \frac{(t_{j,g} - t_{j-1,g})}{2}$ .

For each combination of simulated datasets and missing data patterns, 1000 replications were performed with the objective of evaluating the robustness of the method to test the equality of areas under the curves between the two groups through its type-I error, its power and the bias in the estimation of the difference of nAUC. We compared the results provided by our method with a standard two-sample  $t$ -test for the difference of nAUC between the two groups, i.e.  $H_0 : \overline{n\text{AUC}}_2 - \overline{n\text{AUC}}_1 = 0$  where  $\overline{n\text{AUC}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} n\text{AUC}_i$  with  $n\text{AUC}_i$  defined by equation (1). We performed this test without accounting for missing data and using two common ad hoc approaches: the last observation carried forward (LOCF) where missing data are imputed by the last observed value before the follow-up censoring, and the mean imputation where missing observations are imputed by the mean of the observations before this follow-up censoring.

In addition to the standard two-sample  $t$ -test, we compared our method with the  $t$ -test version of the non-parametric two-sample test proposed by Vardi et al.<sup>25</sup> This test was developed to compare a one-dimensional variable such as AUC between two groups of treatment when individual follow-up is subject to informative homogeneous or heterogeneous censoring. In order to be able to compare the results provided by this test and our method, we applied this test to normalized AUC. The test is based on U-statistics defined as

$$U_{m_1, m_2} = \frac{1}{m_1 m_2} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} D_{i_1, i_2}$$

where  $m_1$  and  $m_2$  are respectively the number of subjects in the first and the second compared groups,  $g_1$  and  $g_2$ , while  $D_{i_1, i_2}$  is defined as the paired cross-treatment contrast for the cross-treatment pair  $(i_1, i_2) \in g_1 \times g_2$

$$\begin{aligned} D_{i_1, i_2} &= \frac{1}{T_{i_1} \wedge T_{i_2}} \int_0^{T_{i_1} \wedge T_{i_2}} [Y_{i_2, g_2}(t) - Y_{i_1, g_1}(t)] dt \\ &= \frac{1}{T_{i_1} \wedge T_{i_2}} \left[ \text{AUC}_{i_2} \Big|_{[0, T_{i_1} \wedge T_{i_2}]} - \text{AUC}_{i_1} \Big|_{[0, T_{i_1} \wedge T_{i_2}]} \right] \end{aligned}$$

where  $T_{i_1} \wedge T_{i_2} = \min(T_{i_1}, T_{i_2})$ . The variable  $D_{i_1, i_2}$  is then defined as the difference of nAUC between the subjects  $i_1$  and  $i_2$ , restricted to their common time of follow-up. Similarly to the simulation studies conducted in their paper, we defined the variance of the U-statistic as equation (2.15) in Vardi's paper<sup>25</sup>

$$\hat{\sigma}_{m_1, m_2}^2 = \sum_{i_1=1}^{m_1} \frac{(\bar{D}_{i_1.} - \bar{D}_{..})^2}{m_1(m_1 - 1)} + \sum_{i_2=1}^{m_2} \frac{(\bar{D}_{.i_2} - \bar{D}_{..})^2}{m_2(m_2 - 1)}$$

where  $\bar{D}_{i_1.} = \sum_{i_2} D_{i_1, i_2} / m_2$ ,  $\bar{D}_{.i_2} = \sum_{i_1} D_{i_1, i_2} / m_1$  and  $\bar{D}_{..} = U_{m_1, m_2}$  and we considered the following null hypothesis  $H_0$ : *the distribution of D is symmetric about 0*.

Five procedures are then compared for testing the equality of nAUC including three ad hoc methods respectively called Indiv. nAUC Data, Indiv. nAUC LOCF and Indiv. nAUC Mean Imp., the non-parametric test called NP nAUC and our approach called MEM nAUC.

### 4.3 Simulation results

The results of our simulations evaluating the robustness of the test of equality of nAUC are displayed in Table 2. Although only results for simulations involving  $n_g = 50$  patients by group are presented in the main body of the article, extended results for  $n_g = 20$  and 100 can be found in Online Appendix C, Tables C.2a and C.3a. In these simulations, as expected with a well-specified model, when there is no censored follow-up and no left censoring using individual nAUC, non-parametric approach or our method based on MEM nAUC are identical in term of type-I error, which are kept to their nominal level of 5% (between 0.044 and 0.06). However, the power seems to be consistently higher for MEM nAUC in particular when the inter-individual variability is high. When introducing the LOD at 50 cp/ml, the results are similar for profile 1 but tend to show a superiority of MEM nAUC for profile 2 in which there are a larger number of left-censored observations due to delay in viral rebound in one group. This is explained by the fact that MEM nAUC, contrary to individual nAUC involved either in indiv. nAUC or NP nAUC methods, accounts for left censoring instead of considering censored data fixed to their censorship level value. When the threshold of HIV RNA defining drop-out,  $\alpha$ , is equal to 100,000 and 50,000 cp/ml, all individual methods (with or without adjustment for missing data) fail in term of type-I error in the second profile with lagged increasing trajectories of viral load (see Figure 1). Even when the type-I error is controlled such as for profile 1 (with the same shape of mean trajectories see Figure 1), the power for raw data and mean imputation approaches is low for most settings. While the NP nAUC method shows controlled type-I error between 0.048 and 0.057 for profiles 1 and 2 when  $\alpha$  is equal to 100,000 cp/ml and for profile 1 when the threshold is equal to 50,000 cp/ml, we observe an inflation of the type-I error up to 0.075 for the second profile. On the contrary, the MEM nAUC method shows type-I error between 0.048 and 0.064 for profiles 1 and 2. When variability is low, the power is also good and higher than 76% for the two methods. In all cases, the power found in these settings is similar in magnitude to the power obtained when there is no censored follow-up and no left censoring for viral load. When the threshold  $\alpha$  is equal to 10,000 copies/ml, while all individual methods and the non-parametric approach fail to control the type-I error for the profile 2, our approach MEM nAUC successfully gets a type-I error around the nominal value for both profiles. This result is mainly driven by the difference of the shapes of the mean trajectories for the two compared groups in Profile 2. In fact, as shown in Figure 1, the difference of nAUC appears as quite homogeneously distributed over the time of follow-up in profile 1 leading to robust results for all methods despite an early drop out for a high percentage of subjects. However, in profile 2, the value of  $\Delta nAUC$  resulting from the compensation of the beginning and the end of the dynamics, only the parametric method is able to capture the true difference of nAUC regardless of the premature censored follow-up for more than 80% of individuals.

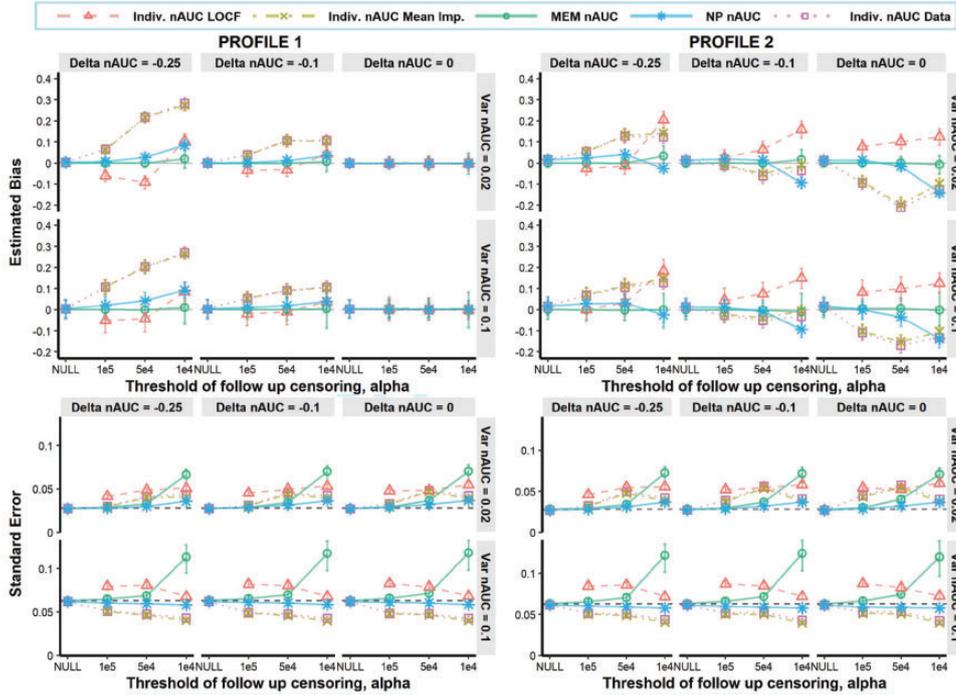
In addition, we graphically illustrated the estimated bias and standard error for  $\Delta nAUC$  obtained for each method in Figure 3. For all profiles, when there is no drop-out or when the threshold  $\alpha$  is high enough (equal to 100,000 and 50,000 cp/ml), the bias is closer to 0 for MEM nAUC compared to other methods. Also, the standard error of  $\Delta nAUC$  calculated with MEM nAUC is similar to the non-parametric approach and closer to all the ad hoc individual methods to the theoretical values of standard error of  $\Delta nAUC$ , respectively 0.028 for  $\text{Var}(nAUC) = 0.02$  and 0.063 for  $\text{Var}(nAUC) = 0.1$ . This mostly explains the comparable robustness between MEM nAUC and NP nAUC and their better performances in term of power compared to individual methods. When  $\alpha$  is equal to 10,000 cp/ml, the inflated type-I errors observed for individual and non-parametric methods

**Table 2.** Comparison of the robustness of the test of equality of nAUC calculated as individual summary measures and mixed model summary statistics.

Data pattern	Methods $\Delta nAUC$	Profile 1				Profile 2								
		Type-I error	Power											
		0	-0.1	-0.25	0	-0.1	-0.25	0	-0.1	-0.25				
LOD	$\alpha$	0.02	0.02	0.1	0.02	0.1	0.02	0.1	0.02	0.1	0.02	0.1	0.02	0.1
	Var(nAUC)	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
	Indiv. nAUC	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
	NP nAUC	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060	0.060
$\emptyset$	MEM nAUC	0.059	0.055	0.055	0.059	0.055	0.055	0.055	0.059	0.055	0.055	0.055	0.059	0.055
	Control	0	0	0	0	0	0	0	0	0	0	0	0	0
	Treatment	0	0	0	0	0	0	0	0	0	0	0	0	0
	Indiv. nAUC	0.056	0.049	0.049	0.056	0.049	0.049	0.049	0.056	0.049	0.049	0.049	0.056	0.049
50	NP nAUC	0.056	0.049	0.049	0.056	0.049	0.049	0.049	0.056	0.049	0.049	0.049	0.056	0.049
	MEM nAUC	0.063	0.053	0.053	0.063	0.053	0.053	0.053	0.063	0.053	0.053	0.053	0.063	0.053
	Control	0	0	0	0	0	0	0	0	0	0	0	0	0
	Treatment	0	0	0	0	0	0	0	0	0	0	0	0	0
Mean missing rate (%)	Indiv. nAUC	0.060	0.054	0.054	0.060	0.054	0.054	0.054	0.060	0.054	0.054	0.054	0.060	0.054
	1. Data	0.052	0.045	0.045	0.052	0.045	0.045	0.045	0.052	0.045	0.045	0.045	0.052	0.045
	2. LOCF	0.059	0.053	0.053	0.059	0.053	0.053	0.053	0.059	0.053	0.053	0.053	0.059	0.053
	3. Mean Imp.	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053
Mean missing rate (%)	NP nAUC	0.064	0.054	0.054	0.064	0.054	0.054	0.054	0.064	0.054	0.054	0.054	0.064	0.054
	MEM nAUC	28	38	38	28	38	27	38	28	38	27	38	28	38
	Control	28	38	38	28	38	27	38	28	38	27	38	28	38
	Treatment	28	38	38	28	38	5	28	28	38	5	28	28	38
Mean missing rate (%)	Indiv. nAUC	0.050	0.052	0.052	0.050	0.052	0.052	0.052	0.050	0.052	0.052	0.052	0.050	0.052
	1. Data	0.046	0.051	0.051	0.046	0.051	0.051	0.051	0.046	0.051	0.051	0.051	0.046	0.051
	2. LOCF	0.051	0.050	0.050	0.051	0.050	0.050	0.050	0.051	0.050	0.050	0.050	0.051	0.050
	3. Mean Imp.	0.048	0.053	0.053	0.048	0.053	0.053	0.053	0.048	0.053	0.053	0.053	0.048	0.053
Mean missing rate (%)	MEM nAUC	0.063	0.060	0.060	0.063	0.060	0.060	0.060	0.063	0.060	0.060	0.060	0.063	0.060
	Control	77	69	69	77	69	77	69	77	69	77	69	77	69
	Treatment	77	69	69	77	69	32	58	77	69	32	41	77	69
	Indiv. nAUC	0.041	0.057	0.057	0.041	0.057	0.057	0.057	0.041	0.057	0.057	0.057	0.041	0.057
Mean missing rate (%)	1. Data	0.058	0.043	0.043	0.058	0.043	0.043	0.043	0.058	0.043	0.043	0.043	0.058	0.043
	2. LOCF	0.039	0.050	0.050	0.039	0.050	0.050	0.050	0.039	0.050	0.050	0.050	0.039	0.050
	3. Mean Imp.	0.055	0.053	0.053	0.055	0.053	0.053	0.053	0.055	0.053	0.053	0.053	0.055	0.053
	NP nAUC	0.059	0.058	0.058	0.059	0.058	0.058	0.058	0.059	0.058	0.058	0.058	0.059	0.058
Mean missing rate (%)	MEM nAUC	100	99	99	100	99	100	99	100	99	100	99	100	99
	Control	100	99	99	100	99	100	99	100	99	100	99	100	99
	Treatment	100	99	99	100	99	100	98	100	99	100	95	100	99
	Indiv. nAUC	0.041	0.057	0.057	0.041	0.057	0.057	0.057	0.041	0.057	0.057	0.057	0.041	0.057
Mean missing rate (%)	1. Data	0.058	0.043	0.043	0.058	0.043	0.043	0.043	0.058	0.043	0.043	0.043	0.058	0.043
	2. LOCF	0.039	0.050	0.050	0.039	0.050	0.050	0.050	0.039	0.050	0.050	0.050	0.039	0.050
	3. Mean Imp.	0.055	0.053	0.053	0.055	0.053	0.053	0.053	0.055	0.053	0.053	0.053	0.055	0.053
	NP nAUC	0.059	0.058	0.058	0.059	0.058	0.058	0.058	0.059	0.058	0.058	0.058	0.059	0.058
Mean missing rate (%)	MEM nAUC	100	99	99	100	99	100	99	100	99	100	99	100	99
	Control	100	99	99	100	99	100	99	100	99	100	99	100	99
	Treatment	100	99	99	100	99	100	98	100	99	100	95	100	99
	Indiv. nAUC	0.041	0.057	0.057	0.041	0.057	0.057	0.057	0.041	0.057	0.057	0.057	0.041	0.057

AUC: area under the curve; nAUC: normalized AUC; NP: non parametric; Individual ad hoc methods (Indiv. nAUC): 1. Data: raw data, 2. LOCF: last observation carried forward, 3. Mean Imp.: mean imputation.

<sup>a</sup> Missing rate: Percentage of subjects dropping out before the end of the study.  
 Note: Individual trajectories are subject to missing data and/or limit of detection. Simulations were performed for  $n_g = 50$  subjects by group, mean trajectories following both profiles and for 1000 replications.



**Figure 3.** Comparison of the estimated bias and standard error of  $\Delta nAUC$  obtained by the three individual methods Indiv. nAUC Data, Indiv. nAUC LOCF, Indiv. nAUC Mean Imp., the non-parametric test Non Param. and our method MEM nAUC. Both criteria were estimated for data subject to a LOD, with or without censored follow-up, with  $n_g = 50$  subjects by group, mean trajectories following both profiles, for 1000 replications. Note: Pink dashed lines and triangles represent Ind. nAUC LOCF, green dot-dashed lines and crosses represent Indiv. nAUC Mean Imp., green solid lines and circles represent MEM nAUC, blue solid lines and stars represent NP nAUC and purple dotted lines with squares represent Indiv. nAUC Data. In standard Error plots, black dashed lines display the theoretical values (0.028 when  $\text{Var}(nAUC) = 0.02$  and 0.063 for 0.1); LOCF: last observation carried forward.

are explained by biased estimates of  $\Delta nAUC$  which are not compensated by an increased value of the standard error, unlike the MEM nAUC method.

#### 4.4 Relaxing the correct model specification assumption

The validity of the method relies on the correct specification of the MEM as described in equation (3) in the section Method. To relax this assumption, we conducted additional simulations to evaluate the method when data are fitted with another MEM. To evaluate the performances in a setting closer to real-data, the number and position of the knots in the MEM defined in equation (9) were also estimated with the data. We used the R-package *freenknotspline* to estimate and replace the two sets of fixed two internal knots (2.0, 4.5) and (5.0, 8.0) involved in the build of group-specific spline curves by a set of knots optimizing the fit of data. Moreover, spline basis was built with external knots chosen as  $(0, T_g)$  instead of  $(0, 24)$  considering the real observed time of follow-up, which can be modified with censored follow-up. For each simulation, the number of internal knots for a given group is optimized between 1 and 3 as well as their position using AIC as optimization criterion. Three other selection criteria have been tested: BIC, adjAIC, adjGCV and compared to AIC. Similar results of power and type-I error have been obtained for the four criteria (results not shown). Spline basis involved in random effects were similarly built chosen  $(0, T_i)$  as boundary knots and the number of internal knots chosen between 1 and 2. This adaptive feature of the model allows to build group-specific spline basis taken into account both left-censored and missing data. The results obtained by this model are displayed in Table 3 for  $n_g = 50$  subjects by group. Similar results are presented in Online Appendix C in Tables C.2b and C.3b for  $n_g = 20$  and 100, respectively.

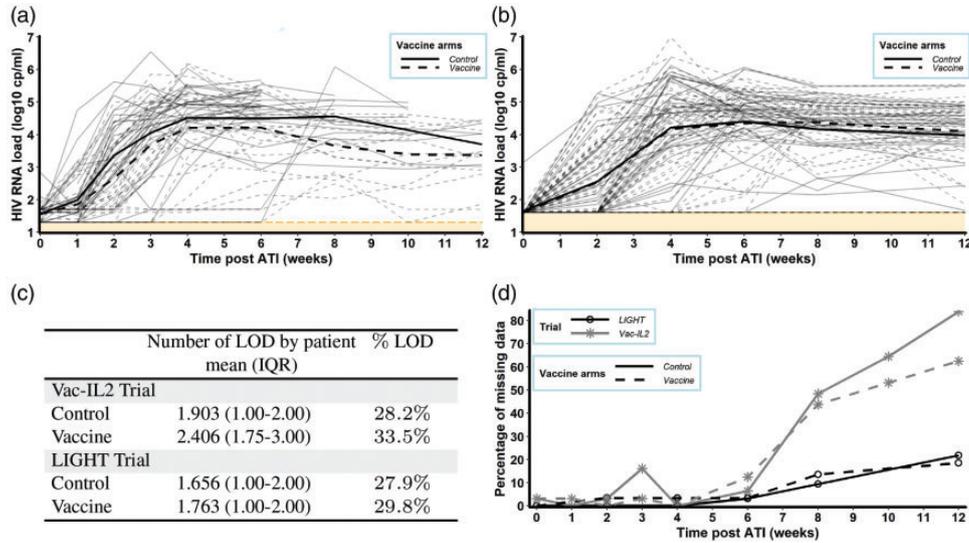
In all settings except for high level of censored follow-up with  $\alpha = 10,000$ , using adaptive MEM led to equivalent type-I error (between 0.046 and 0.063 instead of 0.044 and 0.064) and power than with the well-specified model, for both profiles. Using adaptive MEM slightly increased the type-I error when the threshold for drop-out

**Table 3.** Robustness of the test of equality of nAUC calculated as mixed model summary statistics considering the MEM (9) with adaptive spline basis.

Data pattern	Method	Profile 1			Profile 2								
		Type-I error	Power		Type-I error	Power							
		0	-0.1	-0.25	0	-0.1	-0.25						
LOD	$\alpha$	0.02	0.1	0.02	0.1	0.02	0.1	0.02	0.1	0.02	0.1		
$\emptyset$	$\emptyset$	0.060	0.060	0.41	0.96	1.00	0.99	0.046	0.054	0.95	0.42	1.00	1.00
50	$\emptyset$	0.063	0.056	0.35	0.95	1.00	0.97	0.050	0.051	0.94	0.37	1.00	0.98
	$1 \cdot 10^5$	0.060	0.054	0.33	0.94	1.00	0.96	0.050	0.060	0.91	0.35	1.00	0.97
	$5 \cdot 10^4$	0.060	0.059	0.31	0.84	1.00	0.95	0.049	0.050	0.77	0.30	1.00	0.95
	$1 \cdot 10^4$	0.070	0.061	0.17	0.31	0.89	0.54	0.061	0.078	0.26	0.17	0.83	0.54

AUC: area under the curve; nAUC: normalized AUC; LOD: limit of detection.

Note: Individual trajectories are subject to missing data and/or LOD. Simulations were performed for  $n_g = 50$  subjects by group, mean trajectories following both profiles and for 1000 replications.



**Figure 4.** Exploratory plots and table for the control and vaccine groups from the Vac-IL2 and LIGHT HIV therapeutic vaccine trials. Observations are subject to LODs of 40 cp/ml or 20 and 50 cp/ml for LIGHT and Vac-IL2 trial, respectively. LOD: limit of detection. (a) Outcome trajectories for the control and vaccine groups of the Vac-IL2 HIV therapeutic vaccine trial, with two LOD =  $\log_{10}(50)$  and  $\log_{10}(20)$  cp/ml. (b) Outcome trajectories for the control and vaccine groups of the LIGHT HIV therapeutic vaccine trial, with LOD =  $\log_{10}(40)$  cp/ml. (c) Mean number by patient and global percentage of observations below the LOD. (d) Percentage of missing data over time. Note: In (a) and (b), thick lines describe mean dynamics and thin lines individual ones, solid lines represent control group and dashed lines represent vaccine group. In (d), black lines with circles describe data from LIGHT trial, grey lines with crosses describe data from Vac-IL2 trial, solid lines represent control groups and dashed lines represent vaccine groups.

is 10,000 (between 0.061 and 0.078 instead of 0.051 and 0.073), while the estimated power remained unchanged. Altogether, even when the MEM structure is not known, this simulation shows that it is possible to use adaptive MEM for the modeling of the marker trajectories without invalidating the method, making it more relevant on real data.

## 5 Application on real clinical data

As illustrative examples, we applied the presented approach to the log-transformed HIV RNA load data from the Vac-IL2 and LIGHT trials (see section 2, Motivating Examples). Exploratory plots of the individual and mean HIV RNA load dynamics for control and vaccine groups are shown in Figure 4(a) and (b), for VAC-IL2 and LIGHT trials, respectively. As illustrated in table in Figure 4(c), longitudinal data in both trials are subject to left-censoring. While two values of LOD are considered in Vac-IL2 trial, 20 and 50 cp/ml ( $\sim 1.3$  and  $1.7 \log_{10}$  cp/ml), impacting a total of 28.2% and 33.5% of observations for control and vaccine groups, only a LOD at 40 cp/ml ( $\sim 1.6 \log_{10}$  cp/ml) is involved in LIGHT trial, leading to 27.9% and 29.8% of observations in the respective groups. In addition to left-censoring, those data are impacted by drop-outs. In LIGHT trial, ART resumption was required in case of serious AIDS or non-AIDS adverse events, when two consecutive of CD4+ T cells counted below  $350 \text{ cells/mm}^3$  within at least a two weeks' time interval as well as for specific patient or physician willingness. Approximately 20% of patients were concerned by these rules and resumed ART before the end of the predefined 12 weeks of ATI (see Figure 4(d)) being considered as drop-outs. In Vac-IL2 trial, 63% and 84% of drop-outs occurred in vaccine and control group respectively, as the result of HIV RNA load exceeding 50,000 cp/ml at four or six weeks post-ATI or exceeding 10,000 cp/ml after eight weeks of ART interruption.

We applied the proposed approach discussed in the manuscript using the MEM described by equation (9) where the number and the position of internal knots for both population and individual levels are optimized on data using the R-package *freaknotspline* and AIC criteria. Also, the structure of the covariance matrix of random effects being unknown, we estimated this matrix as unstructured instead of diagonal. Moreover, we verified the applicability of our method on these real data by checking the normality of the distribution of the residuals provided by the MEM as well as the homoscedasticity of its error model for both trials (see Online Appendix E). We compared the results obtained by our approach, where the difference of nAUC between the two groups of

**Table 4.** Summary of results from both Vac-IL2 and LIGHT studies.

Methods	Estimate (SE)	95% CI	$p$ -value	Estimate (SE)	95% CI	$p$ -value
	Vac-IL2 trial			LIGHT trial		
Data	-0.346 (0.170)	[-0.680; -0.013]	0.046	-0.030 (0.175)	[-0.312; 0.372]	0.864
LOCF	-0.382 (0.198)	[-0.770; 0.007]	0.060	-0.018 (0.186)	[-0.382; 0.346]	0.924
Mean Imp.	-0.345 (0.312)	[-0.957; 0.266]	0.276	0.217 (0.245)	[-0.263; 0.697]	0.959
NP nAUC	-0.349 (0.205)	[-0.751; 0.053]	0.089	0.042 (0.178)	[-0.306; 0.390]	0.813
Adap. MEM	-0.459 (0.213)	[-0.877; -0.041]	0.031	0.095 (0.216)	[-0.329; 0.519]	0.660

SE: standard error; CI: confidence interval; NP: non parametric; Individual ad hoc methods (Indiv. nAUC): 1. Data: raw data, 2. LOCF: last observation carried forward, 3. Mean Imp.: mean imputation.

treatment is calculated with fixed parameter estimates, with the traditional ones where the nAUC is calculated using the trapezoidal method for every individual and compared at group level with a two-sample  $t$ -test. Similarly to the study of simulated data, estimates of individual nAUCs are computed using either log-transformed raw data without any transformation, LOCF or mean imputation ad hoc approaches. In addition, we applied the non-parametric approach NP nAUC briefly defined in section 4, Simulation study. The results are gathered in Table 4. In vac-IL2, the proposed approach concluded a significant difference between the two groups of treatment with a  $p$ -value of 0.031. Similar result is obtained with raw data with  $p$ -value slightly lower than 0.05. However, both LOCF, mean imputation ad hoc methods and non-parametric method are unable to reject the null hypothesis. All the tests lead to the same conclusion of no difference between groups in the LIGHT study. Considering the mean trajectories of the control and vaccine groups displayed in Figure 4(a) and (b), all the results obtained with our new approach are consistent with expected conclusions.

## 6 Discussion

In this paper, we proposed a splines-MEM based approach to estimate and compare the normalized area under the longitudinal outcome curve when observations are subject to left-censoring, induced by an LOD, and MAR monotonic missing data, due to drop-out. We demonstrated in a simulation study that incomplete data leads to biased estimates of nAUC resulting in invalid inferences regarding the difference in nAUC between groups with individual methods even when using simple ad hoc missing data correction, such as LOCF and mean imputation. Compared to the latter, we illustrated the superiority of our approach in term of type-I error and power. In addition, although the non-parametric approach developed by Vardi et al.<sup>25</sup> provided as robust statistical properties as our proposed method while the percentages of left-censored data remained lower than 50%, corresponding to a threshold of ART resumption higher than 100,000 copies/ml, the lack of information induced by higher percentages of drop out resulted in weaker results under certain conditions of simulation and more biased estimations of the difference of nAUC. We also highlighted that when the amount of data with drop-out is as high as 80% such as in a situation when ART are resumed if HIV RNA viral load exceeds 10,000 copies/mL in ATI trial, only the parametric approach appeared efficient to compare nAUC between groups. An application of two ATI trials for HIV illustrates the superiority of our method on real data.

Limitations of the proposed method include some assumptions induced by the use of MEM such as the normality and the homoscedasticity. However, we demonstrated that on clinical data these assumptions are realistic. As briefly noticed in section 3 (*Method*), two other versions of the proposed method are presented in Online Appendix replacing the estimation of  $\Delta nAUC$  through the most commonly used trapezoid method by its estimation with either Lagrange or Spline interpolation methods. No significant differences of robustness have been observed in the application of those three methods on our well defined and tightened simulated trial designs. However, Lagrange and Splines methods could present more robust results in case of sparse designs. Also, in our simulations, we assumed a balanced longitudinal design with equal number of measurements and constant time points for every subject. Although clinical trials are commonly designed with the same monitoring for all participants, in reality the observed follow-up may deviate from the expected one. Moreover, some clinical trials could be designed to compare different monitoring designs among group in addition to treatment efficacy. As defined, the proposed method, being based on a discrete method of AUC calculation, should be biased by unbalanced times of measurements among groups with varying number of time points as well as different and irregular time steps between groups. As mentioned by Chandrasekhar et al.,<sup>18</sup> the consideration of time as continuous variable

in the AUC calculation could be a solution to handle this problem. To this end, we could either refine the time grid to mimic continuous time in the AUC calculation step, or use more complex AUC approximation methods such as Gaussian quadrature methods. The choice of Gaussian quadrature methods requires thus the use of a resampling procedure, such as bootstrapping to estimate the standard error. In clinical trials, the sample size calculation, resulting in the determination of the number of participants in each arm needed to detect a clinically relevant treatment effect, is one of the major steps in designing the study. The proposed statistics being defined as classical Z-statistics, typical formulas of sample size calculation can be derived from it. As defined by Hazra et al.,<sup>47</sup> the general formula for two-sided test can be given by  $n = (Z_{1-\alpha/2} + Z_{1-\beta})^2 \times \sigma^2 / \delta^2$  where  $\alpha$  represents the accepted type-I error,  $\beta$  the type-II error,  $\sigma$  the standard deviation of the outcome being studied and  $\delta$  the size effect defined as the targeted  $\Delta nAUC/2$  in our case. Adjusted formulas can also be derived from this latter to account for unequal sized groups or unequal variance of outcomes using pooled variances. Simulations can be found in Online Appendix (see Figure F.1 in Online Appendix F) and showed good concordance between theoretical and practical power when there is no missing data. When missing data arise due to left censoring (LOD) or informative drop out, one need to take it into account in the sample size calculation.

The simulation study has been led under model correct specification assumption, i.e. the model used to analyse the data corresponds to the true data generation process. We further relaxed this assumption by using adaptive splines model for which some parameters, such as the location and number of knots for splines are supposed unknown.

Various extensions of this work could be guided to address the problem when there are a high proportion of drop-outs. The incorporation of prior information could be done through several ways. The study of more constrained splines through the addition of penalty on spline coefficients (P-splines)<sup>48</sup> or monotony and boundary conditions<sup>49</sup> (natural splines) is an option. In the same perspective, future research aims to extend this method to the use of mechanistic models.<sup>50</sup> In addition to introducing biological interpretation of the parameters, these models could incorporate more easily additional information such as asymptotic behaviors with steady states.

## Acknowledgements

The authors thank the Vaccine Research Institute and the ANRS as sponsors and the LIGHT and VAC-IL2 study groups for sharing the data used in this article. Numerical computations were in part carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see <https://www.plafrim.fr/>). The authors thank Dr Torsten Hothorn and two anonymous reviewers for constructive comments on the manuscript. The research has been initiated in the contest of the EHVA T01 trial which is supported by the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 681032) and the Swiss Government (grant number 15 0337).

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Additional information

Web Appendix is available with this paper at the Statistical Methods in Medical Research website. R Code implementing the method is available on github at <https://github.com/marie-alexandre/AUCcomparison.git>. A reference manual has been included in the package ([https://github.com/marie-alexandre/AUCcomparison/blob/master/Reference\\_manual.pdf](https://github.com/marie-alexandre/AUCcomparison/blob/master/Reference_manual.pdf)) describing how to implement the proposed method.

## ORCID iD

Marie Alexandre  <https://orcid.org/0000-0002-3557-7075>

## References

1. Scheff JD, Almon RR, DuBois DC, et al. Assessment of pharmacologic area under the curve when baselines are variable. *Pharmaceut Res* 2011; **28**: 1081–1089.

2. Heldens J, Weststrate M and Van den Hoven R. Area under the curve calculations as a tool to compare the efficacy of equine influenza vaccines – a retrospective analysis of three independent field trials. *J Immunol Methods* 2002; **264**: 11–17.
3. Lydick E, Epstein R, Himmelberger D, et al. Area under the curve: a metric for patient subjective responses in episodic diseases. *Qual Life Res* 1995; **4**: 41–45.
4. Neoptolemos JP, Stocken DD, Friess H, et al. A randomized trial of chemoradiotherapy and chemotherapy after resection of pancreatic cancer. *N Engl J Med* 2004; **350**: 1200–1210.
5. Schleyer E, Kühn S, Rührs H, et al. Oral idarubicin pharmacokinetics – correlation of trough level with idarubicin area under curve. *Leukemia* 1996; **10**: 707–712.
6. Duh MS, Lefebvre P, Fastenau J, et al. Assessing the clinical benefits of erythropoietic agents using area under the hemoglobin change curve. *The Oncologist* 2005; **10**: 438–448.
7. Hothorn LA. Statistical analysis of in vivo anticancer experiments: tumor growth inhibition. *Drug Inform J* 2006; **40**: 229–238.
8. Wu J and Houghton PJ. Interval approach to assessing antitumor activity for tumor xenograft studies. *Pharmaceut Stat* 2010; **9**: 46–54.
9. Qian W, Parmar M, Sambrook R, et al. Analysis of messy longitudinal data from a randomized clinical trial. *Stat Med* 2000; **19**: 2657–2674.
10. Cole SR, Napravnik S, Mugavero MJ, et al. Copy-years viremia as a measure of cumulative human immunodeficiency virus viral burden. *Am J Epidemiol* 2010; **171**: 198–205.
11. Ramos EL, Mitcham JL, Koller TD, et al. Efficacy and safety of treatment with an anti-m2e monoclonal antibody in experimental human influenza. *J Infect Dis* 2015; **211**: 1038–1044.
12. Calfee DP, Peng AW, Cass LM, et al. Safety and efficacy of intravenous zanamivir in preventing experimental human influenza a virus infection. *Antimicrob Agents Chemother* 1999; **43**: 1616–1620.
13. Allison DB, Paultre F, Maggio C, et al. The use of areas under curves in diabetes research. *Diabetes Care* 1995; **18**: 245–250.
14. Venter C, Slabber M and Vorster H. Labelling of foods for glycaemic index-advantages and problems. *South African J Clin Nutr* 2003; **16**: 118–126.
15. Potteiger J, Jacobsen D and Donnelly J. A comparison of methods for analyzing glucose and insulin areas under the curve following nine months of exercise in overweight adults. *Int J Obesity* 2002; **26**: 87–89.
16. Wilding GE, Chandrasekhar R and Hutson AD. A new linear model-based approach for inferences about the mean area under the curve. *Stat Med* 2012; **31**: 3563–3578.
17. Bell ML, King MT and Fairclough DL. Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data: summary measures versus summary statistics. *SAGE Open* 2014; **4**: 2158244014534858.
18. Chandrasekhar R, Shi Y, Hutson AD, et al. Likelihood-based inferences about the mean area under a longitudinal curve in the presence of observations subject to limits of detection. *Pharmaceut Stat* 2015; **14**: 252–261.
19. Little R and An H. Robust likelihood-based analysis of multivariate data with missing values. *Stat Sin* 2004; **14**: 949–968.
20. Jacqmin-Gadda H, Thiébaud R, Chêne G, et al. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* 2000; **1**: 355–368.
21. Thiébaud R and Jacqmin-Gadda H. Mixed models for longitudinal left-censored repeated measures. *Comput Meth Progr Biomed* 2004; **74**: 255–260.
22. Vaida F and Liu L. Fast implementation for normal mixed effects models with censored response. *J Comput Graph Stat* 2009; **18**: 797–817.
23. Schisterman E and Rotnitzky A. Estimation of the mean of a k-sample u-statistic with missing outcomes and auxiliaries. *Biometrika* 2001; **88**: 713–725.
24. Spritzler J, DeGruttola VG and Pei L. Two-sample tests of area-under-the-curve in the presence of missing data. *Int J Biostat* 2008; **4**.
25. Vardi Y, Ying Z and Zhang CH. Two-sample tests for growth curves under dependent right censoring. *Biometrika* 2001; **88**: 949–960.
26. Garner SA, Rennie S, Ananworanich J, et al. Interrupting antiretroviral treatment in HIV cure research: scientific and ethical considerations. *J Virus Eradicator* 2017; **3**: 82.
27. Li JZ, Etemad B, Ahmed H, et al. The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. *AIDS* 2016; **30**: 343.
28. Henderson GE, Peay HL, Kroon E, et al. Ethics of treatment interruption trials in HIV cure research: addressing the conundrum of risk/benefit assessment. *J Med Ethics* 2018; **44**: 270–276.
29. Sneller MC, Justement JS, Gittens KR, et al. A randomized controlled safety/efficacy trial of therapeutic vaccination in HIV-infected individuals who initiated antiretroviral therapy early in infection. *Sci Transl Med* 2017; **9**: eaan8848.
30. Sued O, Ambrosioni J, Nicolás D, et al. Structured treatment interruptions and low doses of il-2 in patients with primary HIV infection. inflammatory, virological and immunological outcomes. *PLoS One* 2015; **10**: e0131651.

31. Lévy Y, Thiébaud R, Montes M, et al. Dendritic cell-based therapeutic vaccine elicits polyfunctional hiv-specific t-cell immunity associated with control of viral load. *Eur J Immunol* 2014; **44**: 2802–2810.
32. Pollard RB, Rockstroh JK, Pantaleo G, et al. Safety and efficacy of the peptide-based therapeutic vaccine for hiv-1, vacc-4x: a phase 2 randomised, double-blind, placebo-controlled trial. *Lancet Infect Dis* 2014; **14**: 291–300.
33. Bar KJ, Sneller MC, Harrison LJ et al. Effect of HIV antibody vrc01 on viral rebound after treatment interruption. *N Engl J Med* 2016; **375**: 2037–2050.
34. Fagard C, Le Braz M, Günthard H, et al. A controlled trial of granulocyte macrophage-colony stimulating factor during interruption of HAART. *AIDS* 2003; **17**: 1487–1492.
35. Palich R, Ghosn J, Chaillon A, et al. Viral rebound in semen after antiretroviral treatment interruption in an HIV therapeutic vaccine double-blind trial. *AIDS* 2019; **33**: 279–284.
36. Lévy Y, Gahéry-Ségard H, Durier C, et al. Immunological and virological efficacy of a therapeutic immunization combined with interleukin-2 in chronically hiv-1 infected patients. *AIDS* 2005; **19**: 279–286.
37. Brundage TM, Vainorius E, Chittick G, et al. Brincidofovir decreases adenovirus viral burden, which is associated with improved mortality in pediatric allogeneic hematopoietic cell transplant recipients. *Biol Blood Marrow Transplant* 2018; **24**: S372.
38. Hill JA, Mayer BT, Xie H, et al. Kinetics of double-stranded DNA viremia after allogeneic hematopoietic cell transplantation. *Clin Infect Dis* 2018; **66**: 368–375.
39. Zecca M, Wynn R, Dalle JH, et al. Association between adenovirus viral load and mortality in pediatric allo-HCT recipients: the multinational advance study. *Bone Marrow Transplant* 2019; **54**: 1632–1642.
40. Kosulin K, Pichler H, Lawitschka A, et al. Diagnostic parameters of adenoviremia in pediatric stem cell transplant recipients. *Front Microbiol* 2019; **10**: 414.
41. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometr Bull* 1946; **2**: 110–114.
42. Hrong-Tai Fai A and Cornelius PL. Approximate f-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *J Stat Comput Simulat* 1996; **54**: 363–378.
43. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**: 983–997.
44. Bailer AJ. Testing for the equality of area under the curves when using destructive measurement techniques. *J Pharmacokinet Biopharmaceut* 1988; **16**: 303–309.
45. Spiriti S, Smith P and Lecuyer P. *freeknotsplines: algorithms for implementing free-knot splines*, 2018, R package version 1.0.1, <https://cran.r-project.org/web/packages/freeknotsplines/index.html>.
46. Vaida F and Liu L. *lme4: linear mixed-effects models with censored responses*, 2012. . R package version 1.0, <https://CRAN.R-project.org/package=lme4> (accessed 8 June 2021).
47. Hazra A and Gogtay N. Biostatistics series module 5: determining sample size. *Ind J Dermatol* 2016; **61**: 496.
48. Eilers PH and Marx BD. Flexible smoothing with b-splines and penalties. *Stat Sci* 1996; **11**: 89–102.
49. Laurini MP and Moura M. Constrained smoothing b-splines for the term structure of interest rates. *Insurance* 2010; **46**: 339–350.
50. Perelson AS and Ribeiro RM. Introduction to modeling viral infections and immunity. *Immunol Rev* 2018; **285**: 5.
51. Berrut JP and Trefethen LN. Barycentric Lagrange interpolation. *SIAM Rev* 2004; **46**: 501–517.