Bayesian generalized method of moments applied to pseudo-observations in survival analysis

Léa Orsini, Caroline Brard, Emmanuel Lesaffre, Guosheng Yin, David Dejardin, Gwénaël Le Teuff

BAYESIAN GENERALIZED METHOD OF MOMENTS APPLIED TO PSEUDO-OBSERVATIONS IN SURVIVAL ANALYSIS

Léa Orsini
Oncostat U1018, Inserm
University Paris-Saclay
Villejuif, France
lea.orsini@gustaveroussy.fr

Caroline Brard
Ipsen Innovation, Clinical Development Organisation
Les Ulis, France
caroline.brard@ipsen.com

Emmanuel Lesaffre
I-Biostat
KU-Leuven
Leuven, Belgium
emmanuel.lesaffre@kuleuven.be

Guosheng Yin
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong, China
gyin@hku.hk

David Dejardin
Product Development, Data Sciences
F. Hoffmann-La Roche AG
Basel, Switzerland.
david.dejardin@roche.com

Gwénaël Le Teuff
Oncostat U1018, Inserm
University Paris-Saclay
Villejuif, France
gwenael.leteuff@gustaveroussy.fr

ABSTRACT

Bayesian inference for survival regression modeling offers numerous advantages, especially for decision-making and external data borrowing, but demands the specification of the baseline hazard function, which may be a challenging task. We propose an alternative approach that does not need the specification of this function. Our approach combines pseudo-observations to convert censored data into longitudinal data with the Generalized Methods of Moments (GMM) to estimate the parameters of interest from the survival function directly. GMM may be viewed as an extension of the Generalized Estimating Equation (GEE) currently used for frequentist pseudo-observations analysis and can be extended to the Bayesian framework using a pseudo-likelihood function. We assessed the behavior of the frequentist and Bayesian GMM in the new context of analyzing pseudo-observations. We compared their performances to the Cox, GEE, and Bayesian piecewise exponential models through a simulation study of two-arm randomized clinical trials. Frequentist and Bayesian GMM gave valid inferences with similar performances compared to the three benchmark methods, except for small sample sizes and high censoring rates. For illustration, three post-hoc efficacy analyses were performed on randomized clinical trials involving patients with Ewing Sarcoma, producing results similar to those of the benchmark methods. Through a simple application of estimating hazard ratios, these findings confirm the effectiveness of this new Bayesian approach based on pseudo-observations and the generalized method of moments. This offers new insights on using pseudo-observations for Bayesian survival analysis.

Keywords Bayesian analysis · Generalized method of moments · Pseudo-observations · Survival analysis

1 Introduction

Bayesian analysis offers many benefits in pharmaceutical research for drug development and clinical trials. The Bayesian computation methods are flexible and allow for fitting every model by estimating the posterior distribution of the parameters through a sampling procedure, even when no closed-form formula is available for this particular problem. Consequently, it allows the estimation of the posterior tail probability for any given threshold that may be clinically relevant, which is particularly useful for decision-making (Held, 2020). It also provides adaptive design methods for clinical trials that are naturally suited for interim analysis. In addition, Bayesian methods are advantageous in the context of rare diseases or precision medicine, where external information can be incorporated through the prior definition (Lesaffre et al., 2024).

Despite those advantages, Bayesian survival analysis is yet rarely used in survival analysis (Chevret (2011), Brard et al. (2017)). One reason may be that contrary to the frequentist framework where the partial likelihood of the Cox proportional hazard model can be used to estimate the regression coefficients of covariates on the survival outcome from right censored data (Cox, 1972), in Bayesian inference, the baseline hazard function is usually modeled and associated with priors (Biard et al., 2021), introducing nuisance parameters in this setting. Numerous Bayesian models have been proposed using parametric distributions (exponential or Weibull) and other functions (monotone or polynomials) referenced in Ibrahim et al. (2001), Chapters 2 and 3. According to the literature review of Fors and González (2020), the most common model in randomized clinical trials is the piecewise exponential model, which assumes the baseline hazard function to be constant on intervals. More complex models have been developed using splines, which allow more flexibility (Murray et al., 2016). Non-parametric alternatives exist but involve many parameters and are computationally intensive. The Gamma process is chosen in many applications, and Cox's partial likelihood can be seen as the limiting case of this Bayesian process by allowing the prior precision to approach zero (Kalbfleisch, 1978).

Over the past twenty years, the use of pseudo-observations in the frequentist framework has become an attractive research field in survival analysis since it offers a flexible and unique framework to directly estimate quantities of interest such as the survival probability, the cumulative incidence, the transition and state-occupation probabilities in multi-states models, or the restricted mean survival time (Andersen and Pohar-Perme, 2010). Pseudo-observations are computed for a specific quantity of interest and a straightforward regression model, with pseudo-observations as outcome, is used to directly estimate the association between the covariates and this quantity. Although other approaches exist to model these different quantities, covariate adjustment is difficult with non-parametric estimators, while the assumption of fully parametric estimators may be challenged by the data (Sachs and Gabriel, 2022). Thus, by transforming (right or interval) censored data into pseudo-observations, survival analysis turns into a standard regression problem.

When traditional Bayesian survival methods involve the formulation of the full likelihood, including the specification of the baseline hazard function in the setting of regression coefficient estimation, which may be challenging due to censoring, the transformation of censored data into pseudo-observations may be advantageous in overcoming this issue. This paper presents a methodology to analyze pseudo-observations in the Bayesian framework, creating an alternative approach for Bayesian survival analysis, which does not require specifying the baseline hazard function. Currently, pseudo-observations are usually analyzed as an outcome of a generalized linear model using the Generalized Estimating Equations (GEE), introduced in Liang and Zeger (1986). This marginal approach does not involve a likelihood function and is consequently not easily translatable to the Bayesian framework. Our approach relies on the Generalized Methods of Moments (GMM) for which a Bayesian version based on a pseudo-likelihood function has been developed by Yin (2009). The GMM method has been defined by Hansen (1982) and is widely used in econometrics. GMM is defined by specifying multiple moments from the data. GMM estimates are obtained by minimizing a quadratic inference function that combines these moments.

In this paper, we assess the usefulness of the frequentist and Bayesian GMM in the particular context of estimating hazard ratios with pseudo-observations. The rest of the paper is organized as follows. Section 2 presents the theoretical aspects of pseudo-observations analysis using GEE and the innovating application of GMM to analyze pseudo-observations, comparisons of the GMM models to benchmark methods are presented through simulations in Section 3, and illustrations through real-data examples in Section 4. We conclude with some final remarks and future extensions of the proposed approach in Section 5.

2 Methods

2.1 Pseudo-observations computation to estimate hazard ratios

As previously stated, pseudo-observations have been used in many applications in survival analysis. Suppose that $T_1, ..., T_n$ are n independent and identically distributed time to event variables and $\widehat{\theta}$ is an unbiased estimator of a quantity of interest $\theta = \mathbb{E}(h(T_i))$, where h is a known function. For individual i, the pseudo-observation is calculated as:

$$\widehat{\theta}_i = n\widehat{\theta} - (n-1)\widehat{\theta}^{-i} \tag{1}$$

where $\widehat{\theta}^{-i}$ is the value of the estimator when the i-th individual is removed from the data set. From this definition, $\widehat{\theta}_i$ is an approximately unbiased estimator of θ. Pseudo-observations can be interpreted as an individual contribution to the overall estimate of the quantity of interest.

In the case of estimating hazard ratios from a proportional hazard model, the pseudo-observation of the i-th individual at time t_k is defined as:

$$y_{ik} = n\widehat{S}(t_k) - (n-1)\widehat{S}^{-i}(t_k) \tag{2}$$

where $\widehat{S}(t_k)$ is the Kaplan-Meier estimator of the survival probability at time t_k and $\widehat{S}^{-i}(t_k)$ is the Kaplan-Meier estimator of the survival probability at time t_k after removing the i-th individual from the data set.

From this definition, pseudo-observations can take values around 0 and 1 that vary over the follow-up time depending on the status (censored / uncensored) of each patient. For all the individuals at risk at time t_k, their pseudo-observation is greater than one. If one individual experiences an event, the corresponding pseudo-observations will be negative for all times after this event. If one individual is censored, the pseudo-observations will always be positive and will decrease towards 0 for all times after the last event time of the data set which has occurred before its censoring time (Andersen and Pohar-Perme, 2010). These pseudo-observations are then analyzed as an outcome variable in a generalized linear model with a cloglog link function to interpret the estimated regression coefficients as hazard ratios from a Cox model. Below is the justification for choosing this particular link function.

Since the Cox proportional hazard model with covariates $\mathbf{X_i}$ can be written as $S(t|\mathbf{X_i}) = S_0(t)^{\exp(\boldsymbol{\beta}\mathbf{X_i})}$. Applying the complementary log-log link function $g(x) = \log(-\log(x))$ to the previous equation results in

$$g(S(t|\mathbf{X_i})) = \log(H_0(t)) + \boldsymbol{\beta}\mathbf{X_i} \tag{3}$$

where $H_0(t) = -\log(S_0(t))$ is the cumulative baseline hazard. Assuming that the censoring does not depend on covariates and the event times, Graw et al. (2009) developed theoretical justifications to prove the approximate unbiasedness of pseudo-observations given the covariates (i.e. $E(y_i|\mathbf{X_i}) \approx S(t|\mathbf{X_i})$). For additional information on the theoretical proprieties of pseudo-observations, refer to the discussion in Andersen and Pohar-Perme (2010) and to Overgaard et al. (2017). Consequently, pseudo-observations can be analyzed as an outcome variable in the generalized linear model:

$$g(E(y_i|\mathbf{X_i})) = \log(H_0(t)) + \boldsymbol{\beta}\mathbf{X_i}. \tag{4}$$

We can compute the pseudo-observations not only at one time point t but at different time points $(t_k, k = 1, \ldots, K)$ for each individual. A multivariate model for $S(t_1|\mathbf{X}), \ldots, S(t_K|\mathbf{X})$ can be analyzed similarly, where $\mathbf{y_i}$ is now a K-dimensional vector since several pseudo-observations are defined for each individual. This model, extended for multiple time points, corresponds to a Cox model where the β's can be interpreted as hazard ratios.

2.2 Generalized Estimating Equations (GEE)

Andersen et al. (2003) suggest analyzing pseudo-observations as an outcome variable in a regression model using the generalized estimated equations from Liang and Zeger (1986). This marginal approach is based on quasi-likelihood functions where only the moments are defined (McCullagh and Nelder, 1991). Suppose that $\mathbf{X_i} = (\mathbf{X_{i1}}, \ldots, \mathbf{X_{iK}})^T$, $\mathbf{y_i} = (y_{i1}, \ldots, y_{iK})^T$, and $\boldsymbol{\mu_i} = (\mu_{i1}, \ldots, \mu_{iK})^T$ are the covariates matrix (of dimension $K \times P$), the outcome vector, and the mean vector for the i-th individual, respectively. The mean model is specified as:

$$E(\mathbf{y_i}|\mathbf{X_i}) = \text{cloglog}^{-1}(\beta_0 + \beta_1\mathbf{X_{i1}} + \beta_2\mathbf{X_{i2}} + \cdots + \beta_K\mathbf{X_{iK}}) \tag{5}$$

with β_0 the intercept, β_1 the treatment effect, and $\beta_2, \ldots, \beta_K$ the time effects of the $K-1$ dummy variables, derived from the indicator of the time of which the pseudo-observation is defined. In practice, $K = 5$ time points equally spaced on the event time scale are sufficient to capture all the information from the Kaplan-Meier curve (Klein et al., 2014). The coefficient of interest in this model is the treatment effect (β_1) since the $K-1$ dummy time variables only serve as adjustment variables. Although more covariates may be added to account for other explanatory features, as

illustrated in the real-data examples (see Section 4), for now, only the treatment effect is considered. Therefore, we note $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)^T$ the vector of parameters to estimate, of dimension $P = K + 1$ in this particular case.

The vector $\boldsymbol{\beta}$ is estimated by solving the score equations:

$$\mathbf{U_n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u_i}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D_i}^T \mathbf{R}^{-1}(\alpha)(\mathbf{y_i} - \boldsymbol{\mu_i}) = 0, \tag{6}$$

where $\mathbf{D_i} = \partial \boldsymbol{\mu_i}/\partial \boldsymbol{\beta}^T$ is a $K \times P$ matrix and the working correlation matrix $\mathbf{R}(\alpha)$ is assumed to be of specific forms: the two common ones are the independence form where $\mathbf{R}(\alpha)$ equals the identity matrix and the exchangeable matrix defined as 1 on the diagonal and α elsewhere.

The nuisance parameter α is estimated alternatively with $\boldsymbol{\beta}$, switching between estimating $\boldsymbol{\beta}$ for fixed values of $\widehat{\alpha}$, and estimating α for fixed values of $\widehat{\boldsymbol{\beta}}$. Using a consistent estimator of α suggested in Liang and Zeger (1986), the GEE estimator $\widehat{\boldsymbol{\beta}}$ is also consistent, even if the working correlation matrix is misspecified. When applying GEE to pseudo-observations, the working correlation matrix is usually assumed to be independent, even if pseudo-observations are correlated by definition (Klein et al., 2008). The GEE estimator converges in distribution to a normal distribution:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Gamma}), \tag{7}$$

with $\boldsymbol{\Gamma} = \lim_{n\to+\infty} \mathbf{\Gamma_0}^{-1}\mathbf{\Gamma_1}\mathbf{\Gamma_0}^{-1}$, where $\mathbf{\Gamma_0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D_i}^T\mathbf{R}^{-1}(\alpha)\mathbf{D_i}$, and $\mathbf{\Gamma_1} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D_i}^T\mathbf{R}^{-1}(\alpha)(\mathbf{y_i} - \boldsymbol{\mu_i})(\mathbf{y_i} - \boldsymbol{\mu_i})^T\mathbf{R}^{-1}(\alpha)\mathbf{D_i}$. The estimator of the $\boldsymbol{\Gamma}$ matrix, referred to as the sandwich or robust variance estimator, is obtained by evaluating the matrices $\mathbf{\Gamma_0}$ and $\mathbf{\Gamma_1}$ at their empirical estimates. Jacobsen and Martinussen (2016) have shown that this estimator is slightly conservative, but the correction of the variance proposed by the authors is numerically small. Therefore, the sandwich estimator is still commonly used (Bouaziz, 2023).

2.3 Frequentist Generalized Method of Moments

The generalized method of moments (GMM) is defined by Hansen (1982) and is widely used in econometrics contrary to biostatistics. Ziegler (1995) showed that the GMM and the GEE approaches give asymptotically equivalent estimators. The principle of GMM is to combine multiple moments through score equations. The system of equations becomes over-identified as the number of equations exceeds the number of unknown parameters. Therefore, the exact solution cannot be found anymore. The estimates are then found by minimizing an objective function defined using the score vector and a weight matrix that gives more weights to the equations with less variability.

Qu et al. (2000) proposed a GMM approach for longitudinal data with a theoretical efficiency improvement under correlation misspecification. In this particular case, only the first moment (ie. the mean model $\boldsymbol{\mu_i}$) is specified identically to GEE. This approach can be viewed as an extension of GEE since the general idea is to express the inverse of the working correlation matrix, $\mathbf{R}$, as a linear combination of J basis matrices, $\mathbf{R}^{-1} \approx \sum_{j=1}^{J} a_j \mathbf{M_j}$. The inverses of the different working correlation matrices specified in the GEE approach can be expressed as a sum of the basis matrices. For example, the inverse of the independence matrix is expressed as $\mathbf{R}^{-1} = a_1\mathbf{M_1}$, the inverse of the exchangeable matrix as $\mathbf{R}^{-1} = a_1\mathbf{M_1} + a_2\mathbf{M_2}$, where $\mathbf{M_1}$ is the identity matrix and $\mathbf{M_2}$ is the matrix with 0 on the diagonal and 1 elsewhere. The first-order auto-regressive (AR-1) working correlation matrix is defined with coefficients $r_{ij} = \alpha^{|i-j|}$, with i the line and j the column number. Its inverse $\mathbf{R}^{-1}$ can be approximated by two working correlation matrices: $\mathbf{M_1}$ is the identity, and $\mathbf{M_2}$ is the matrix with 1 on the two diagonals on both sides of the main diagonal and 0 elsewhere.

With the GMM approach, $\mathbf{u_i}(\boldsymbol{\beta})$ is now a $(J \times P)$-dimensional score vector defined as

$$\mathbf{u_i}(\boldsymbol{\beta}) = \begin{Bmatrix} \mathbf{D_i}^T\mathbf{M_1}(\mathbf{y_i} - \boldsymbol{\mu_i}) \\ \mathbf{D_i}^T\mathbf{M_2}(\mathbf{y_i} - \boldsymbol{\mu_i}) \\ \vdots \\ \mathbf{D_i}^T\mathbf{M_J}(\mathbf{y_i} - \boldsymbol{\mu_i}) \end{Bmatrix}, \tag{8}$$

and the objective function (quadratic inference function) is written as

$$Q_n(\boldsymbol{\beta}) = \mathbf{U_n}^T(\boldsymbol{\beta})\mathbf{C_n}^{-1}(\boldsymbol{\beta})\mathbf{U_n}(\boldsymbol{\beta}), \tag{9}$$

where $\mathbf{U_n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u_i}(\boldsymbol{\beta})$, and $\mathbf{C_n}(\boldsymbol{\beta}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{u_i}(\boldsymbol{\beta})\mathbf{u_i}^T(\boldsymbol{\beta})$. Contrary to GEE, the vector $\mathbf{U_n}(\boldsymbol{\beta})$ now contains more equations than unknown parameters, and the $\boldsymbol{\beta}$'s are estimated by minimizing the quadratic inference

function $\widehat{\boldsymbol{\beta}} = \operatorname{argmin}(\mathrm{Q_n}(\boldsymbol{\beta}))$. The Newton-Raphson algorithm can be used to minimize this function, with starting values usually chosen to be the least squared estimates.

A consistent variance estimator can also be derived with a sandwich form:

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\left[\{\partial \mathbf{U_n}(\widehat{\boldsymbol{\beta}})^T/\partial\boldsymbol{\beta}\}{\mathbf{C_n}}^{-1}\{\partial \mathbf{U_n}(\widehat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}^T\}\right]^{-1}. \tag{10}$$

Under some regulatory conditions, Yu, Li, and Turner (2020) have shown that this approach produces identical point estimates compared to GEE and robust covariances with an independence or exchangeable working matrix. Regarding the analysis of pseudo-observations with GMM, the mean model is identical to the one of the GEE approach, with a cloglog link function to interpret the regression coefficients as hazard ratios.

2.4 Bayesian Generalized Method of Moments

2.4.1 Model

The formulation of the Bayesian generalized method of moments can be derived by considering that the minimization problem of the GMM can be converted to a Bayesian sampling problem. By applying the Central Limit Theorem,

$$\mathbf{U_n}(\boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}(\boldsymbol{\beta})), \text{ as } n \to \infty \tag{11}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \lim_{n\to\infty} \boldsymbol{\Sigma_n}(\boldsymbol{\beta})$, then

$$Q_n(\boldsymbol{\beta}) \xrightarrow{d} \chi^2_{(J-1)\times P}. \tag{12}$$

A chi-squared test can be derived, analogous to the usual likelihood ratio test, where $Q_n(\boldsymbol{\beta})$ behaves like $-2\log L(y|\boldsymbol{\beta})$ with $L(y|\boldsymbol{\beta})$ being the likelihood function (Hansen, 1982). Thus, the GMM approximates the likelihood for selected moments of the data without specifying the full likelihood (Chernozhukov and Hong, 2003).

Given these theoretical results, Yin (2009) presented a Bayesian version of GMM by defining a pseudo-likelihood function as follows:

$$\tilde{L}(y|\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}{\mathbf{U_n}}^T(\boldsymbol{\beta}){\boldsymbol{\Sigma_n}}^{-1}(\boldsymbol{\beta})\mathbf{U_n}(\boldsymbol{\beta})\}, \tag{13}$$

with $\boldsymbol{\Sigma_n}(\boldsymbol{\beta}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{u_i}(\boldsymbol{\beta}){\mathbf{u_i}}^T(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{U_n}(\boldsymbol{\beta}){\mathbf{U_n}}^T(\boldsymbol{\beta})$. Note that ${\boldsymbol{\Sigma_n}}^{-1}(\boldsymbol{\beta})$ in the quasi-likelihood has an additional term compared to the empirical covariance matrix $\mathbf{C_n}(\boldsymbol{\beta})$ in the quadratic inference function $Q_n(\boldsymbol{\beta})$.

Yin (2009) showed the validity of the posterior distribution resulting from this pseudo-likelihood. However, this pseudo-likelihood function, $\tilde{L}(y|\boldsymbol{\beta})$, is only defined on the support of $\mathbb{R}^P$ where $\boldsymbol{\Sigma_n}$ is invertible, which is restricted due to the cloglog link function used for pseudo-observations analysis. For example, Figure 1 represents the pseudo-likelihood function as a function of the treatment effect β_1; all other parameters are fixed at their GEE estimates. The gray zone indicates the values of β_1 for which the matrix $\boldsymbol{\Sigma_n}$ is not invertible. Thus, convergence issues may occur when parameter values fall outside this local support. The inverse link function being $x \to \exp(-\exp(x))$ may result in extreme values of the parameters. In practice, when the Bayesian sampler draws a value of $\boldsymbol{\beta}$ far from the true value, $\boldsymbol{\Sigma_n}$ becomes non-invertible. Consequently, it is essential to calibrate the Bayesian algorithm well by choosing appropriate priors and starting values for each model parameter. Below, we specify (i) how to choose appropriate prior distributions and (ii) the algorithm to generate sensible starting values.

2.4.2 Choosing appropriate priors

Choosing appropriate priors for the cloglog scale partially resolves the convergence issue previously mentioned. Gelman et al. (2008) proposed to use Cauchy(0, 2.5) for all regression coefficients as default priors in generalized linear regression models after centering and re-scaling all the input variables. These weakly informative priors reflect the fact that large changes on the logit or cloglog scale are rare. Using weak Gaussian priors such as $N(0, 10)$ or $N(0, 1)$, as recommended by the Stan Development Team (2020), can provide an alternative to Cauchy priors. They may be more adapted to the pseudo-likelihood defined on a small support because they have lighter distribution tails (See Figure 1). We do not recommend using extremely vague priors, for example, $N(0, 1000)$, as they correspond to unrealistic values on the probability scale. As we estimate hazard ratios, such large priors are unreasonable. Although weak priors are more informative than flat priors, they are vague enough compared to the pseudo-likelihood (Gelman, Simpson, and Betancourt, 2017).

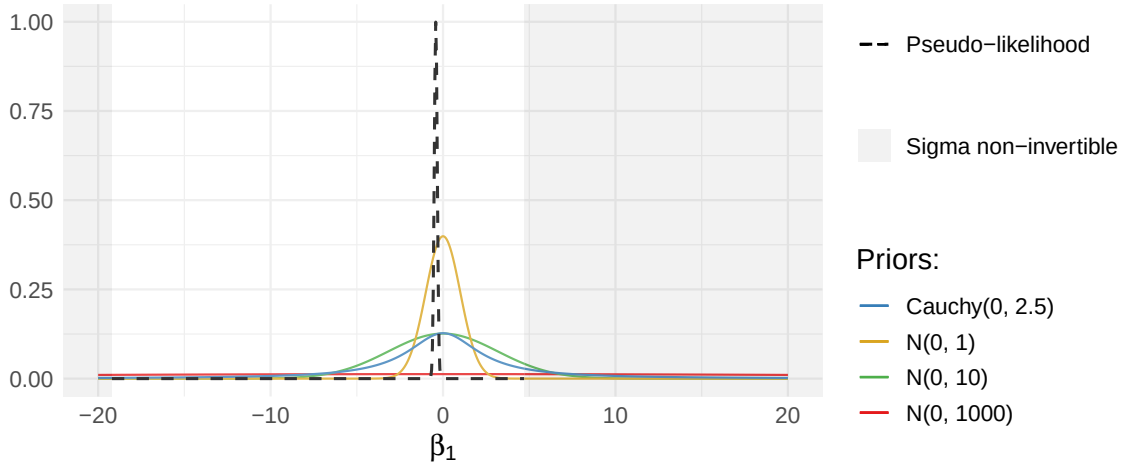

Figure 1: Example of the pseudo-likelihood function (black dashed line) depending on the treatment effect (β_1), all the other parameters are fixed to their GEE estimations. Solid lines represent different priors that have been investigated. Gray zone represents values β_1 where $\mathbf{\Sigma_n}$ is non-invertible and thus, the likelihood function is not defined.

2.4.3 Starting values

Setting starting values randomly is not optimal as they might fall outside the definition support of the pseudo-likelihood. Even if they are on the edge of the support, the $\boldsymbol{\beta}$ values may fall outside the support after a few iterations, especially during the warm-up period where the No-U-turn sampler (NUTS) chooses the step size adaptively (Hoffman and Gelman, 2014). Consequently, the step size may be too large for some iterations, resulting in poor convergence. To overcome this issue, we propose to generate starting values of the NUTS in a similar manner to the one used for the frequentist GMM, which are based on the least square estimates, while taking into account the cloglog link function. The generation of these initial values includes three steps described as follows:

1. Truncate the pseudo-observations to $[\epsilon, 1-\epsilon]$ where $\epsilon > 0$ takes a small value. Pseudo-observations above $1-\epsilon$ are set to $1-\epsilon$, and pseudo-observations below ϵ are set to ϵ. This step is needed to apply the cloglog function to pseudo-observations as it is defined on $]0, 1[$.
2. Apply the cloglog function to the truncated pseudo-observations.
3. Perform a linear regression model using these modified pseudo-observations as a continuous outcome and treatment factor and dummy time variables as covariates, then use the ordinary least square estimates as starting values.

Different values of the truncation parameter ϵ can be chosen, resulting in different starting values for each chain of the NUTS sampling. We emphasize the point that this process does not give correct estimation (1) the pseudo-observations have been truncated to $[\epsilon, 1-\epsilon]$ which may induce a bias in the estimates, (2) the cloglog is applied to the pseudo-observations themselves and not to the mean vector, and (3) the correlation between pseudo-observations of the same individual is not taken into account using the least square estimation. So, this is only a generic and straightforward process to generate starting values in the definition support of the pseudo-likelihood to improve the convergence.

3 Simulation study

We performed a simulation study to assess the performance of the frequentist and Bayesian GMM applied to pseudo-observations in order to estimate the hazard ratio of the treatment effect. The purpose of this simulation study was (a) to assess the validity of the pseudo-observations analysis using the frequentist and Bayesian GMM; (b) to compare the performances of the GMM models to the three benchmark methods: the Cox proportional hazard model, and the GEE approach based on pseudo-observations in the frequentist framework, and to the piecewise exponential model in the Bayesian framework; and (c) to evaluate the impact on the estimation of the different choices made in the pseudo-observations based models, i.e., the number of time points and the form of the working correlation matrix.

3.1 Settings

Simulations were based on a two-arm randomized clinical trial with a time-to-event outcome. Event times were generated from a Weibull distribution $f(t|a,b) = (a/b)(t/b)^{a-1}\exp((t/b)^a)$ with shape parameter $a = 0.6$ and scale parameter $b = \exp(-\frac{\beta_1 X_1}{a})$ depending on the treatment indicator X_1 coded 1 for experimental arm and 0 for control arm. This corresponds to a randomized control trial with a median survival time of approximately 6 months in the control arm for the core scenario (i.e. with $n = 500$, a censoring rate of 20%, and $\log \mathrm{HR} = -0.3$). No other explanatory variables were considered in the simulations. Censoring times were generated independently following a uniform distribution. The parameter of the uniform distribution was chosen according to the desired censoring rate, following Wan (2017). The simulation parameters were the sample size, varying from 50 to 1000, the censoring rate, ranging from 5% to 70%, and the true treatment effect varying from -0.5 to -0.1 (log scale) corresponding to hazard ratio from 0.6 to 0.9, approximately. These specifications represent different scenarios of randomized clinical trials with different sizes of the treatment effect between the experimental and control arms. Bias, average standard error (ASE), root-mean-square error (RMSE), and coverage rate from 95% equal-tailed intervals were calculated from $n_{\mathrm{sim}} = 1000$ replications for each scenario.

All computations were performed using the R Language for Statistical Computing (R Core Team (2021), version 4.1.2). Pseudo-observations have been computed using the R package `pseudo` (Pohar-Perme, Gerster, and Rodrigues, 2017), with $K = 5$ time points. Because the R package `qif` by Jiang, Song, and Kleinsasser (2019), only allows the use of the canonical links (identity, log, logit, inverse), we developed an R script to implement the frequentist GMM with a cloglog link function.

We also developed a specific script to implement the Bayesian GMM using the Stan software (Carpenter et al., 2017). The model was then compiled via the `rstan` R package, (Stan Development Team, 2023). The NUTS sampling was performed with 3 chains of each 5000 iterations after a warm-up of 1000 iterations, and thinning of 5, yielding 3000 iterations overall. As mentioned in Section 2, weakly informative priors were specified for all parameters (intercept, treatment effect, and dummy time variables). In scenarios with $n = 50$ or $n = 100$, a prior distribution of $N(0,1)$ was specified for all parameters; in all the other scenarios, $N(0,10)$ prior was specified. Initial parameters were set by fixing the truncation parameters $\epsilon \in (0.01, 0.05, 0.1)$ for the three chains, respectively. The convergence diagnoses were performed through trace plots checking and $\widehat{R}$ estimation (Vehtari, 2021).

The R package `geepack` was used to implement the GEE approach on pseudo-observations (Højsgaard et al., 2022). Multiple jackknife variance estimators are given in addition to the sandwich variance estimator. The approximate jackknife variance estimator is recommended to analyze pseudo-observations, following suggestions in Klein et al. (2008). All the estimators are equivalent for large samples, as referenced in Yan and Fine (2004). We used the `spBayesSurv` R package by Zhou, Hanson, and Zhang (2020) to implement the Bayesian piecewise exponential model. The number of intervals for the time partition was chosen according to the number of events following the rule in Murray et al. (2014) (i.e., $M = \max\{5, \min(\frac{r}{8}, 20)\}$) where M is the total number of intervals, and r is the observed number of events in the trial data set. The baseline hazard is assumed constant within each interval: $h_0(t) = \sum_{m=1}^{M} h_m I\{t \in I_m\}$. All the priors were kept as default, i.e., the priors for the baseline hazard were $h_m \sim \Gamma(1, \widehat{h})$ with $\widehat{h}$ the maximum likelihood estimate of the rate parameter from fitting an exponential proportional hazard model, and the priors for the log hazard ratio was $\beta_1 \sim N(0, 10^5)$.

3.2 Results

Table 1 represents the estimates of the hazard ratio (on a log scale) for different scenarios with a substantial treatment effect of $HR = 0.74$ ($\log HR = -0.3$), a censoring rate of 20% and different sample sizes. Overall, GMM approaches (frequentist and Bayesian) produce valid inferences with a bias that decreases toward zero as the sample size increases. From small and moderate sample sizes ($n = 50$, 100, and 200), Bayesian GMM results in slightly higher bias (varying from -0.0852 for $n = 50$ to -0.0155 for $n = 200$) compared to frequentist GMM (varying from -0.0149 to 0.0004) and similar standard errors. The coverage rates are close to the nominal coverage rate of 95% for large sample sizes ($n \geq 500$). When comparing these performances with the ones of the three benchmark models: the Cox model, the pseudo-observations-based GEE model, and the piecewise exponential model, GMM gives similar results (bias close to zero, similar standard errors and RMSEs, coverage rate close to 95%). We note, however, a slight difference for the scenarios with small sample sizes ($n \leq 100$). For these scenarios, as expected, frequentist GMM and GEE give similar results with a higher variance than the estimates of the Cox and Bayesian exponential piecewise models. For example, the average standard error is 0.257 for frequentist GMM compared to 0.228 for the Cox model for $n = 100$. This result is consistent with Andersen et al. (2003). This results in a higher RMSE for pseudo-observations-based models. The estimates from Bayesian methods are more biased than the frequentist approaches, especially for $n = 50$ with a higher bias for the Bayesian GMM (-0.0852 for Bayesian GMM, -0.0574 for piecewise exponential model versus -0.0203

for Cox). In addition, this bias decreases when the treatment effect decreases (toward 0) for the piecewise exponential model, while it remains constant for the Bayesian GMM when $n \leq 100$.

Table 1: Performances of the frequentist and Bayesian GMM compared to the Cox, GEE, and piecewise exponential model (PEM) with a true log hazard ratio of -0.3 (HR=0.74), a censoring rate of 20% and different sample sizes.

n	Methods	Bias	ASE[1]	RMSE[2]	Coverage
50	**Frequentist**				
	Cox	-0.0203	0.326	0.332	95.2
	GEE	-0.0149	0.355	0.385	92.6
	GMM	-0.0149	0.367	0.385	93.5
	Bayesian				
	PEM	-0.0574	0.332	0.367	92.7
	GMM	-0.0852	0.354	0.386	91.9
100	**Frequentist**				
	Cox	0.0054	0.228	0.245	93.2
	GEE	0.0100	0.253	0.270	93.6
	GMM	0.0100	0.257	0.270	94.0
	Bayesian				
	PEM	-0.0178	0.233	0.264	91.0
	GMM	-0.0341	0.253	0.273	92.8
200	**Frequentist**				
	Cox	0.0003	0.160	0.161	94.1
	GEE	0.0004	0.180	0.188	93.5
	GMM	0.0004	0.181	0.188	93.5
	Bayesian				
	PEM	-0.0202	0.163	0.174	93.4
	GMM	-0.0155	0.189	0.195	93.1
500	**Frequentist**				
	Cox	0.0028	0.101	0.100	95.0
	GEE	0.0032	0.114	0.112	95.4
	GMM	0.0032	0.114	0.112	95.5
	Bayesian				
	PEM	-0.0063	0.102	0.104	94.0
	GMM	-0.0028	0.116	0.113	95.4
1000	**Frequentist**				
	Cox	0.0032	0.071	0.072	94.8
	GEE	0.0005	0.081	0.082	95.2
	GMM	0.0005	0.081	0.082	95.2
	Bayesian				
	PEM	-0.0017	0.071	0.073	94.9
	GMM	-0.0026	0.082	0.082	95.2

[1] ASE = Average Standard Error

[2] RMSE = Root Mean Square Error

When varying the censoring rate for the core scenario ($\log \mathrm{HR} = -0.3$, $\mathrm{HR} = 0.74$ and $n = 500$), the performances of GMM (frequentist and Bayesian) are similar to the three benchmark methods (Table 2). The more pronounced differences between these pseudo-observation-based approaches (GEE and GMM) and the Cox and Bayesian piecewise exponential methods occur with higher average standard error and RMSE for large censoring rates (30% and 70%). For example, the average standard error of the Bayesian GMM is 0.194 compared to 0.166 for the Bayesian piecewise exponential model for a censoring rate of 70%.

Table 2: Performances of the frequentist and Bayesian GMM compared to the Cox, GEE, and piecewise exponential model (PEM) with a true log hazard ratio of -0.3 (HR=0.74), different censoring rates, and a sample size of 500.

CR[1]	Methods	Bias	ASE[2]	RMSE[3]	Coverage
5%	**Frequentist**				
	Cox	0.0016	0.093	0.094	94.0
	GEE	0.0032	0.107	0.105	95.1
	GMM	0.0032	0.107	0.105	95.1
	Bayesian				
	PEM	-0.0080	0.094	0.097	93.6
	GMM	-0.0036	0.109	0.108	95.1
10%	**Frequentist**				
	Cox	0.0013	0.095	0.096	94.4
	GEE	0.0032	0.109	0.107	95.5
	GMM	0.0032	0.109	0.107	95.5
	Bayesian				
	PEM	-0.0080	0.096	0.098	94.0
	GMM	-0.0032	0.111	0.108	95.3
20%	**Frequentist**				
	Cox	0.0028	0.101	0.100	95.0
	GEE	0.0032	0.114	0.112	95.4
	GMM	0.0032	0.114	0.112	95.5
	Bayesian				
	PEM	-0.0063	0.102	0.104	94.0
	GMM	-0.0028	0.116	0.113	95.4
30%	**Frequentist**				
	Cox	0.0039	0.108	0.107	94.8
	GEE	0.0021	0.121	0.119	95.0
	GMM	0.0021	0.121	0.119	95.2
	Bayesian				
	PEM	-0.0055	0.109	0.111	94.0
	GMM	-0.0041	0.123	0.120	95.1
70%	**Frequentist**				
	Cox	0.0018	0.165	0.165	94.3
	GEE	0.0006	0.184	0.185	94.9
	GMM	0.0006	0.185	0.185	94.9
	Bayesian				
	PEM	-0.0068	0.166	0.170	93.9
	GMM	-0.0119	0.194	0.188	95.0

[1] CR = Censoring Rate

[2] ASE = Average Standard Error

[3] RMSE = Root Mean Square Error

We evaluated the impact of different effect sizes from small to important (HR=0.90, 0.74, and 0.60) on the performances of the frequentist and Bayesian GMM approaches for the core scenario ($n = 500$, censoring rate $= 20\%$). Performances are similar to the three benchmark methods (Table 3). The supporting information Tables S1a, S1b, S2a, and S2b show the performances in all the other scenarios. For a given censoring rate and sample size, the size of the treatment effect did not affect the performances of all the models.

No convergence issue was observed through all scenarios and replicates, with $\widehat{R}$ close to 1. The Bayesian GMM was run with a parallelized code using a server HPE DL385 (2.0 GHz) with 150 virtual cores. In the core scenario (with

Table 3: Performances of the frequentist and Bayesian GMM compared to the Cox, GEE, and piecewise exponential model (PEM) with different treatment effects (log hazard ratio of -0.1 (HR=0.9), -0.3 (HR=0.74), and -0.5 (HR=0.60), a censoring rate of 20% and a sample size of 500.

HR[1]	Methods	Bias	ASE[2]	RMSE[3]	Coverage
0.9	**Frequentist**				
	Cox	0.0029	0.100	0.100	94.1
	GEE	0.0041	0.114	0.112	94.9
	GMM	0.0041	0.114	0.112	95.0
	Bayesian				
	PEM	-0.0006	0.101	0.103	94.0
	GMM	0.0012	0.116	0.113	94.9
0.74	**Frequentist**				
	Cox	0.0028	0.101	0.100	95.0
	GEE	0.0032	0.114	0.112	95.4
	GMM	0.0032	0.114	0.112	95.5
	Bayesian				
	PEM	-0.0063	0.102	0.104	94.0
	GMM	-0.0028	0.116	0.113	95.4
0.6	**Frequentist**				
	Cox	0.0034	0.102	0.101	94.8
	GEE	0.0020	0.115	0.112	95.0
	GMM	0.0020	0.115	0.112	95.1
	Bayesian				
	PEM	-0.0107	0.102	0.104	94.1
	GMM	-0.0071	0.118	0.114	95.6

[1] HR = Hazard ratio

[2] ASE = Average Standard Error

[3] RMSE = Root Mean Square Error

$n = 500$ patients), the pseudo-observations computation took less than 1 second and the median running time of one chain was 12 minutes.

3.3 Sensitivity analysis on the number of time points

When computing pseudo-observations, the choice of the time points, K, remains arbitrary. Some authors (Klein and Andersen (2005), and Andersen et al. (2003)) suggested that $K = 5$ is sufficient to obtain asymptotically unbiased estimates and, therefore, this value has been considered as the default. Intuitively, one wants to choose time points equally spaced on the event-times scale to capture most of the information from the Kaplan-Meier estimate. To assess the impact of the number of time points, we transformed survival data generated from the core scenario (a true log hazard ratio of -0.3 ($\mathrm{HR} = 0.74$), a censoring rate of 20% and a sample size of 500) into $K = 5$, 7, and 10 pseudo-observations separately. We analyzed the transformed data with the GEE, the frequentist GMM, and the Bayesian GMM models.

Figure 2 shows that increasing the number of time points had no impact on the median of the log hazard ratios estimated from 1000 replicates and a minor impact on the variability of these estimates, whatever the method. Supporting information Table S3 details the different performances of each model with different numbers of time points. Hence, our findings align with the previous sensitivity analysis from the literature. Thus, using $K > 5$ time points is not recommended as the gain in efficiency is negligible compared to the complexity induced and the increase of the running time of the NUTS algorithm for the Bayesian GMM (data not shown).

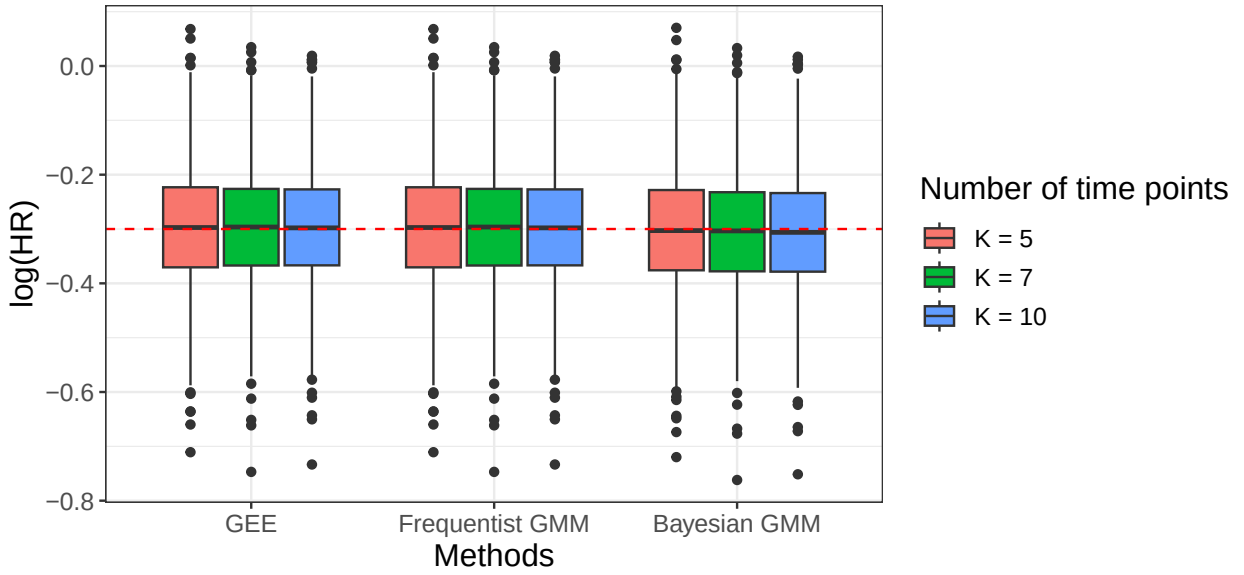

Figure 2: Sensitivity analysis on the number of time points ($K = 5$, 7, and 10) used for the computation of pseudo-observations. Box plots represent the hazard ratio (log scale) estimated from GEE, frequentist, and Bayesian GMM from 1000 replicates, with a true log hazard ratio of -0.3 ($HR = 0.74$), a censoring rate of 20% and a sample size of 500. The horizontal red line represents the true log hazard ratio.

3.4 Sensitivity analysis on the choice of the working correlation matrix

The choice of the working correlation matrix is one of the assumptions required when applying GEE or GMM. Previous results from pseudo-observations-based approaches are obtained with the independent working correlation matrix. This structure is often chosen in practice when using GEE to analyze pseudo-observations because the GEE method yields unbiased estimates even when the working correlation matrix is misspecified. The frequentist GMM has been developed to produce more efficient estimates than GEE for longitudinal continuous data when the working correlation matrix is misspecified (Qu et al., 2000). Thus, we analyzed the impact of different correlation matrices on hazard ratio estimates for the GEE, frequentist, and Bayesian GMM based on pseudo-observations. We firstly limit this analysis to the core scenario (true log hazard ratio of -0.3, censoring rate of 20% and $n = 500$). Table 4 shows that GMM approaches produce unbiased estimates whatever the working matrix and similar standard errors between the three structures. Similar results are obtained with different treatment effects (See supporting information Tables S4a and S4b). In this context, the differences in the precision of the estimations between the GEE and GMM approaches were marginal. These results concord with Yu et al. (2020), who compared the GEE and the frequentist GMM approaches to analyze longitudinal outcomes in randomized clinical trials.

4 Illustration on real-data examples

For illustration, post-hoc efficacy analyses were performed on three randomized clinical trials ($R1$, $R2_{loc}$, and $R2_{pulm}$) involving patients with Ewing Sarcoma to evaluate different consolidation treatments. After receiving intensive induction chemotherapy and surgery, patients were included in one of these trials according to prognostic factors and the response after surgery. In all these trials, the main endpoint was the event-free survival (EFS), defined as the time from random assignment to the first occurrence of any of the following events: relapse, second malignancy, or death from any cause, and the secondary endpoint was overall survival (OS), considering all causes of death.

The $R1$ trial was a phase III non-inferiority trial, which included standard-risk patients with small localized tumors or good histologic response to chemotherapy. The cyclophosphamide-based experimental arm was compared to the Ifosfamide-based control arm (Le Deley et al., 2014). This trial recruited 856 patients ($n = 431$ received Vincristine-Actinomycine-Cyclophosphamide (VAC), and $n = 425$ received Vincristine-Actinomycine-Ifosfamide (VAI)). The median follow-up was 5.9 years, and the censoring rate was 73% for the main endpoint. The $R2_{loc}$ trial, a phase III superiority trial, included high-risk patients with large localized tumors or poor histologic response. Busulfan and Melphalan (BuMel) were compared with the standard chemotherapy VAI (Whelan et al., 2018). This trial recruited

Table 4: Comparison of the performances of GEE and GMM models with different working correlation matrices: Independence (IND), Exchangeable (EXCH) and first-order auto-regressive (AR-1) for a true log hazard ratio of -0.3 (HR=0.74), a censoring rate of 20% and sample size of 500.

Methods	WCM[1]	Bias	ASE[2]	RMSE[3]	Coverage
Frequentist					
GEE	IND	0.0032	0.114	0.112	95.4
GEE	EXCH	0.0021	0.113	0.112	95.1
GEE	AR-1	0.0024	0.111	0.110	95.5
GMM	IND	0.0032	0.114	0.112	95.5
GMM	EXCH	0.0001	0.111	0.111	95.3
GMM	AR-1	-0.0017	0.111	0.112	95.3
Bayesian					
GMM	IND	-0.0028	0.116	0.113	95.4
GMM	EXCH	-0.0055	0.113	0.113	95.0
GMM	AR-1	-0.0073	0.113	0.113	94.7

[1] WCM = Working Correlation Matrix

[2] ASE = Average Standard Error

[3] RMSE = Root Mean Square Error

240 patients ($n = 122$ received BuMel and $n = 118$ received VAI). The median follow-up was 7.8 years, and the censoring rate was 56% for the main endpoint. The $R2_{pulm}$ trial, a phase III superiority trial, enrolled patients with only pulmonary or plural metastases and compared VAI + BuMel with VAI + pulmonary radiotherapy (RT) (Dirksen et al., 2019). This trial included 287 patients ($n = 144$ receiving VAI+BuMel and $n = 134$ received VAI+RT). The median follow-up was 8.1 years, and the censoring rate was 50% for the main endpoint.

The same methods and settings from the simulation study were used to analyze the EWING data. Supporting information Figure S1, shows the Kaplan-Meier curves for the three trials and the corresponding pseudo-observations profiles for all patients. Focusing on the R_1 trial, most of the events were observed between 0 and 3 years post-randomization. Consequently, the last time point to compute pseudo-observations is at 2.81 year. As most individuals are censored, we observed most of the pseudo-observations above 1 or between 0 and 1. Similar observations can be drawn from the $R2_{loc}$ and $R2_{pulm}$ trials.

Figure 3 depicts the estimates of the hazard ratios with their 95% confidence intervals for EFS and OS, produced by the frequentist and Bayesian GMM and the three benchmark methods (Cox, GEE, and piecewise exponential models) without adjustment on covariates contrary to the published results. The results of the different methods are consistent, supporting the validity of the GMM approaches for analyzing pseudo-observations. Frequentist GMM and GEE give similar results with a higher variance, as expected, compared to the Cox proportional hazard model. The results from the Bayesian GMM and piecewise exponential model are also similar, with a slightly higher variance for the former. According to the trace plots of the NUTS sampling for the Bayesian GMM, the 3 chains mixed well and appeared stationary, suggesting no divergence issue (see supporting information Figures S2a to S2f).

While the previous results are obtained with an independent working correlation matrix for GEE and GMM approaches (which is used by default for pseudo-observations analysis), our analysis was firstly extended using different correlation assumptions, as the generalized methods of moments also allow the definition of complex working correlation structures using multiple base matrices. We implemented the exchangeable (EXCH) and the first-order auto-regressive (AR-1) correlation matrices so that the results remain comparable with the GEE approach. However, the choice is not limited to these cases. Only one of the three chains did not converge for the Bayesian GMM with an exchangeable matrix for EFS. This issue was resolved by changing ϵ from 0.05 to 0.03 for generating initial values of the NUTS algorithm. Overall, the results using different working correlation matrices are similar within each approach and between approaches, except for the exchangeable matrix for situations where the treatment effect β_1 is close to 0 (log scale). The results reported in Table 5 for $R1$ trial (supporting information Tables S5a and S5b for $R2_{loc}$ and $R2_{pulm}$) suggest that increasing the complexity of the working correlation matrix gives a negligible increase in precision.

To further illustrate the versatility of our approach, the analysis of the three trials was secondly extended by including age as a covariate. Age was a stratification variable, reported as a binary variable (< 25 or ≥ 25) years in the $R1$ trial and as a four-categorical variable (< 12, $12-18$, $18-25$, and > 25) years in $R2_{loc}$ and $R2_{pulm}$ trials. The age-adjusted treatment effect is similar across the methods within a trial, with a larger variance for the Bayesian GMM in the $R2_{pulm}$

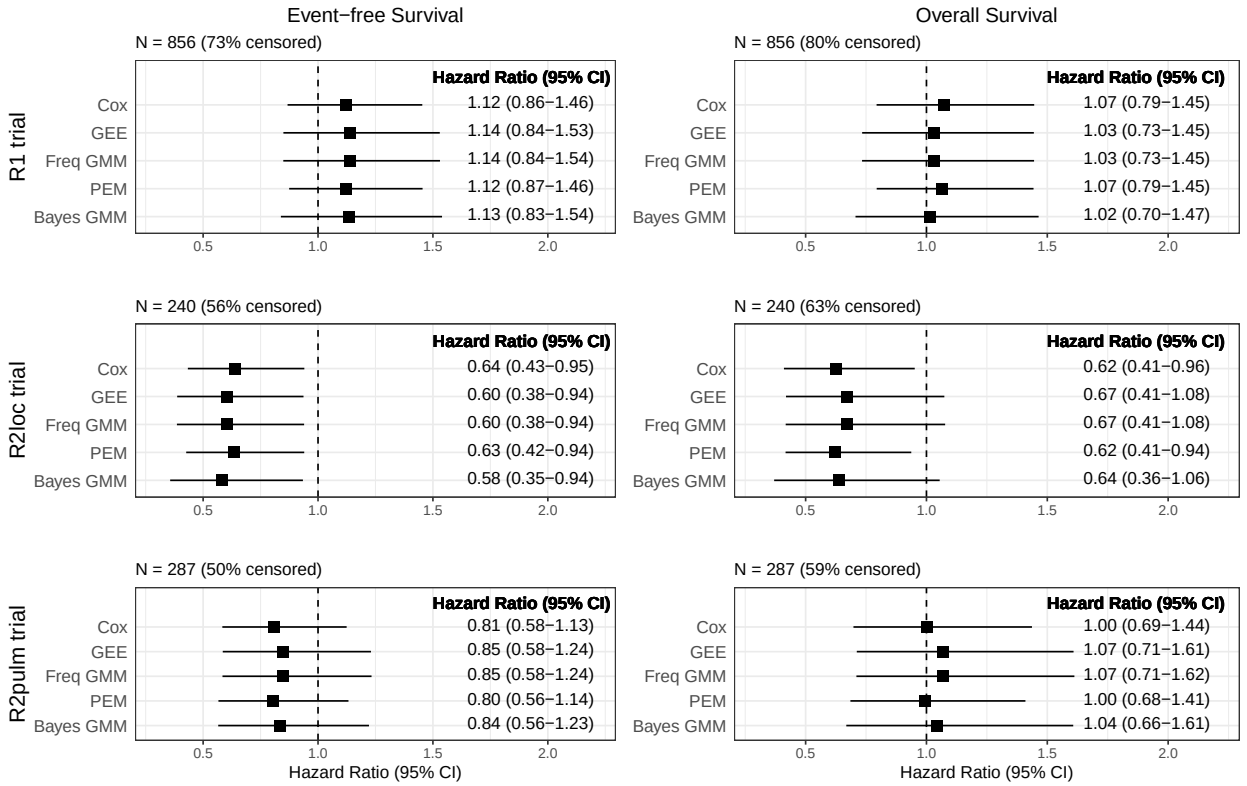

Figure 3: Hazard ratio estimates (and 95% confidence intervals) of treatment effect from the Cox proportional hazard, GEE, frequentist (Freq) GMM, piecewise exponential and Bayesian (Bayes) GMM models in the three EWING trials ($R1$, $R2_{loc}$, or $R2_{pulm}$) for event-free survival (left part) and overall survival (right part). These analyses were performed with no adjustment on covariates. The independent working correlation matrix is used for GEE and GMM approaches. The vertical dashed line represents the null effect.

trial (supporting information Figure S3). One relevant advantage of the Bayesian inference over the frequentist approach is that one can better characterize the treatment effect through its posterior distribution, estimated with the piecewise exponential model and the Bayesian GMM based on pseudo-observations (supporting information Figure S4 for EFS endpoint). For example, in $R1$ trial, the posterior probability of the log HR to be below the noninferiority margin ($\log(1.43)$) is 0.973 (se=0.002) and 0.906 (se=0.005) for piecewise exponential and GMM models, respectively. In $R2_{loc}$, the posterior probability of a log HR to be below the log HR under H_1 ($\log(0.60)$) is 0.365 (se=0.005) and 0.552 (se=0.009) for piecewise exponential and GMM models, respectively. In $R2_{pulm}$, the posterior probability of a log HR to be below the log HR under H_1 ($\log(0.65)$) is 0.091 (se=0.003) and 0.148 (se=0.006) for piecewise exponential and GMM models, respectively.

5 Discussion

In this paper, we propose a new and practical approach for Bayesian survival regression modeling based on pseudo-observations. This method does not require specifying the full likelihood, contrary to the usual Bayesian parametric and semi-parametric regression models, where nuisance parameters are specified to model the baseline hazard function. With this new approach, we bypass this specification by transforming time-to-event data into pseudo-observations, then analyzed by the generalized method of moments. The generalized method of moments applied to pseudo-observations was evaluated for estimating the hazard ratio from a two-arm randomized clinical trial in a frequentist and Bayesian framework. This approach results in valid inferences and comparable results to those produced by the Cox, GEE, and piecewise exponential models. In the frequentist framework, GMM and GEE have similar results regardless of the treatment effects, censoring rates, and sample sizes. More interestingly, in the Bayesian framework, the GMM gives unbiased results when a reasonable number of events is observed. The hazard ratio estimations are, as expected, less efficient compared to the piecewise exponential model, but remain acceptable given that the model does not require any assumption on the full likelihood contrary to the piecewise exponential model. The effect size of the treatment did not

Table 5: Log hazard ratio and standard error of treatment effect estimated by GEE and GMM with different correlation matrices: Independence (IND), Exchangeable (EXCH) and first-order auto-regressive (AR-1) in R1 trial for event-free survival and overall survival

Methods	WCM[1]	log(HR)	SE[2]
Event-free survival			
Frequentist			
GEE	IND	0.1309	0.150
GEE	EXCH	0.1015	0.148
GEE	AR-1	0.1268	0.148
GMM	IND	0.1309	0.150
GMM	EXCH	0.0819	0.149
GMM	AR-1	0.1223	0.146
Bayesian			
GMM	IND	0.1253	0.156
GMM	EXCH	0.0758	0.140
GMM	AR-1	0.1200	0.156
Overall survival			
Frequentist			
GEE	IND	0.0291	0.173
GEE	EXCH	0.0164	0.170
GEE	AR-1	0.0251	0.171
GMM	IND	0.0291	0.173
GMM	EXCH	0.0068	0.171
GMM	AR-1	0.0110	0.168
Bayesian			
GMM	IND	0.0155	0.184
GMM	EXCH	-0.0124	0.181
GMM	AR-1	0.0144	0.184

[1] WCM = Working Correlation Matrix

[2] SE = Standard Error

influence these results, and there was no inflation of the standard errors estimated by the Bayesian GMM compared to the frequentist GMM. Under the exchangeable or the first-order auto-regressive assumption, the GMM approach gave results similar to the GEE's.

Although our approach makes valid inferences, special care is needed in setting adequate priors and starting values of the regression coefficients to avoid convergence problems during the sampling. These convergence issues occur when one value of a regression coefficient falls outside the support of the pseudo-likelihood function. Priors have to be appropriately chosen according to a range of reasonable values that can be taken by the regression coefficient. In our simulation study of two-arm randomized clinical trials, we built a procedure as a preliminary step to specify the initial values of the regression coefficients carefully. So, both adequate priors and well-chosen initial values ensure the robustness of the Bayesian results. Although no additional covariate was included in the simulation study, the possibility of performing Bayesian inference with multiple covariates was reported in the real-data applications. As our approach was based on a pseudo-likelihood, it does not allow us to make predictions, while Bayesian deep learning algorithms using pseudo-observations have been proposed for this purpose (Zhao and Feng, 2020).

In conclusion, this paper proposes the first Bayesian modeling of pseudo-observations using the generalized method of moments, combining the advantages of pseudo-observations with those of Bayesian inference. The Bayesian generalized method of moments was based on pseudo-observations to estimate hazard ratios, one of the most straightforward applications of pseudo-observations. Although this application is not complex and classical Bayesian survival models may also be used, the aim of this work is to be a proof of concept that pseudo-observations can be analyzed in the Bayesian framework. It serves as a starting point before its extension to other applications where pseudo-observations are useful, such as the restricted mean survival time estimation, the estimation of transition and state probabilities derived from multi-state models, or the analysis for interval-censored data in a Bayesian framework, and unlocks various options not available with traditional proportional hazard models. Compared to the frequentist methods, this

approach offers not only new insights on interpretation via the estimated posterior distribution but may also overcome their limitations in some situations, especially in randomized clinical trials for rare diseases where it allows enriching the analysis by incorporating external data.

Additional information

The R code to analyze pseudo-observations using frequentist and Bayesian GMM is available on the Oncostat team's GitHub https://github.com/Oncostat/pseudo_gmm

Conflict of interest

The authors declare that they have no conflict of interest.

Funding

This study was funded by PhD grant MESRI from the doctoral School of Public Health, Paris-Saclay University.

References

Andersen P. K., Klein J. P., and Rosthøj S. (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1), 15–27.

Andersen, P. K., and Pohar-Perme, M. (2010) Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1), 71–99.

Biard, L., Bergeron, A., Lévy, V., and Chevret, S. (2021) Bayesian survival analysis for early detection of treatment effects in phase 3 clinical trials. *Contemporary Clinical Trials Communications*, 21, 100709.

Bouaziz, O. (2023) Fast approximations of pseudo-observations in the context of right censoring and interval censoring. *Biometrical Journal*, 65(4), 2200071.

Brard, C., Le Teuff, G., Le Deley, M.-C., and Hampson, L. V. (2017) Bayesian survival analysis in clinical trials: What methods are used in practice? *Clinical Trials*, 14(1), 78–87.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017) Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 1-32.

Chernozhukov V., and Hong, H. (2003) An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2), 293-346.

Chevret, S. (2012). Bayesian adaptive clinical trials: A dream for statisticians only? *Statistics in Medicine*, 31(11–12), 1002–1013.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.

Dirksen, U., Brennan, B., Le Deley, M.-C., Cozic, N., van den Berg, H., Bhadri, V., et al. (2019). High-Dose Chemotherapy Compared With Standard Chemotherapy and Lung Radiation in Ewing Sarcoma With Pulmonary Metastases: Results of the European Ewing Tumour Working Initiative of National Groups, 99 Trial and EWING 2008. *Journal of Clinical Oncology*, 37(34), 3192–3202.

Fors, M., and González, P. (2020) Current status of Bayesian clinical trials for oncology. *Contemporary Clinical Trials Communications*, 20, 100658.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 1360-1383.

Gelman, A., Simpson, D., and Betancourt, M. (2017) The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10), 555.

Graw, F., Gerds, T. A., and Schumacher, M. (2009) On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2), 241–255.

Hansen, L. P. (1982) Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4), 1029.

Book chapter Held, L. (2020). Bayesian Tail Probabilities for Decision Making. Lesaffre, E., Baio, G., and Boulanger, B. (Eds.). (2020). *Bayesian Methods in Pharmaceutical Research.*. Chapman and Hall/CRC.

Hoffman, M. D., and Gelman, A. (2014) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1351–1381.

Højsgaard, S., Halekoh, U., Yan, J., and Ekstrøm, C. T. (2022) *geepack* R package version 1.3.9, https://cran.r-project.org/web/packages/geepack/ (accessed on October 2022).

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001) *Bayesian survival analysis.*. Springer.

Jacobsen, M., and Martinussen, T. (2016) A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations. *Scandinavian Journal of Statistics*, 43(3), 845–862.

Jiang, Z., Song, P., and Kleinsasser, M. (2019) *qif: Quadratic Inference Function* R package version 1.5, https://CRAN.R-project.org/package=qif (Accessed on April 2023)

Kalbfleisch, J. D. (1978). Non-Parametric Bayesian Analysis of Survival Time Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2), 214–221.

Klein, J. P., and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1), 223–229.

Klein, J. P., Logan, B., Harhoff, M., and Andersen, P. K. (2007). Analyzing survival curves at a fixed point in time. *Statistics in Medicine*, 26(24), 4505–4519.

Klein, J. P., Gerster, M., Andersen, P. K., Tarima, S., and Perme, M. P. (2008) SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine*, 89(3), 289 – 300.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2014) *Handbook of survival analysis* CRC Press, Taylor and Francis Group.

Le Deley, M.-C., Paulussen, M., Lewis, I., Brennan, B., Ranft, A., Whelan, J., et al. (2014).Cyclophosphamide compared with ifosfamide in consolidation treatment of standard-risk Ewing sarcoma: Results of the randomized noninferiority Euro-EWING99-R1 trial. *Journal of the American Society of Clinical Oncology*, 32(23), 2440–2448.

Lesaffre, E., Qi, H., Banbeta, A., and Rosmalen, J. van. (2024). A review of dynamic borrowing methods with applications in pharmaceutical research. *Brazilian Journal of Probability and Statistics*, 38(1), 1 – 31.

Liang, K.-Y. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13 – 22.

McCullagh, P., and Nelder, J. A. (1991) *Generalized Linear Models (2nd ed)*. London: Chapman & Hall.

Murray, T. A., Hobbs, B. P., Lystig, T. C., and Carlin, B. P. (2014). Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data: Semiparametric Bayesian Commensurate Survival Model. *Biometrics*, 70(1), 185–191.

Murray, T. A., Hobbs, B. P., Sargent, D. J., and Carlin, B. P. (2016). Flexible Bayesian survival modeling with semiparametric time-dependent and shape-restricted covariate effects. *Bayesian Analysis*, 11(2), 381-402.

Overgaard, M., Parner, E. T., and Pedersen, J. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45(5), 1988–2015.

Pohar-Perme, M., Gerster, M., and Rodrigues, K. (2017) *pseudo: Computes Pseudo-Observations for Modeling* R package version 1.4.3 URL: https://cran.r-project.org/web/packages/pseudo/

Qu, A., Lindsay, B. G., and Li, B. (2000) Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4), 823 – 836.

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

Sachs, M. C., and Gabriel, E. E. (2022) Event History Regression with Pseudo-Observations: Computational Approaches and an Implementation in R. *Journal of Statistical Software*, 102(9), 1–34.

Stan Development Team. (2020) Prior Choice Recommendations. Available online https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations (accessed on April 2023).

Stan Development Team (2023). RStan: the R interface to Stan, R package version 2.21.7, https://mc-stan.org/ (accessed on October 2022).

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2), 667–718.

Wan, F. (2017) Simulating survival data with predefined censoring rates for proportional hazards models. *Statistics in Medicine*, 36(5), 838–854.

Whelan, J., Le Deley, M.-C., Dirksen, U., Le Teuff, G., Brennan, B., Gaspar, N., et al. (2018). High-dose chemotherapy and blood autologous stem-cell rescue compared With standard chemotherapy in localized high-risk Ewing sarcoma: results of Euro-E.W.I.N.G.99 and Ewing-2008. *Journal of Clinical Oncology*, 36(31).

Yan, J., and Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine*, 23(6), 859–874.

Yin, G. (2009) Bayesian generalized method of moments. *Bayesian Analysis*, 4(2), 191 – 208.

Yu, H., Li, F., and Turner, E. L. (2020). An evaluation of quadratic inference functions for estimating intervention effects in cluster randomized trials. *Contemporary Clinical Trials Communications*, 19, 100605.

Zhao, L., and Feng, D. (2020). Deep neural networks for survival analysis using pseudo values. *IEEE Journal of Biomedical and Health Informatics*, 24(11), 3308–3314.

Zhou, H., Hanson, T., and Zhang J. (2020) spBayesSurv: Fitting Bayesian spatial survival models using R. *Journal of Statistical Software*, 92(9), 1-33.

Ziegler, A. (1995) The different parameterizations of the GEE1 and the GEE2. *Lecture Notes in Statistics*, 104, 315–324.