Received: 2 August 2019 | Revised: 24 April 2020 | Accepted: 14 June 2020

DOI: 10.1002/pst.2047

WILEY

MAIN PAPER

Score and deviance residuals based on the full likelihood approach in survival analysis

Susan Halabi[1] | **Sandipan Dutta**[2] | **Yuan Wu**[1] | **Aiyi Liu**[3]

[1]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina

[2]Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia

[3]Biostatics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, Maryland

Correspondence
Susan Halabi, Department of Biostatistics and Bioinformatics, Duke University Medical Center, 2424 Erwin Road, Hock Plaza, Durham, NC 27705.
Email: susan.halabi@duke.edu

Funding information
National Institutes of Health, Grant/Award Number: R21 CA195424; Prostate Cancer Foundation; United States Army Medical Research, Grant/Award Numbers: W81XWH-15-1-0467, W81XWH-18-1-0278

Summary

Assuming the proportional hazards model and non-informative censoring, the full likelihood approach is used to obtain two new residuals. The first residual is based on the ideas used in obtaining score-type residuals similar to the partial likelihood approach. The second type of residual is based on the concept of deviance residuals. Extensive simulations are conducted to compare the performance of the residuals from the full likelihood-based approach with those of the partial likelihood method. We demonstrate through simulation studies that the full likelihood-based residuals are more efficient than their partial likelihood counterpart in identifying potential outliers when the censoring proportion is high. The graphical techniques are used to illustrate the applications of these residuals using some examples.

KEYWORDS

deviance residuals, full likelihood, non-informative censoring, partial likelihood, proportional hazards, score-type residuals

1 | INTRODUCTION

Residuals are important tools for checking the goodness-of-fit of a model, and identifying potential outliers and influential observations to a fitted model. A number of residuals for the Cox model[1] have been proposed for model diagnostics, identifying influential points and outliers.[2-4] These include martingale residual, deviance residual, and the score residual.[4-6] Schoenfeld's residuals (1982) are estimated through the partial likelihood function and are used to check the assumption of the proportional hazards model.[2] Martingale residuals were originally proposed to check for the overall goodness-of-fit of a proportional hazards model with respect to a data.[4] Although martingale residuals are a good measure for the overall goodness-of-fit, they suffer from lack of symmetry, making them difficult to be used for outlier detection. Therneau and Grambsch (2000) proposed deviance residual for Cox proportional hazards model in order to improve on the lack of symmetry drawback.[4] Other methods for identifying influential observations are based on the jackknife approach.[7]

The score residuals help in identifying influential or extreme observations with respect to every covariate in the fitted model, and in determining which of the covariates do not fit well in the proportional hazards model. Large magnitude of the deviance residual for an observation indicates that it is a potential outlier to the model. Similarly, large magnitude of the score residual of an individual with respect to a particular covariate indicates heavy influence of that

individual in the estimation of the regression effect of that covariate. The deviance and score residuals for the proportional hazards model are based on the partial likelihood approach.

We are interested in detecting outliers in a phase III clinical trial where the censoring proportion was high. We develop innovative score-type and deviance residuals using the full likelihood function in the proportional hazards model. We further demonstrate through simulation studies that the full likelihood-based residuals have higher area under the curve (AUC) than their partial likelihood counterpart in identifying potential outliers. The rest of this article is organized as follows. We derive the score-type and the deviance residuals based on the full likelihood function in Section 2. We conduct simulation studies in Section 3 where we generate survival data under the proportional hazards assumption and compare the performances of the different types of residuals in identifying potential outliers. We discuss the results of the simulations and we apply the full likelihood-based score-type and deviance residuals to two real datasets from the primary biliary cirrhosis and breast cancer studies. Finally, we discuss the results and implications in Section 5.

2 | METHODS

We first introduce notations and then we describe the full and the partial likelihood functions. The full likelihood function is simplified for the case of the proportional hazards model in order to obtain the score-type and deviance residuals.

2.1 | Full likelihood function

Assume that there are n individuals in the study. For the ith individual, let T_i and C_i be the failure time and censoring time, respectively. Assume further that: (a) the failure time T_i has the probability density function (p.d.f) $f(t)$ and survival function $S(t)$; (b) the censoring time has p.d.f $g(t)$ and survival function $G(t)$; and (c) the two random variables, T_i and C_i, are independent. Let δ_i be the indicator variable taking the value of 1 (or 0) if the i^{th} individual has failed (or not) and let t_i be the observed failure time $\min(T_i, C_i)$.

Assuming that the censoring is non-informative and using Lawless formula of the full likelihood function,[8] the full likelihood function is given by

$$L = \prod_{i=1}^{n} (f(t_i))^{\delta_i} (S(t_i))^{1-\delta_i} . \tag{1}$$

For the proportional hazards model, the p.d.f. and survival function are given by

$$f(t) = \lambda_0(t) e^{x'\beta} S(t) . \tag{2}$$

and

$$S(t) = e^{-\Lambda_0(t) e^{x'\beta}} . \tag{3}$$

where

$\lambda_0(t)$ is the baseline hazard function at time t with x=0,
$\Lambda_0(t)$ is the baseline cumulative hazard function at time t, where $\Lambda_0(t) = \int_0^t \lambda_0(u) du$,
$x' = (x_1, x_2, ..., x_p)$ is the vector of p covariates associated with the i^{th} individual,
$\beta' = (\beta_1, \beta_2, ..., \beta_p)$ is the vector of p regression coefficients common to all the individuals.
Then, the likelihood function (1), L, simplifies to

$$L = \prod_{i=1}^{n} \left(\lambda_0(t_i) e^{x_i'\beta}\right)^{\delta_i} \left(e^{-\Lambda_0(t_i) e^{x_i'\beta}}\right). \tag{4}$$

and the log-likelihood is

$$\log L = \sum_{i=1}^{n} \delta_i \left(log(\lambda_0(t_i)) + e^{x_i'\beta}\right) - \sum_{i=1}^{n} \Lambda_0(t_i) e^{x_i'\beta}. \tag{5}$$

2.2 | Score-type residuals

In this section, the score-type residuals are obtained. On differentiating log$L$ (5) with respect to β_j one obtains the j^{th} score function ($j = 1,..., p$), which is given by

$$U_j(\beta) = \frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \left(\delta_i - \Lambda_0(t_i) e^{x_i'\beta}\right). \tag{6}$$

Proposition 1 $\mathrm{E}(U_j(\beta)) = 0$

Proof: We provide a sketch of the proof in Appendix A1. The unknown parameters of the score function are estimated by solving the score equation, $E(U_j(\beta)) = 0$, where the score vector $U_j(\beta) = (U_1(\beta), ..., U_p(\beta))'$. The score equation involves two different parameters, Λ_0 and β, that need to be estimated using a two-step procedure:

First, one needs to estimate $\Lambda_0(t)$, say $\hat{\Lambda}_0(t)$, and we used both the Kaplan-Meier[9] (KM) $\hat{\Lambda}_{oKM}(t)$ and the Nelson-Aalen[10] (NA) estimators $\hat{\Lambda}_{oNA}(t)$ for the baseline cumulative hazard function.

Second, given the estimates $\hat{\Lambda}_0(t_i)$, we estimate β by solving the first and second derivatives of the p non-linear equations

$$\begin{aligned} \sum_{i=1}^{n} x_{i1} \left(\delta_i - \hat{\Lambda}_0(t_i) e^{x_i'\beta}\right) &= 0, \\ \sum_{i=1}^{n} x_{i2} \left(\delta_i - \hat{\Lambda}_0(t_i) e^{x_i'\beta}\right) &= 0, \\ \sum_{i=1}^{n} x_{\mathrm{ip}} \left(\delta_i - \hat{\Lambda}_0(t_i) e^{x_i'\beta}\right) &= 0. \end{aligned} \tag{7}$$

$$\frac{\partial^2 \log L}{\partial \beta_j \beta_k} = \sum_{i=1}^{n} x_{ij} \left(-\Lambda_0(t_i) e^{x_i'\beta} x_{ik}\right). \tag{8}$$

And, therefore, replacing β with $\hat{\beta}$ in Equation (8), we obtain.

$$-\frac{\partial^2 \log L}{\partial \hat{\beta}_j \hat{\beta}_k} = \sum_{i=1}^{n} x_{ij} x_{ik} \left(\hat{\Lambda}_0(t_i) e^{x_i'\hat{\beta}}\right) \tag{9}$$

for $j = 1, 2,..., p$ and $k = 1, 2,...p$.

If $\hat{\beta}$ is the solution to the p non-linear equations, then we obtain a $n \times p$ matrix of score residuals E, defined by

$$E = \begin{pmatrix} \hat{e}_{11} & \cdots & \hat{e}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{e}_{n1} & \cdots & \hat{e}_{np} \end{pmatrix} \tag{10}$$

where $\hat{e}_{ij} = x_{ij}\left(\delta_i - \hat{\Lambda}_0(t_i)e^{x_i'\hat{\beta}}\right)$. It should be noted that every column of the residual matrix E in (10) adds up to 0, that is,

$$\sum_{i=1}^{n}\hat{e}_{i1} = 0, \sum_{i=1}^{n}\hat{e}_{i2} = 0, ..., \sum_{i=1}^{n}\hat{e}_{ip} = 0$$

2.3 | Deviance residual

If L is the likelihood function, then by definition[11-14] the deviance function D, is given by

$$D = \sup 2\log L_{\text{Satuarted model}} - \sup 2\log L_{\text{usual}}, \tag{11}$$

where a "saturated model" is a model that perfectly reproduces the data. We show in Appendix A2 that the deviance function D, for the proportional hazards model is expressed by

$$D = 2\sum_{i=1}^{n}\left[(\delta_i\log\delta_i - \delta_i) - \delta_i\left(x_i'\hat{\beta} + \log\hat{\Lambda}_0(t_i)\right) + \hat{\Lambda}_0(t_i)e^{x_i'\hat{\beta}}\right], \tag{12}$$

where $\hat{\beta}$ and $\hat{\Lambda}_0$ are the estimates for β and Λ_0, respectively. Then, the deviance residual for the i^{th} individual is

$$D_i = \text{sign}\sqrt{2}\left[(\delta_i\log\delta_i - \delta_i) - \delta_i\left(x_i'\hat{\beta} + \log\hat{\Lambda}_0(t_i)\right) + \hat{\Lambda}_0(t_i)e^{x_i'\hat{\beta}}\right]^{1/2}. \tag{13}$$

If the i^{th} individual is censored ($\delta_i = 0$) and then $\delta_i\log\delta_i = 0$ and the expression in (13) becomes

$$D_i = \text{sign}\sqrt{2}\left[\hat{\Lambda}_0(t_i)e^{x_i'\hat{\beta}}\right]^{1/2}. \tag{14}$$

Whereas if the i^{th} individual is a failure ($\delta_i = 1$), then the expression in (13) is

$$D_i = \text{sign}\sqrt{2}\left[-\left(x_i'\hat{\beta} + \log\hat{\Lambda}_0(t_i)\right) + \hat{\Lambda}_0(t_i)e^{x_i'\hat{\beta}} - 1\right]^{1/2}. \tag{15}$$

2.4 | The partial likelihood approach

Cox[15] proposed a partial likelihood approach for the proportional hazards model to estimate β without involving $\lambda_0(t)$. The partial likelihood function is a product of the failure times of the conditional probabilities of observed individuals, chosen from the risk set to experience an event. Let $R(t) = \{i : t_i > t\}$ denote the risk set, that is, the set of individuals who are "at-risk" for failure at time t. Suppose individual (j) fails at t_j, then the conditional probability is

$$PL_m = P(\text{individual } (m) \text{ fails} \,|\, \text{one individual from } R(t_m) \text{ fails at } t_m)$$

$$\approx \frac{P(\text{individual } (m) \text{ fails at } t_m | (m) \text{ at risk})}{\sum_{l \in R(t_m)} P(\text{individual l fails at } t_m | \text{l at risk})}$$

$$\approx \frac{\lambda\left(t_m | x_{(m)}\right)}{\sum_{l \in R(t_m)} \lambda(t_m | x_l)}.$$

Under the proportional hazards model, the partial likelihood is

$$PL = \prod_{m=1}^{k} \frac{e^{x'_{(m)}\beta}}{\sum_{x_l \in R(t_m)} e^{x'_l \beta}} = \prod_{i=1}^{n} \left(\frac{e^{x'_i \beta}}{\sum_{x_l \in R(t_i)} e^{x'_l \beta}} \right)^{\delta_i}.$$

Then, the relationship between the full likelihood and the partial likelihood function is

$$L = \prod_{i=1}^{n} \left(f(t_i | x_i) \right)^{\delta_i} \left(S(t_i | x_i) \right)^{1-\delta_i} = \prod_{i=1}^{n} \left(\lambda(t_i | x_i) \right)^{\delta_i} S(t_i | x_i)$$

$$= PL \prod_{i=1}^{n} \left(\sum_{l \in R(t_i)} \lambda(t_i | x_l) \right)^{\delta_i} S(t_i | x_i).$$

Cox[1] argued that the partial likelihood term in the full likelihood contains almost all the information about β, which validates the partial likelihood approach.

2.4.1 | Residuals based on the partial likelihood approach

In what follows, we briefly review several widely used residuals based on the partial likelihood approach, with which we compare our proposed methods in the simulation and real examples.

The maximum likelihood estimation for β based on the partial likelihood is the solution for

$$\sum_{\delta_i = 1} (x_i - E(x_i | R(t_i))) = 0,$$

where $E(x_i | R(t_i)) = \frac{\sum_{x_l \in R(t_i)} x_l e^{x'_l \beta}}{\sum_{x_l \in R(t_i)} e^{x'_l \beta}}$ and the left-hand side is the score function for the partial likelihood. For each i with $\delta_i = 1$, Schoenfeld residual[2] is defined as

$$r_i = x_i - E(x_i | R(t_i)),$$

where for each component j with $1 \le j \le p$, $r_{ij} = x_{ij} - E(x_{ij} | R(t_i))$. Grambsch and Therneau[16] proposed the scaled Schoenfeld residual to assess the proportional hazards assumption. The martingale residual can assist in assessing the proportional hazards assumption as well, which was discussed by Lagakos[17] and is based on the following martingale process:

$$\hat{M}_i(t) = N_i(t) - \int_0^t I_{[t_i \ge s]} e^{x'_i \hat{\beta}} d\hat{\Lambda}_0(s), i = 1, \cdots, n,$$

where $N_i(t) = I_{[t_i \le t, \delta_i = 1]}$. The martingale residual for individual i is defined as:

$$\hat{M}_i = \hat{M}_i(\infty) = \delta_i - \int_0^{\infty} I_{[t_i \ge s]} e^{x'_i \hat{\beta}} d\hat{\Lambda}_0(s).$$

As indicated above, one concern in using the martingale residual is that it tends to be asymmetric. To overcome this problem, Therneau et al[18] introduced the deviance residual as a transformation of the martingale residual, and it is defined as

$$d_i = \text{sign}\left(\hat{M}_i\right)\left[-2\left\{\hat{M}_i + \delta_i \log\left(\delta_i - \hat{M}_i\right)\right\}\right]^{\frac{1}{2}}.$$

All the aforementioned residuals are helpful for assessing the proportional hazards assumption. The score residual[4] is useful for studying the influence of one observation. Similar to the Schoenfeld residual, the score residual is based on the score function. Since the score function $\sum_{\delta_i=1}(x_i - E(x_i|R(t_i)))$ can be rewritten as

$$\begin{gathered}\sum_{i=1}^{n}\int_0^{\infty}(x_i - E(x_i|R(s)))d\hat{M}_i(s), \text{ by the fact that} \\ \int_0^{\infty}(x_i - E(x_i|R(s)))dN_i(s) = x_i - (x_i|R(t_i)) \text{ for } \delta_i = 1 \text{ and} \\ \int_0^{\infty}(x_i - E(x_i|R(s)))dN_i(s) = 0 \text{ for } \delta i = 0.\end{gathered}$$

The score residual for individual i is defined as

$$\int_0^{\infty}(x_i - E(x_i|R(s)))d\hat{M}_i(s).$$

3 | SIMULATION STUDIES

3.1 | Simulation design

We describe different simulation scenarios conducted in this section. The survival data are generated under the proportional hazards assumption in order to compare the performances of residuals estimated from the full likelihood and the partial likelihood functions. We consider a sample size n with p different covariates. In the scenario of the proportional hazards, the true survival time for individual i, T_i, is generated as $T_i = {\Lambda_0}^{-1}\left[-\log(V_i)e^{x_i'\beta}\right]$,[19] where V is a uniform random variable, and Λ_0 is the cumulative hazard function such that $\Lambda_0(t) = \int_0^t \lambda_0(y)dy$. If C_i represents the censoring time for individual i, there the observed event (death) time for individual i is given by $t_i = \min(T_i, C_i)$. We consider exponential hazard as well as the Weibull hazard functions as choices for λ_0 as well for the scale parameter (2) and shape parameter (0.5 for the Weibull distribution). The censoring times $C_1, \ldots, C_n$ are generated from uniform distributions independent of the true survival times, with four different censoring proportions (0, 0.10, 0.20, and 0.50) as described in Halabi and Singh.[20] Three thousand simulations were generated for each scenario.

Due to the lack of guidelines in the literature for simulating outliers in a regression study, we assume that the outliers are generated through the extreme observations in the covariate space, also known as high leverage points. In general, high leverage points may or may not lead to outliers and influential observations. However, for simplicity, we assume that the high leverage points are related to the outliers in our simulation settings. We refer the reader to the following references[6, 11–13] for detailed discussions on outliers, leverages, and influential observations.

To assess the performance of the two methods in detecting outliers, covariate values are generated from three different normal distributions such that the 95% of the sample values are from a multivariate normal distribution with mean μ_1 and covariance matrix Σ, while the remaining 5% are generated from a multivariate normal distribution with mean μ_2 and covariance matrix Σ. We have assumed $n = 300$, $p = 3$, $\mu_1 = (0, 0, 0)'$, $\mu_2 = (-2, -3, -4)'$, and

$$\Sigma = \begin{pmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}.$$

We consider two different sets for β: $\beta = (1, 2, -1)'$ and $\beta = (0.2, 0.4, -0.2)'$.

We follow two approaches for classifying outliers based on the residuals. Suppose s_{ij} is the full likelihood-based score residual obtained for the j^{th} covariate component of the individual ($j = 1, 2, 3$; $i = 1, \ldots, 300$). If Q_{1j} and Q_{3j} are the

first and the third quartiles based on the residuals $\{s_{1j}, s_{2j}, ..., s_{nj}\}$ of the j^{th} covariate component, then the inter-quartile range (IQR) is obtained as $Q_{3j} - Q_{1j}$. Following Tukey's approach[21] on detecting outliers based on quartile measures, we construct the interval $[Q_{1j} - 1.5(Q_{3j} - Q_{1j}), Q_{3j} + 1.5(Q_{3j} - Q_{1j})]$ where a typical k observation is classified as a potential outlier for the j^{th} covariate component if the corresponding residual value s_{kj} is not contained in the aforementioned interval. We have also considered another interval $[Q_{1j} - 0.5(Q_{3j} - Q_{1j}), Q_{3j} + 0.5(Q_{3j} - Q_{1j})]$. We refer to these intervals as quartile based threshold intervals. We construct threshold intervals for the partial likelihood score residuals. We also consider thresholds based on the median absolute deviation (MAD).[22-23] If M_j is the median based on the j^{th} covariate component residuals $\{s_{1j}, s_{2j}, ..., s_{nj}\}$, then the corresponding MAD is defined as the median of $\{|s_{1j} - M_j|, |s_{2j} - M_j|, ..., |s_{nj} - M_j|\}$. A threshold interval for detecting potential outlier for the j^{th} covariate component is obtained as $[M_j - 3D_j, M_j + 3D_j]$ and $[M_j-, M_j + D_j]$, where D_j is the MAD for the j^{th} covariate component. Any k observation is classified as a potential outlier for the j^{th} covariate component if the corresponding residual value, s_{kj}, is not contained by this interval. We also construct similar MAD-based threshold intervals for detecting potential outliers based on the partial likelihood score residuals.

We compute the deviance residuals based on the full likelihood as well as the partial likelihood approach for detecting potential outliers. Unlike the score residuals, which are calculated for every covariate for each individual, the deviance residuals have only one value per individual. Suppose that $\{v_1, v_2, ..., v_n\}$ is the set of deviance residuals obtained from the whole data, M_V and D_V are the median and MAD based on the set of deviance residual, then a threshold interval based on MAD for detecting potential outlier observations is obtained as $[M_V - D_V, M_V + D_V]$. Any k observation is considered a potential outlier if the corresponding deviance residual, v_k, is not contained in this interval. These threshold intervals are constructed using both the partial as well as the full likelihood-based deviance residuals.

We apply Equations (7) and (11) to compute the score and the deviance residuals using the full likelihood approach. On the other hand, we fit the Cox proportional hazards model and compute the score and deviance residuals using the partial likelihood function. The function "residual" of the R package "survival"[24] computes the aforementioned residuals for the partial likelihood estimation, one may specify residual types as "martingale", "deviance", "score", "schoenfeld", and "scaledsch", where "scaledsch" denotes the scaled Schoenfeld residual and the other types are self-explanatory. We compare our proposed score and deviance residuals with the counterparts based on the partial likelihood approach using this R package.

We investigate the performance of the score residuals for each of the three covariates as well of the deviance residuals for the two different likelihood by computing the area under the receiver operating characteristic curve (AUC) using the library pROC in R.[25] There is only one decision interval for a data sample (or for each coefficient of a data set), that is, we only have one estimation pair of the sensitivity and the specificity. Therefore, the AUC estimate is based on the ROC curve using the only estimation pair of the sensitivity and the specificity and two end points, (0, 0) and (1, 1), on the sensitivity vs the 1-specificity plane. The AUC values for the full likelihood and the corresponding partial likelihood residuals are presented. Empirical standard errors of the AUC are also provided based on Monte Carlo simulations and the results are averaged over 3000 simulations. The R codes were written by the second author and are available on https://duke.box.com/s/kjnruu2e9corg1rrv0g0uncikucu41u7.

3.2 | Simulation results

3.2.1 | Small effect size

We present the simulation results for the small effect size where $\beta = (0.2, 0.4, -0.2)'$. In computing the full likelihood score-type and deviance-type residuals, we have utilized the Kaplan-Meier estimator of the baseline cumulative hazard function denoted as KM in the tables. We observe that as the censoring proportion increases the performance of score residuals, estimated from the full likelihood, becomes better than that score residuals from the partial likelihood. This is evident when the censoring proportion is 0.50 for the Weibull distribution (Table S1, Figure 1) and for 0.20 and 0.50 censoring proportion when the failure rate was exponentially distributed (Table S2, Figure 2). On the other hand, the partial likelihood score residuals have higher levels of AUC than the full likelihood residuals under lower censoring proportion (namely, 0 and 0.10). When we narrow the thresholds, the AUC values of both the full likelihood and the partial likelihood residuals increase. However, the full likelihood approach outperforms the partial likelihood residuals at high censoring proportion where the AUC values of the full likelihood residuals reach as high as 0.81 (Table S2, Figure 2).

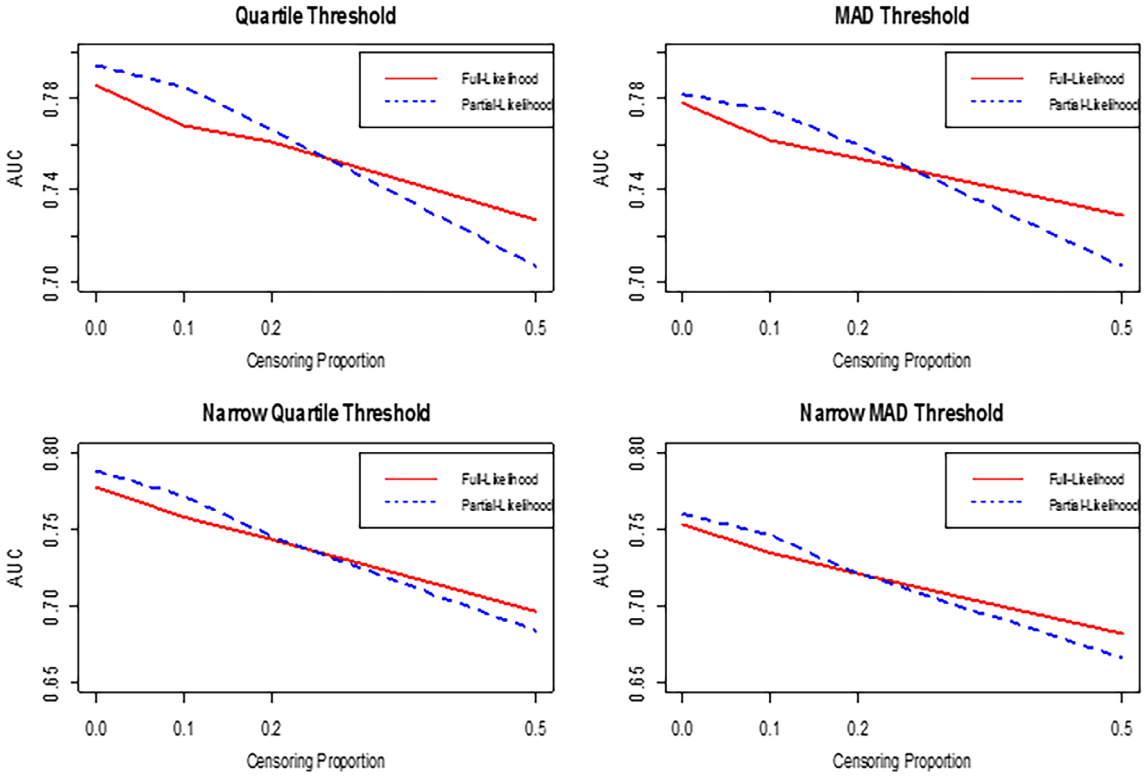

FIGURE 1 Plot summarizing the AUC levels for the score residuals computed using the full and the partial likelihood functions involving the covariate X.3

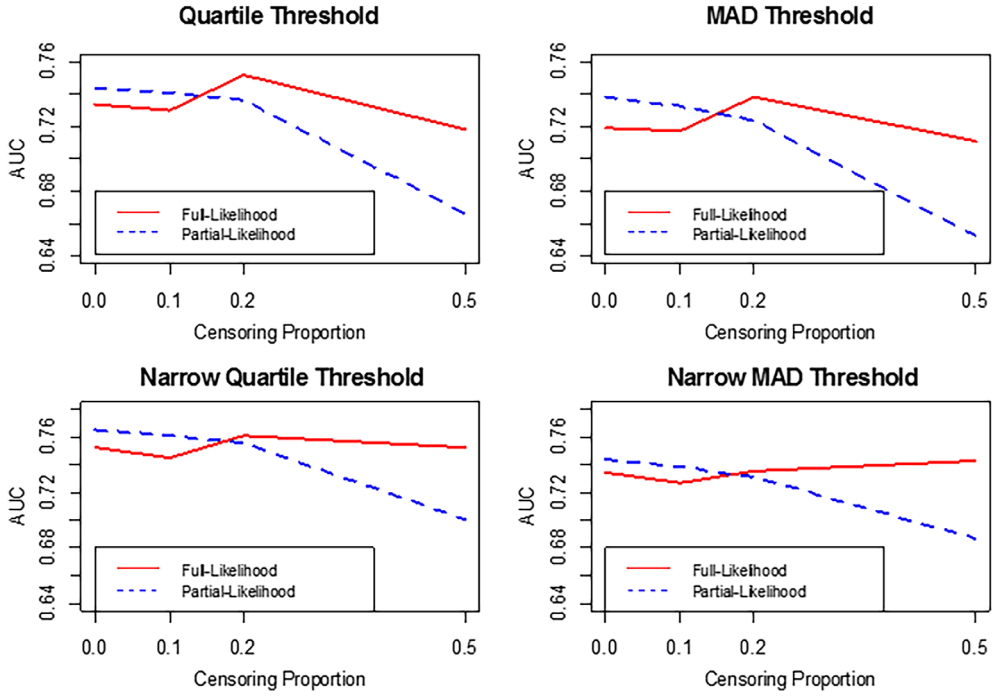

FIGURE 2 Plot summarizing the AUC levels for the score residuals computed using the full and the partial likelihood functions involving the covariate X.2

We assess the performances of the full likelihood-based deviance residuals with the partial likelihood-based deviance residuals, assuming the failure times follow exponential and Weibull distributions, respectively (Table 1). We observe that the full likelihood-based deviance residuals have higher AUC levels than the partial likelihood-based deviance residuals in identifying outliers irrespective of the censoring proportion and failure time distributions (Table 1).

TABLE 1 AUC and its standard error (SE) based on the partial and full likelihood methods of the deviance residuals where $\beta = (0.2, 0.4, -0.2)'$, mean of the covariate distribution of the true outliers as $\mu_2 = (-2, -3, -4)'$ and based on 3000 Monte-Carlo simulations

	AUC (SE)			
Method	**0 censoring**	**0.10 censoring**	**0.20 censoring**	**0.50 censoring**
Exponential distribution				
KM full likelihood deviance	0.621 (0.066)	0.611 (0.065)	0.598 (0.065)	0.517 (0.114)
Partial likelihood deviance	0.502 (0.059)	0.519 (0.061)	0.521 (0.061)	0.458 (0.051)
Weibull distribution				
KM full likelihood deviance	0.621 (0.066)	0.605 (0.065)	0.579 (0.066)	0.548 (0.094)
Partial likelihood deviance	0.501 (0.058)	0.503 (0.060)	0.503 (0.060)	0.470 (0.061)

Abbreviation: KM, Kaplan-Meier estimator for the cumulative baseline hazard function.

We have observed similar results for the AUC levels of the score (and deviance) residual based on the full likelihood function utilizing the Kaplan-Meier (KM) and the Nelson-Aalen (NA) estimators for the baseline cumulative hazard function (Table S2). Therefore, we report the results from the KM estimator for the full likelihood residual results in the simulation studies to avoid redundancy.

3.2.2 | Large effect size

We next consider the scenario where $\beta = (1, 2, -1)'$ and present the AUC values for the different score residuals, assuming Weibull failure times in Table 2. We observe that the AUC of the full likelihood score residual is greater than the partial likelihood score residuals irrespective of the censoring proportion. The AUC values of the score residuals in this setting of large effect size appear to be smaller than the corresponding AUC values under similar censoring proportion in the setting of small effect size, that is, $\beta = (0.2, 0.4, -0.2)'$. This is more evident in the AUC values of the partial likelihood score residuals. Particularly, at the high censoring proportion of 0.50, the AUC values of the partial likelihood residuals remain below 0.5, making them practically ineffective. On the other hand, the full likelihood score residual maintains high AUC values with increased censoring proportion.

4 | DATA ANALYSIS

4.1 | Primary biliary cirrhosis data analysis

We computed the score and deviance residuals to identify outliers using the primary biliary cirrhosis (PBC), a well-studied example in the literature.[4] The original data set includes information on survival time, survival status (dead or censored), disease type, treatment type, age, gender, and some baseline laboratory measurements of 424 patients. The censoring proportion in this data set was 0.61. We are interested in examining the effects of a patient's age, albumin level, and bilirubin level on the outcome overall survival. These three continuous variables have observed values for all the patients. We fit a proportional hazards model with the three covariates and we evaluate whether there is any outlier to the fitted model. We apply our proposed full likelihood-based residuals, as described in Section 2. In addition, we compare the full likelihood residuals with the partial likelihood score residuals.

We present the plots of the full likelihood-based score residuals using the KM cumulative hazard estimators for the variables age, albumin, and bilirubin, respectively (Figure 3). The left panel shows the full likelihood score residuals, whereas the right panel presents the score residuals estimated using the partial likelihood function. In each of the figures, the middle horizontal line refers to the residual value of zero, while the uppermost and the lowermost horizontal lines refer to the upper and lower limits of the quartile based threshold interval of the full likelihood score residuals as described in Section 3. We observe that the score residuals for all the three variables are symmetrically distributed around zero, which is expected. However, the dispersion of the full likelihood-based score residuals is large compared with the partial likelihood-based residuals. The score residuals for age and albumin covariates are randomly distributed

TABLE 2 AUC and its standard error (SE) for different methods of score residuals assuming Weibull failure times, $=(1, 2, -1)^{'}$, assuming mean of the covariate distribution of the true outliers as $\mu_2 = (-2, -3, -4)^{'}$ based on 3000 Monte-Carlo simulations

Method	AUC (SE)											
	0 censoring			0.10 censoring			0.20 censoring			0.50 censoring		
	$X_{.1}$	$X_{.2}$	$X_{.3}$	$X_{.1}$	$X_{.2}$	$X_{.3}$	$X_{.1}$	$X_{.2}$	$X_{.3}$	$X_{.1}$	$X_{.2}$	$X_{.3}$
KM score residual[a]	0.617 (0.061)	0.698 (0.066)	0.779 (0.065)	0.592 (0.058)	0.678 (0.066)	0.760 (0.066)	0.568 (0.053)	0.648 (0.065)	0.745 (0.065)	0.522 (0.067)	0.578 (0.087)	0.704 (0.095)
KM score residual[b]	0.604 (0.059)	0.684 (0.065)	0.764 (0.067)	0.578 (0.055)	0.662 (0.066)	0.744 (0.067)	0.558 (0.050)	0.634 (0.064)	0.729 (0.065)	0.529 (0.069)	0.603 (0.099)	0.693 (0.096)
KM score residual[c]	0.657 (0.065)	0.725 (0.058)	0.783 (0.051)	0.648 (0.066)	0.736 (0.060)	0.782 (0.053)	0.627 (0.067)	0.729 (0.061)	0.777 (0.056)	0.548 (0.097)	0.634 (0.112)	0.731 (0.079)
KM score residual[d]	0.652 (0.062)	0.714 (0.054)	0.759 (0.044)	0.652 (0.063)	0.725 (0.053)	0.761 (0.046)	0.636 (0.066)	0.726 (0.054)	0.758 (0.048)	0.588 (0.098)	0.675 (0.092)	0.723 (0.069)
Partial likelihood score residual[a]	0.573 (0.053)	0.670 (0.063)	0.762 (0.062)	0.554 (0.051)	0.639 (0.062)	0.714 (0.063)	0.536 (0.050)	0.606 (0.059)	0.662 (0.064)	0.442 (0.044)	0.472 (0.050)	0.488 (0.055)
Partial likelihood score residual[b]	0.566 (0.052)	0.668 (0.064)	0.751 (0.062)	0.551 (0.049)	0.635 (0.062)	0.708 (0.062)	0.535 (0.048)	0.602 (0.059)	0.661 (0.062)	0.455 (0.042)	0.490 (0.048)	0.495 (0.054)
Partial likelihood score residual[c]	0.606 (0.064)	0.719 (0.062)	0.761 (0.054)	0.562 (0.064)	0.660 (0.066)	0.696 (0.062)	0.520 (0.062)	0.600 (0.066)	0.631 (0.066)	0.388 (0.051)	0.417 (0.057)	0.445 (0.062)
Partial likelihood score residual[d]	0.604 (0.065)	0.701 (0.058)	0.736 (0.050)	0.550 (0.066)	0.636 (0.065)	0.672 (0.061)	0.502 (0.064)	0.574 (0.067)	0.609 (0.066)	0.372 (0.054)	0.402 (0.060)	0.425 (0.064)

Abbreviation: KM, Kaplan-Meier estimator for the cumulative baseline hazard function.
[a]Quartile-based threshold $[Q_{1j} - 1.5(Q_{3j} - Q_{1j}), Q_{3j} + 1.5(Q_{3j} - Q_{1j})]$.
[b]MAD-based threshold $[M_j - 3D_j, M_j + 3D_j]$.
[c]Quartile-based narrower threshold $[Q_{1j} - 0.5(Q_{3j} - Q_{1j}), Q_{3j} + 0.5(Q_{3j} - Q_{1j})]$.
[d]MAD-based narrower threshold $[M_j - D_j, M_j + D_j]$.

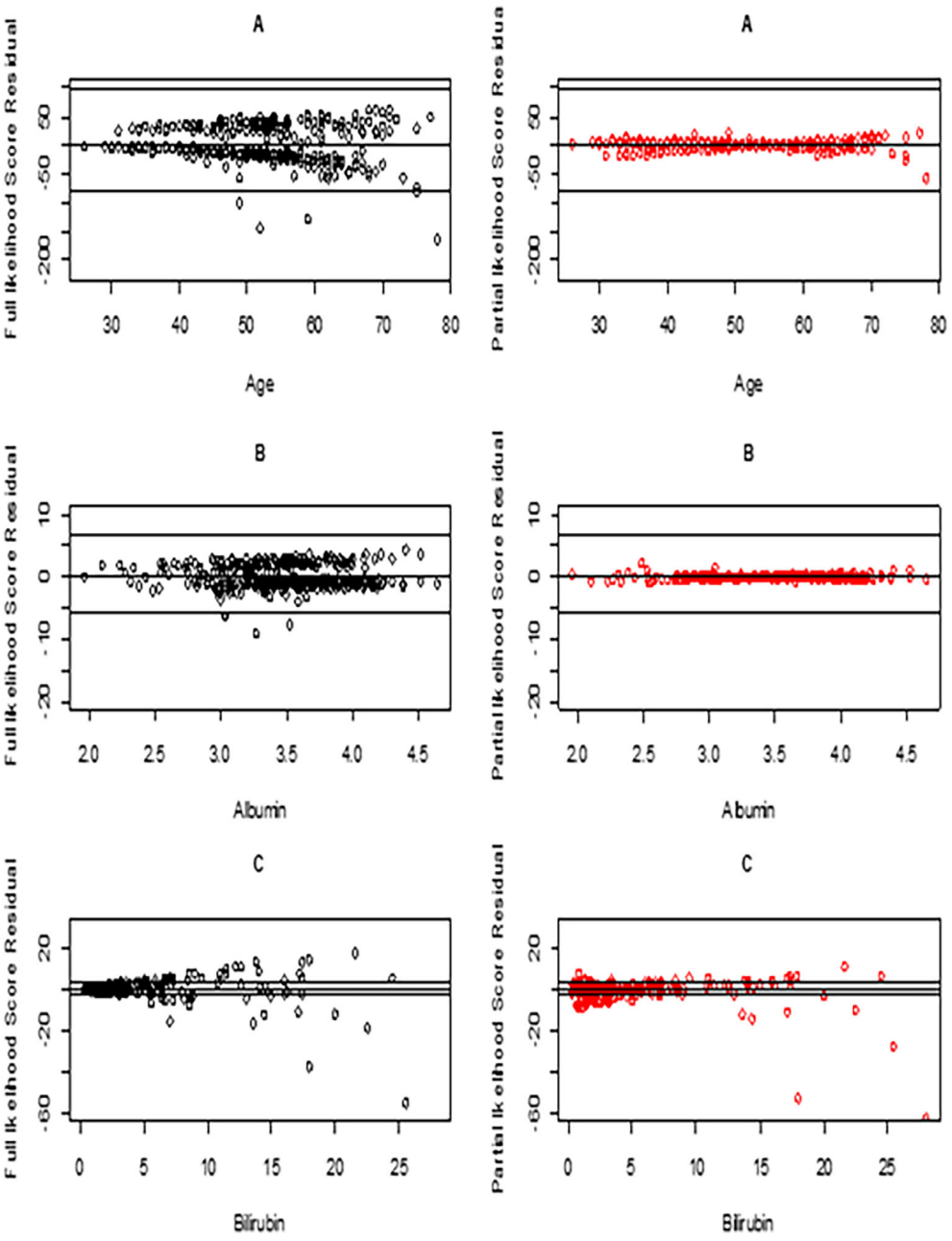

FIGURE 3 A-C, Plots of the full likelihood and the partial likelihood score residuals against age, albumin, and bilirubin values in PBC data

around zero with wide dispersion (Figure 3A,B), with few residual values falling below the lower limit of the threshold interval. This suggests that, despite the presence of a few outliers, the fitted proportional hazards model with age and albumin may be a good fit.

On the other hand, we observe that the score residuals of bilirubin are close to zero at the lower values of bilirubin, but gradually deviate from zero as the bilirubin levels increase (Figure 3C). This leads to a high number of outliers for larger values of bilirubin that are outside the limits of the threshold interval. This suggests that bilirubin should be transformed and not modeled in its original scale.

Furthermore, we present the plot of the deviance residual against risk scores. The risk score variable for every sample is constructed based on the three given covariates (age, bilirubin, and albumin). This is done by multiplying the covariate value with the corresponding regression coefficient estimated from the full likelihood approach and then taking the sum over the products (Figure 4). The uppermost and the lowermost horizontal lines represent the upper and lower threshold limits at y-axis values of 3 and −3, respectively, while the horizontal line at the middle represents the y-axis value of zero. We note that the deviance residual with respect to the linear predictor has a

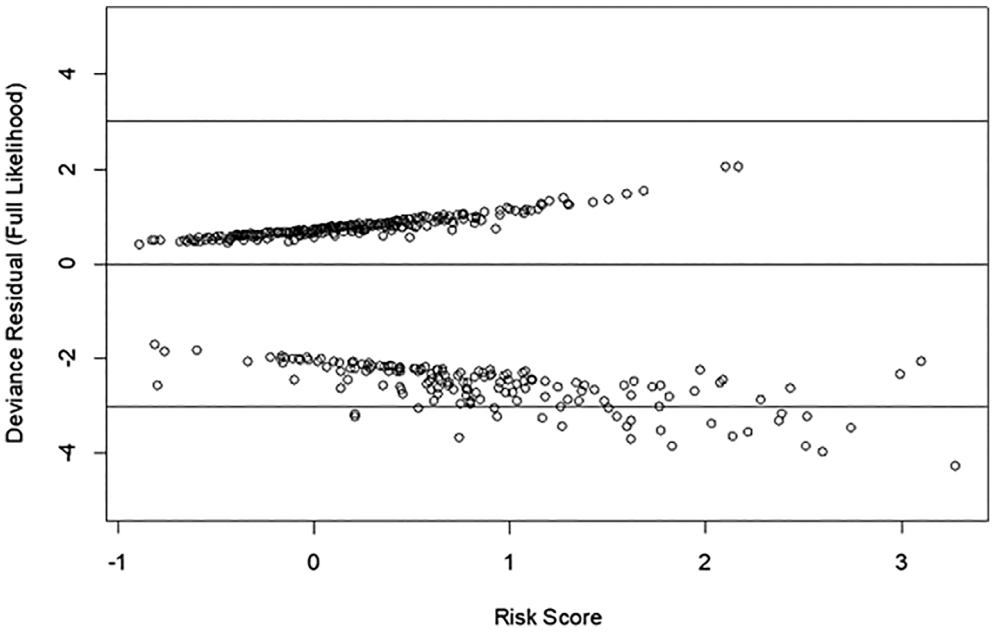

FIGURE 4 Plot of the full likelihood deviance residual against the risk score (linear predictor) based on age, albumin, and bilirubin values in the PBC data

distinct pattern; in that, the deviance residual values tend to be more dispersed with the increased values of the risk score.

4.2 | German breast cancer data analysis

The German Breast Cancer study[26] data comprise of 686 individuals affected with breast cancer, where we have information on the survival time, censoring (right censoring) status, age, menopause status, tumor grade, tumor size, hormone therapy status (whether or not hormone therapy received), progesterone receptor level, and estrogen receptor level of each of the patients. The censoring proportion in this data set is 0.56.

Figure 5 shows the score residual for age, progesterone, and estrogen. The left panel

presents the results from the full likelihood residuals, whereas the right panel is for the partial likelihood residuals. For the age covariate, overall, the patterns of the score residuals based on the full likelihood and the partial likelihood are similar (Figure 5A). While both sets of score residuals are randomly and symmetrically placed around the zero line, the full likelihood residuals are more widely spread than the partial likelihood residuals. Moreover, we note that the full likelihood score residuals for age are within the quartile-based thresholds as marked by the uppermost and the lowermost horizontal lines (Figure 5A). This implies that there are no significant outliers for age with respect to the fitted model.

We present the score residual plots for progesterone and estrogen receptor levels in Figure 5B,C. We observe that the full likelihood score residuals and the partial likelihood score residuals follow similar patterns. These residuals, however, are not randomly distributed and tend to deviate further away from the zero line as the values of the covariates increase. This also leads to an increased number of outliers from the plots as many observations fall outside the quartile-based thresholds. From the above analysis, we conclude that modelling progesterone and estrogen receptors in their original scales is not appropriate in the proportional hazards model.

5 | DISCUSSION

We have used the full likelihood approach to develop innovative residuals for the proportional hazards model. We computed the score-type residuals by solving the score equations from the full likelihood function. In addition, we calculated the deviance residuals from the deviance function based on the full likelihood function. We compared the performances of the score and deviance residuals computed from the full likelihood function with the score and deviance residuals computed from the partial likelihood function in identifying outliers. Overall, we found that the score

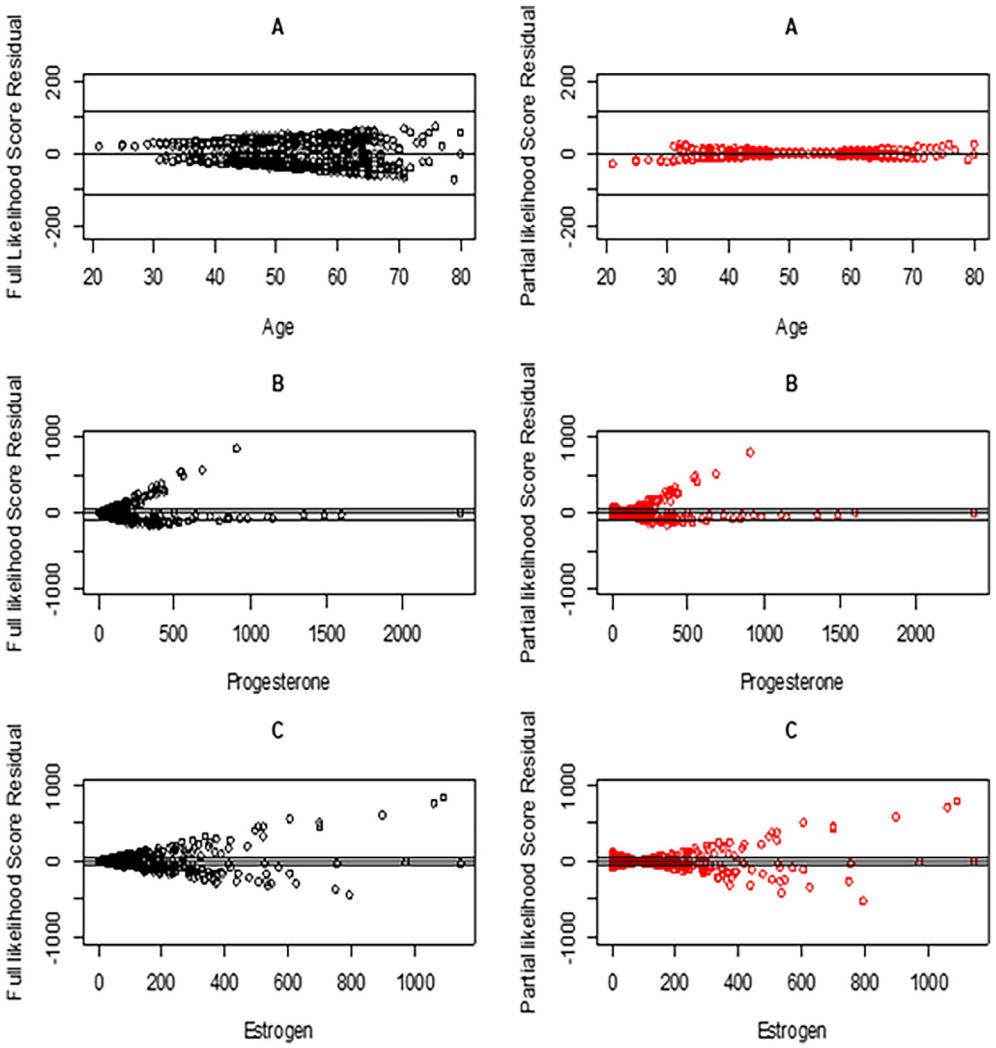

FIGURE 5 A-C, Plots of the full likelihood and the partial likelihood score residuals for age, progesterone receptor, and estrogen receptor levels in the German breast cancer data

residuals derived from the full likelihood function have higher area under the curve (AUC) values in detecting outliers than the partial likelihood approach. Specifically, the full likelihood score residuals outperform the competing methods when the censoring proportion is high. In addition, the deviance residual based on the full likelihood approach had higher values in the AUC than the deviance residual derived from the partial likelihood function. It is noteworthy to point out that residuals from the two methods have different formulas and it is likely that that the residuals from the full likelihood function have large variability compared to that from the partial likelihood function due to the baseline cumulative hazard.

To the best of our knowledge, there are very few articles that have simulated data in studying residuals. One of the main challenges in detecting outliers is how to determine an optimal threshold limit for the appropriate detection of outliers. In our simulation settings, we have generated extreme observations in the covariate space to indicate the presence of outliers. We have used Tukey's quantile-based threshold and the median-based threshold limits for comparing the AUC of the full likelihood and the partial likelihood residuals. Such thresholds have been utilized in previous studies[16-17] and are useful in comparing the detection ability among the different methods. It is worth indicating that the AUC values for both the full and partial likelihood methods, obtained from the simulation studies, depend on the choice of threshold used for detecting outliers. In other words, one can attain higher overall detection accuracies by using some variations of these thresholds. For example, when we restrict the distance between the lower and upper thresholds, the values of the AUCs from both methods increase. We note that the AUC levels for the score residuals in detecting outliers based on the full likelihood approach are also higher than the partial likelihood approach. Most of the existing residual methods tend to use graphical representations to identify outliers. While visual methods are useful, they tend to be subjective. Nevertheless, we have provided an

objective framework to study outliers. More research, however, is needed to set a standard definition for thresholds that can help in identifying outlier efficiently.

In summary, we have employed a two-step procedure in estimating the score and deviance residuals based on the full likelihood function. In the first step, one needs to estimate the baseline cumulative hazard function in constructing the residuals. We have considered both the Kaplan-Meier and the Nelson-Aalen estimators for this purpose. The results showed that there is no difference in the performances of the new residuals based on the two different estimators of the baseline cumulative hazard function. Residuals derived from the full likelihood function have higher AUC levels in detecting outlier than the partial likelihood approaches when the censoring proportion is high. They are robust and can be easily computed. Investigators are encouraged to use them in identifying outliers when the censoring proportion is high.

ACKNOWLEDGEMENTS

This research was supported in part by National Institutes of Health Grants R21 CA195424, U01CA157703, United States Army Medical Research W81XWH-15-1-0467 and W81XWH-18-1-0278, and the Prostate Cancer Foundation Challenge Award. Research of A. Liu was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The authors would like to thank Dr. Bahadur Singh for his insightful comments and helpful suggestions on this manuscript prior to his death.

DATA AVAILABILITY STATEMENT

We have used datasets that are available in the public domain. All the programs and the datasets are available on this url link https://duke.app.box.com/folder/74677308943.

ORCID

Susan Halabi https://orcid.org/0000-0003-4135-2777
Sandipan Dutta https://orcid.org/0000-0002-7211-2752
Yuan Wu https://orcid.org/0000-0001-8925-555X
Aiyi Liu https://orcid.org/0000-0002-6618-5082

REFERENCES

1. Cox DR. Regression models and life tables (with discussion). *J Roy Stat Soc Series B*. 1972;74:187-220.
2. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika*. 1982;69:239-241.
3. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. New York: Wiley; 1991.
4. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
5. Barlow WE, Prentice RL. Residuals for relative risk regression. *Biometrika*. 1988;75:65-74.
6. Colette D. *Modelling Survival Data in Medical Research*. 3rd ed. New York: Chapman and Hall/CRC; 2014.
7. Cain KC, Lange NT. Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*. 1984;40:493-499.
8. Lawless JF. *Statistical Models and Methods for Lifetime Data*. New York: Wiley; 1982.
9. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.
10. Aalen O, Borgan O, Gjessing H. *Survival and Event History Analysis: A Process Point of View (Statistics for Biology and Health)*. New York: Springer; 2008.
11. Belsley DA, Kuh E, Welsh RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley; 1980.
12. Draper NR, Smith H. *Applied Regression Analysis*. 2nd ed. New York: Wiley; 1981.
13. Cook RD, Weisberg S. *Residuals and Influence in Regression*. New York: Chapman and Hall; 1982.
14. Agresti A. *An Introduction to Categorical Data Analysis*. 3rd ed. New York: Wiley; 2019.
15. Cox DR. Partial likelihood. *Biometrika*. 1975;62:269-276.
16. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81:515-526.
17. Lagakos SW. The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika*. 1981;68:93-98.
18. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. *Biometrika*. 1990;77:147-160.
19. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24:1713-1723.
20. Halabi S, Singh B. Sample size determination for comparing several survival curves with unequal allocations. *Stat Med*. 2004;23:1793-1815.
21. Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. *Am Stat*. 1989;43:50-54.
22. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Am Stat Ass*. 1993;88:1273-1283.
23. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Experimental Social Psychol*. 2013;49:764-766.

24. Therneau TM, Lumley T, Elizabeth A, Cynthia C. Survival: Survival Analysis. R Package Version 3. 2020:1-12.
25. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.
26. Schumacher M, Bastert G, Bojar H. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *J Clin Oncol*. 1994;12:2086-2093.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Halabi S, Dutta S, Wu Y, Liu A. Score and deviance residuals based on the full likelihood approach in survival analysis. *Pharmaceutical Statistics*. 2020;1–15. https://doi.org/10.1002/pst.2047