ORIGINAL RESEARCH ARTICLE

Z-Residual Diagnostic Tool for Assessing Covariate Functional Form in Shared Frailty Models*

Tingxuan Wu[a,b], Longhai Li[a] and Cindy Feng[c]

[a]Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, CA;
[b]School of Public Health, University of Saskatchewan, Saskatoon, CA;
[c]Department of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, CA.

ARTICLE HISTORY
Compiled October 3, 2024

ABSTRACT
Survival analysis often involves modelling hazard functions while considering frailty to account for unobserved cluster-level factors in clustered survival data. Shared frailty models have gained popularity for this purpose, but assessing covariate functional form in these models presents unique challenges. Martingale and deviance residuals are commonly used for visually assessing covariate functional form against continuous covariates. Nevertheless, their subjective nature and lack of a reference distribution make it challenging to derive numerical statistical tests from these residuals. To address these limitations, we propose "Z-residuals", a novel diagnostic tool designed for shared frailty models, leveraging the concept of randomized survival probability and introducing both graphical and numerical tests. To implement this approach, we develop an R package to compute Z-residuals for shared frailty models. Through extensive simulation studies, we demonstrate the high power of our derived numerical test for assessing the functional form of covariates. To validate the effectiveness of our method, we apply it to a real data application concerning the modelling of survival time for acute myeloid leukemia patients. Our Z-residual diagnosis results reveal the inadequacy of log-transformation of the covariate, highlighting the limitations of other diagnostic methods for effectively assessing covariate functional form in shared frailty models.

KEYWORDS
Survival analysis; clustered survival data; shared frailty models; covariate functional form; residual diagnostics; z-residuals

1. Introduction

Survival data often exhibits a multilevel structure, commonly seen in scenarios where patients are clustered within hospitals. In such cases, the hazard of events may vary across clusters due to unobserved cluster-level factors. Traditional survival analysis methods, like the Cox proportional hazard models [6] and accelerated failure time models [30], assume independence among subjects. However, to account for cluster-level heterogeneity, it becomes necessary to incorporate random effects into survival

*This is the accepted version of the paper: Wu, T., Li, L., and Feng, C. (2024). Z-residual diagnostic tool for assessing covariate functional form in shared frailty models. *Journal of Applied Statistics*, forthcoming. https://doi.org/10.1080/02664763.2024.2355551

CONTACT Cindy F. Author. Email: cindy.feng@dal.ca

models. Shared frailty models offer a robust solution to address this cluster-level heterogeneity in survival analysis. These models extend classic survival approaches by introducing random effects (frailties) that act multiplicatively on the baseline hazard function [44]. These frailties are shared among individuals within a cluster or group, allowing us to estimate the extent of unobserved heterogeneity within the clusters. [4, 12, 18, 24]. The frailty model has gained significant popularity across various fields, including public health, environmental studies, and ecological research [1, 5, 42, 47], for its ability to handle clustered survival data and account for the underlying heterogeneity.

Despite the popularity of shared frailty models for clustered survival data, comprehensive tests to examine covariate functional forms are lacking. Accurate diagnosis of functional forms is crucial for model reliability. Residual diagnostics are commonly used for goodness of fit and detecting model misspecifications, including issues related to covariate effects. However, conducting residual diagnostics with censored observations in clustered survival data presents challenges. The widely used Cox-Snell (CS) residual [5, 7] is defined as the negative logarithm of the estimated survival probability. In the absence of censored observations, CS residuals follow an exponential distribution when the model is accurate. However, with censored observations, CS residuals deviate from the exponential distribution due to non-uniform survival probabilities. To address censored observations, diagnostic procedures compare the cumulative hazard plot of CS residuals, estimated using the Kaplan-Meier method, to the expected cumulative hazard of the standard exponential distribution. This comparison helps identify potential deviations from the expected distribution and assess the adequacy of the model in capturing the functional form of covariate effects in clustered survival data.

While overall goodness-of-fit checks, like examining the cumulative hazard plot of CS residuals, are commonly used for diagnosing survival models, they may not provide sufficient information about specific model inadequacies. To address this, tailored graphical and numerical diagnostic tools are necessary, particularly when assessing the functional form of covariates. Several residual diagnostic tools have been proposed for assessing the functional form of covariates [5], including martingale residuals [43] and deviance residuals [32, 42], which are widely used in survival analysis. Martingale residuals measure the difference between a subject's observed failure indicator and its expected value, integrated over the time the patient was at risk. They enable the assessment of the functional form of covariates and the identification of potential outliers in survival data. Deviance residuals, a normalized transformation of martingale residuals, exhibit a mean of zero when the fitted model is appropriate and is approximately symmetrically distributed around zero. While both types of residuals are widely accessible in the `survival` package in R software, they have their limitations. For instance, martingale residuals are asymmetric and lack a lower bound, making visual inspection challenging. Conversely, deviance residuals exhibit less skewness and approximate a more normal distribution, enhancing visual assessment. To gain further insights from residual plots, researchers often employ locally weighted scatterplot smoothing (LOWESS) lines on scatterplots of residuals against continuous covariates. However, visual interpretation of these lines can still be subjective. Addressing the need for a more objective approach, researchers seek numerical measures of statistical significance to quantify observed trends in residual plots. However, deriving such tests for martingale and deviance residuals is challenging due to censoring, as they lack a reference distribution.

In an effort to bridge this gap, a recent study by Li et al. 2021 [27] introduced

the concept of randomized survival probabilities (RSPs) to define residuals for checking model assumptions in accelerated failure time (AFT) models. The RSP approach involves replacing the survival probability of a censored failure time with a uniform random number between 0 and the survival probability of the censored time. As RSPs are uniformly distributed under the true model, they can be transformed into normally distributed residuals using the normal quantile function, resulting in normally-transformed RSP (NRSP) residuals. By having a reference distribution with NRSP residuals, statistical tests can be derived to assess model assumptions, including distributional assumptions and the functional form of covariates. However, the extension of NRSP residuals to diagnose Cox proportional hazard models or shared frailty models remains an unexplored area, calling for further research to adapt and apply NRSP residuals specifically to these complex survival models.

In this study, we extend the concept of NRSP residuals and develop residual diagnostic tools tailored specifically for assessing the functional form of covariates in shared frailty models. To simplify the terminology, we refer to these extended residuals as Z-residuals, adopting the convention of using 'Z' to represent a standard normal random variable. To facilitate the implementation of this approach, we develop an R package for calculating these conditional Z-residuals based on the output of the `coxph` function in the `survival` package in R. Additionally, we propose a non-homogeneity test to examine whether discernible trends exist in the Z-residuals. To evaluate the performance of our Z-residual diagnostic tool in detecting misspecification of covariate functional forms, we conduct extensive simulation studies. Furthermore, we illustrate the effectiveness of Z-residuals in diagnosing the functional form of covariates through a real data analysis focused on the mortality risk of acute myeloid leukemia patients [15, 19].

The remainder of this paper is structured as follows: In Section 2, we provide a concise overview of shared frailty models. Section 3 reviews existing residual diagnostic methods for shared frailty models. Next, in Section 4, we present the definition of Z-residuals and introduce the non-homogeneity test based on these residuals. Section 5 presents the results of our simulation studies, evaluating the performance of the Z-residual diagnostic tool. We then demonstrate the application of the Z-residual diagnostic tool in real data analysis in Section 6. Finally, in Section 7, we conclude the paper by summarizing our findings and highlighting the significance of our study in advancing residual diagnostics for assessing covariate functional forms in shared frailty models.

2. Notation and Shared Frailty Model

A shared frailty model incorporates common or shared frailties among individuals within groups. In the context of clustered failure survival data, the formulation of a frailty model can be defined as follows. Suppose there are g groups of individuals, with each group containing n_i individuals, indexed as $i = 1, 2, \ldots, g$. In the case where all groups consist of a single subject, we have a univariate frailty model [24]. However, if there is more than one subject per group, the model is known as a shared frailty model [12, 21, 22], where all individuals within the same cluster share the same frailty value, denoted as z_i. The true failure time for the jth individual from the ith group is denoted as T^*_{ij}, which we assume to be a continuous random variable, where $j = 1, 2, \cdots, n_i$.

In a shared frailty model, the true failure times T^*_{ij} within a cluster, conditional on

the observed covariates, are assumed to be independent. A conventional shared frailty model also assumes that the groups are independent of one another, so the frailty in one group is unrelated to the frailty in another group. Let t_{ij}^* be the realization of T_{ij}^*. C_{ij} is the corresponding censoring time, assumed to be independent of T_{ij}^*. In the scenario of right censoring, we observe the event time T_{ij}, which is the minimum of the true failure time T_{ij}^* and the censoring time C_{ij}. Specifically, $T_{ij} = \min(T_{ij}^*, C_{ij})$. The non-censoring indicator is denoted as $\delta_{ij} = I(T_{ij}^* < C_{ij})$. Collectively, the observed failure times are represented by the pairs (T_{ij}, δ_{ij}). The observed data can be succinctly expressed as $t = (t_{11}, \cdots, t_{gn_g})$, and the corresponding non-censoring indicators as $\delta = (\delta_{11}, \cdots, \delta_{gn_g})$.

In a shared frailty model, the conditional hazard function of the failure time T_{ij}^* for the jth individual, $j = 1, 2, \cdots, n_i$, in the ith group, denoted as $h_{ij}(t|x_{ij}, z_i)$ and abbreviated as $h_{ij}(t)$ for simplicity, is given by

$$h_{ij}(t) = z_i \exp(\beta' x_{ij}) h_0(t), \tag{1}$$

and the conditional survival function for the jth individual of the ith group at time t, denoted as $S_{ij}(t|x_{ij}, z_i)$ and abbreviated as $S_{ij}(t)$ for simplicity, follows:

$$S_{ij}(t) = \exp\left\{ -\int_0^t h_{ij}(s)\, \mathrm{d}s \right\} = \exp\left\{ - z_i \exp(\beta' x_{ij}) H_0(t) \right\}, \tag{2}$$

where x_{ij} is a column vector of values of p explanatory variables for the jth individual in the ith group, i.e., $x = (x_{11}, \cdots, x_{gn_g})^T$; β is a column vector of regression coefficients; $h_0(t)$ is the baseline hazard function, $H_0(t)$ is the baseline cumulative hazard function (CHF), and z_i is the frailty term that is common for all n_i individuals within the ith group.

The baseline CHF is estimated using the Breslow estimator [11, 28], suitable for continuous event times with few or no tied event times. This method is commonly used when estimating the baseline hazard. Efron's method [13, 14] is an alternative that is more accurate when a large number of ties are present. Various frailty distributions can be adopted, including Gamma, Gaussian, and t distributions, with Gamma being one of the most commonly used distributions [5] due to its closed-form representation of the observable survivor and hazard functions.

Several methods exist for estimating a shared frailty model. The Expectation-maximization (EM) algorithm [10, 12] and the Penalized Partial Likelihood (PPL) method [12, 33] are notable among them. In a comprehensive comparison study of R packages for shared frailty models [45], the EM algorithm implemented in the `frailtyEM` package [2] and the PPL method in the `survival` package [41] in R yielded nearly identical estimates. However, caution is advised with the `frailtyEM` package, especially in cases of small sample sizes with high censoring rates, due to its lower convergence rate. Consequently, the preferred choice for parameter estimation in a shared frailty model is the `survival` package, known for its computational efficiency and high convergence rate. Therefore, in this study, parameter estimation in shared frailty models is performed using the `coxph` function from the widely used `survival` package, utilizing PPL method to estimate model parameters.

3. Review of Existing Residuals for Survival Models

In this section, we will review some existing residuals used in survival analysis, particularly focusing on their application to diagnose the goodness-of-fit (GOF) of survival models. A central concept in these residuals is formulated based on the survival probability, which is a fundamental component in survival analysis. One of the widely used residuals in survival analysis is the Cox-Snell (CS) residual [5, 7], defined as $r^c_{ij}(t_{ij}) = -\log(\hat{S}_{ij}(t_{ij}))$, where $\hat{S}_{ij}(t_{ij})$ is estimated survival function of the jth individual from the ith cluster. In the absence of censored observations, the survival probability follows a uniform distribution under the true model [8], and as a result, the CS residuals are exponentially distributed. A graphical check can be performed by plotting the CHF against the true failure time, which should result in a straight line through the origin with a unit slope if the CS residuals are exponentially distributed as expected in a correctly specified survival model. Additionally, numerical goodness-of-fit tests, such as the Kolmogorov-Smirnov (KS) test [31], can be applied to assess the exponential distribution of CS residuals. However, in the presence of censored failure times, the distribution of $\hat{S}_{ij}(t)$ is no longer uniformly distributed, and the CS residuals lose their exponential distribution property. Nevertheless, we can still compute the Kaplan-Meier (KM) [23] estimate of the survivor function for CS residuals and compare it against the 45° straight line as a diagnostic tool for survival models.

Apart from transforming SPs into exponentially distributed CS residuals, there are other options available. For example, one can also transform SPs using the quantile of standard normal distribution [35], defined as $r^n_{ij}(t_{ij}) = -\Phi^{-1}(\hat{S}_{ij}(t_{ij}))$, $\hat{S}_{ij}(t_{ij})$ is estimated survival function of the jth individual from the ith cluster. We will call it **censored Z-residuals** in this paper. The diagnosis of the GOF of $S_{ij}(t_{ij})$ can be converted to the diagnosis of the normality of $r^n_{ij}(t_{ij})$. The function `gofTestCensored` in R package `EnvStats` [34] provides an SF test for testing the normality of multiply censored data. Hence, `gofTestCensored` can be applied to check the normality of censored Z-residuals for checking the overall GOF of survival models. We will refer to this test using the **CZ-CSF** method in this paper.

While these overall GOF checking methods evaluate the residuals' distribution, they cannot assess specific model assumptions, such as the functional form of covariates. To assess specific model assumptions, such as the functional form of covariates, tailored graphical and quantitative diagnostics tools are required. Two commonly used residuals for this purpose are martingale and deviance residuals. Martingale residuals [43] measure the discrepancy between the predicted and observed number of deaths in the interval $(0, T_{ij})$, taking values of 1 or 0. They are defined as $r^M_{ij} = \delta_{ij} - r^c_{ij}$, where δ_{ij} is the event indicator (1 for an event, 0 for censored), and r^c_{ij} is the Cox-Snell residual. Martingale residuals sum to zero but are not symmetrically distributed about zero [5]. Deviance residuals [32, 42] aim to achieve symmetry and are defined as $r^D_{ij} = sgn(r^M_{ij})[-2(r^M_{ij} + \delta_{ij}\log(\delta_{ij} - r^M_{ij}))]^{\frac{1}{2}}$, where r^M_{ij} is the martingale residual, and $sgn(.)$ is the sign function [5]. While other residual-based diagnostic tools have been proposed for censored survival models [9, 16, 17, 20, 25, 26, 29, 36, 40] a common drawback is their complicated distributions under the true model due to censoring. As a result, these residuals cannot be characterized by known distributions or probability tables, making it challenging to devise numerical tests based on them for diagnosing survival models.

4. Z-residual

4.1. Definition of Z-residual

In this paper, we extend the concept of Z-residual [27], to diagnose shared frailty models in a Cox proportional hazard setting with an unspecified baseline function. The normalized randomized survival probabilities (RSPs) for t_{ij} in the shared frailty model are defined as:

$$S_{ij}^{R}(t_{ij},\delta_{ij},U_{ij}) = \begin{cases} \hat{S}_{ij}(t_{ij}), & \text{if } t_{ij} \text{ is uncensored, i.e., } \delta_{ij}=1, \\ U_{ij}\,\hat{S}_{ij}(t_{ij}), & \text{if } t_{ij} \text{ is censored, i.e., } \delta_{ij}=0, \end{cases} \tag{3}$$

where U_{ij} is a uniform random number on the interval $(0,1)$, and $\hat{S}_{ij}(\cdot)$ is the estimated survival function for t_{ij} given x_{ij} and z_i. $S_{ij}^{R}(t_{ij},\delta_{ij},U_{ij})$ is a random number between 0 and $S_{ij}(t_{ij})$ when t_{ij} is censored. RSPs have been proven to be uniformly distributed on the interval $(0,1)$ given x_i under the true model [27] with independent survival times. In this paper, we extended this theory to scenarios with clustered survival times, demonstrating that RSPs defined with $\hat{S}_{ij}$'s, the survival functions conditional on the cluster indicators, are also independently and uniformly distributed on the interval $(0,1)$. A detailed theoretical proof of the uniformity of RSPs is provided in Section 4.2 below. Given the uniformity of the RSP under the correctly specified model, the transformation of RSPs into residuals with any desired distribution becomes feasible. For our analysis, we opt to perform this transformation using the normal quantile:

$$r_{ij}^{Z}(t_{ij},\delta_{ij},U_{ij}) = -\Phi^{-1}(S_{ij}^{R}(t_{ij},\delta_{ij},U_{ij})), \tag{4}$$

which yields a distribution that conforms to a normal distribution under the true model. Consequently, Z-residuals for censored data can be utilized for model diagnostics, akin to conducting diagnostics for a normal regression model. Transforming RSPs into Z-residuals offers several advantages. First, the diagnostics methods for checking normal regression are well-established in the literature. Second, transforming RSPs into normal deviates facilitates the identification of extremely small and large RSPs. The frequency of such small RSPs may be too low to be highlighted by the plots of RSPs alone. However, the presence of such extreme survival probabilities (SPs), even in very few instances, can indicate model misspecification. The normal transformation helps highlight these extreme RSPs. The Z-residual package can be downloaded from https://github.com/tiw150/Zresidual. We have also provided a detailed demonstration of how to use this package for detecting the covariate functional form, available from this link: https://tiw150.github.io/Zresidual_demo.html.

4.2. Proof of the Independence and Uniformity of RSPs

In a shared frailty model, the frailty term introduces cluster-specific variability. This implies that the survival function for each cluster may differ due to the presence of the random effect. The frailty term represents unobservable characteristics shared by individuals within the same cluster. This term is modeled as a random effect, capturing the between-cluster heterogeneity. Additionally, within each cluster, the failure times T_{ij}^{*} are assumed to be independent, given the true frailty value z_i and the covariate x_{ij}. The subscript ij in the survival function, $S_{ij}(\cdot)$, emphasizes that the survival function depends on the covariate x_{ij} and the cluster indicator.

In the following proof, we assume the true survival function $S_{ij}(t)$ for T^*_{ij} is applied to the definition of RSPs (equation (3)). Specifically, we will utilize true values for β and z_i in $S_{ij}(\cdot)$ (equation (2)). In computing the RSPs (equation (3)), we substitute the estimated regression coefficient vector $\hat{\beta}$ and frailties $\hat{z}_i$ into the survival function $S_{ij}(\cdot)$. When utilizing $\hat{\beta}$ and $\hat{z}_i$ in $S_{ij}(\cdot)$, the resulting RSPs exhibit an approximate uniform distribution between 0 and 1. However, in cases of small sample size or cluster size, caution is warranted due to potential optimistic bias resulting from the dual use of data for estimating $S_{ij}(\cdot)$ and validating the proposed model. A detailed discussion of this bias issue can be found in [46].

We first assume that there is no censoring and prove the uniformity and independence of $S_{ij}(T^*_{ij})$ under the true model for shared frailty models. The survival probabilities, $S_{ij}(T^*_{ij})$, as a function of the random variable T^*_{ij}, maintain uniformity over the interval $(0, 1)$. This uniformity can be proven using conditional probability as follows:

$$P(S_{ij}(T^*_{ij}) < t|\ x_{ij}, z_i) = P(T^*_{ij} > S^{-1}_{ij}(t)|\ x_{ij}, z_i) = S_{ij}(S^{-1}_{ij}(t)) = t, \text{ where } t \in (0, 1). \tag{5}$$

This equation indicates that, $S_{ij}(T^*_{ij})$, for $j = 1, \ldots, n_i$, are independent conditional on covariates x_{ij} and z_i, since T^*_{ij}, for $j = 1, \ldots, n_i$, are independent within a cluster, and the T^*_{ij}'s across clusters are independent of each other.

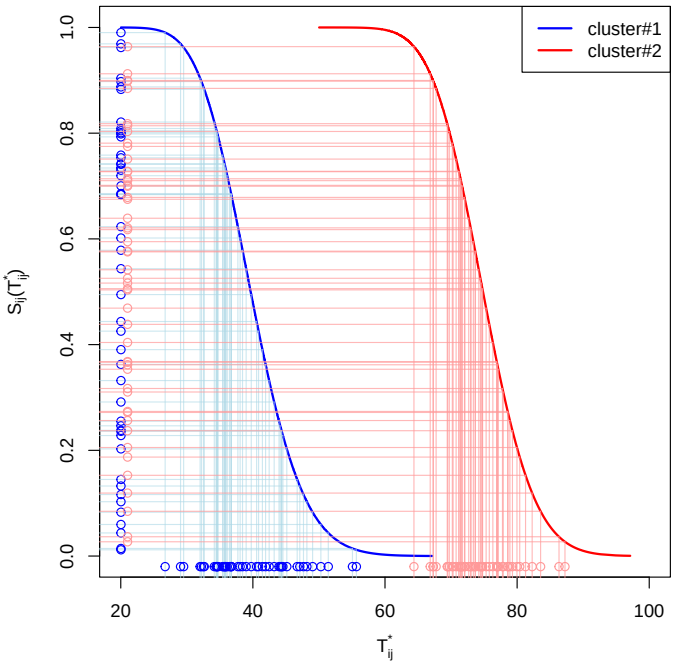

Figure 1. Illustration of the uniformity of SPs based on synthetic data simulated from two gamma distributions for two clusters. The colours of the points depict their cluster indicators.

To illustrate the above probabilistic statement, we simulated synthetic true survival times from a Gamma distribution for two cluster effects but no covariate. The first cluster has a true mean survival time of 40 with a variance of 40, while the second

cluster has a true mean survival time of 75 with a variance of 37.5. We then randomly simulated 50 data points from the true distributions for each cluster and calculated the corresponding survival probabilities. Figure 1 shows the scatterplots of T^*_{ij} and $S_{ij}(T^*_{ij})$. This figure illustrates that the survival probabilities of each cluster follow the uniform distribution on $(0, 1)$ although the failure times of the two clusters have different means. This demonstrates that the uniformity and independence of cumulative probabilities persist irrespective of the specific form of the survival function and the presence of frailty.

Due to the uniformity of $S_{ij}(T^*_{ij})$, we can transform $S_{ij}(T^*_{ij})$ into a random variable following any desired distribution with its quantile function. In particular, the Z-residual Z_{ij} for the true failure time T^*_{ij} as defined by $-\Phi^{-1}(S_{ij}(T^*_{ij}))$ is distributed as the standard normal. Z-residuals are akin to conditional Pearson residuals in linear mixed-effects models. When the model is correctly specified and underlying assumptions are met, conditional Pearson residuals tend to be approximately independent. This independence stems from conditioning on both fixed and random effects, mitigating correlation induced by the data's hierarchical structure. We illustrate the connection between Z-residuals and conditional Pearson residuals using log-normal models. Let T^*_{ij} follow a lognormal distribution with $\log(T^*_{ij})$ being normally distributed with mean μ_{ij} and standard deviation σ_{ij}. In this scenario, the survival function $S_{ij}(T^*_{ij})$ is given as $S_{ij}(T^*_{ij}) = 1 - \Phi\left(\frac{\log(T^*_{ij})-\mu_{ij}}{\sigma_{ij}}\right)$, where Φ is the cumulative distribution function of the standard normal distribution. The Z-residual for this scenario can be derived as:

$$Z_{ij} = -\Phi^{-1}(S_{ij}(T^*_{ij})) = -\left[\Phi^{-1}\left(1 - \Phi\left(\frac{\log(T^*_{ij}) - \mu_{ij}}{\sigma_{ij}}\right)\right)\right] = \frac{\log(T^*_{ij}) - \mu_{ij}}{\sigma_{ij}}.$$

This equation elucidates the relationship between Z-residuals Z_{ij} and conditional Pearson residuals. The connection between Z-residuals and Pearson residuals in survival analysis can be adapted to different distributions. However, the transformation specifics depend on the characteristics of the chosen distribution and typically do not have a closed-form formula as for log-normal models.

Now, we consider the scenario with censoring. We assume that T^*_{ij} and C_{ij} are independent, indicating that C_{ij} is non-informative for the original failure times. We redefine the RSP in equation (3) with the true survival function $S_{ij}(t)$. Furthermore, this version of RSP is a function of the following random variables: the original uncensored failure time T^*_{ij}, censoring time C_{ij}, and a uniform random number U_{ij}, as follows:

$$S^{R'}_{ij}(T^*_{ij}, C_{ij}, U_{ij}) = \begin{cases} S_{ij}(T^*_{ij}), & \text{if } T^*_{ij} \le C_{ij} \\ U_{ij}\, S_{ij}(C_{ij}), & \text{if } T^*_{ij} > C_{ij}. \end{cases} \tag{6}$$

Given covariates and cluster indicators, T^*_{ij}'s are independently distributed and are independent of the censoring times C_{ij}'s. Hence, the distribution of $S_{ij}(T^*_{ij})$ given $C_{ij} = c$ is also a uniform distribution. More specifically, given $T^*_{ij} \le c$, the RSP, $S_{ij}(T^*_{ij})$ (equation (6)), is uniformly distributed on $(S_{ij}(c), 1)$. For $T^*_{ij} > c$, the RSP is $U_{ij}\, S_{ij}(c)$, uniformly distributed on $(0, S_{ij}(c))$ due to the uniformity of U_{ij}. Using $\lambda(B)$ to denote the length of an interval B on $(0, 1)$, we derive the conditional probability

$P(S_{ij}^{R'}(T_{ij}^*, C_{ij}, U_{ij}) \in B \mid C_{ij} = c, x_{ij}, z_i)$:

$$
\begin{aligned}
& P(S_{ij}^{R'}(T_{ij}^*, C_{ij}, U_{ij}) \in B \mid C_{ij} = c, x_{ij}, z_i) \\
=& P(S_{ij}(T_{ij}^*) \in B \mid C_{ij} = c, x_{ij}, z_i, T_{ij}^* \leq c) \times P(T_{ij}^* \leq c \mid C_{ij} = c, x_{ij}, z_i) + \\
& P(U_{ij} S_{ij}(c) \in B \mid C_{ij} = c, x_{ij}, z_i, T_{ij}^* > c) \times P(T_{ij}^* > c \mid C_{ij} = c, x_{ij}, z_i) \\
=& \frac{\lambda(B \cap (S_{ij}(c), 1))}{1 - S_{ij}(c)} \times (1 - S_{ij}(c)) + \frac{\lambda(B \cap (0, S_{ij}(c)))}{S_{ij}(c)} \times S_{ij}(c) \\
=& \lambda(B \cap (S_{ij}(c), 1)) + \lambda(B \cap (0, S_{ij}(c))) \\
=& \lambda(B)
\end{aligned}
\tag{7}
$$

Now, having established that the conditional distribution of $S_{ij}^{R'}(T_{ij}^*, C_{ij}, U_{ij})$ given $C_{ij} = c$, x_{ij}, and z_i, is uniform on the interval $(0, 1)$. Consequently, we can extend this uniformity by applying the total probability rule and marginalizing away C_{ij}.This results in the marginal distribution of $S_{ij}^{R'}(T_{ij}^*, C_{ij}, U_{ij})$, which is also uniform on the interval$(0, 1)$. This completes the proof that the RSPs are uniformly distributed on $(0, 1)$ conditional on the cluster indicators. The independence of RSPs follows the independence of T_{ij}^* for different cases given the covariates and cluster indicators.

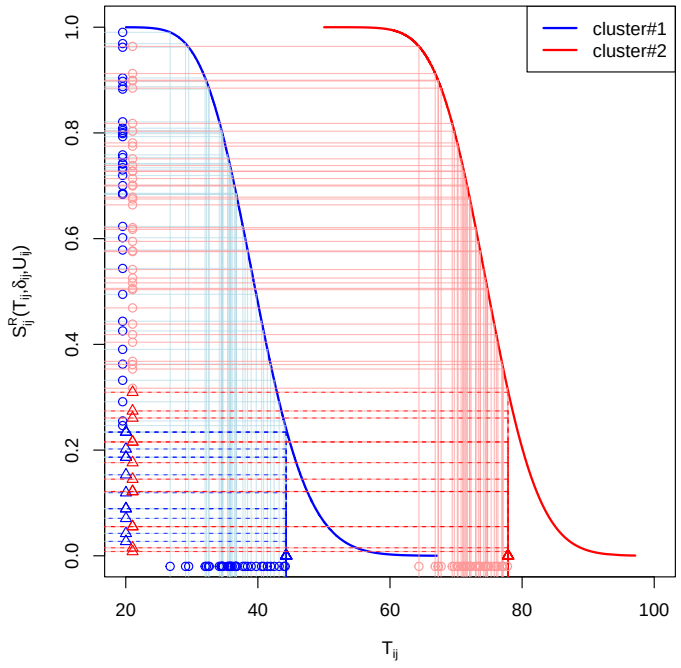

Figure 2. Illustration of the uniformity of RSPs based on synthetic data simulated from two gamma distributions for two clusters. The colours of the points depict their cluster indicators.

To enhance understanding, Figure 2 provides a visual representation of the foundational concepts supporting the independence and uniform distribution of RSPs within the true model. The synthetic dataset used for this illustration aligns with the one used to demonstrate the uniformity and independence of SPs. However, in this case,

T_{ij}^* is censored when it exceeds its fourth quartile for each cluster. The figure illustrates that for these censored observations, RSPs are randomly distributed between 0 and $S_{ij}(C_{ij})$. Consequently, RSPs are independent and follow a uniform distribution under the true model.

4.3. Diagnosis of the Functional Form of Covariates using Z-residuals

Z-residuals offer a powerful tool for diagnosing the functional form of covariates in survival models. To assess the model's overall goodness-of-fit (GOF), a QQ plot based on Z-residuals can be used graphically. Additionally, Shapiro-Wilk (SW) or Shapiro-Francia (SF) normality tests can be applied to Z-residuals to conduct numerical tests for the overall GOF. To specifically check the functional form of covariates, we can plot Z-residuals against the covariates and linear predictors. In a correctly specified model, we expect no discernible trend in these scatterplots. However, visually inspecting these plots may not be sufficient to determine if any observed trend is due to the chance or a genuine misspecification in the covariate function. Hence, we propose a formal test for this purpose.

The Z-residuals can be divided into k groups by equally spacing the covariates or linear predictors, as demonstrated in Figure 3. This figure shows two scatterplots of Z-residuals from two models: a linear effect model (the left plot) and a nonlinear effect model (the right plot). The covariate x_{ij} is sampled from a positive Normal(0, 1) distribution, and we generate the failure times t_{ij} from a shared frailty model with the hazard function $h_{ij}(t_{ij}) = z_i \exp(\beta \log(x_{ij}))h_0(t_{ij})$, where $h_0()$ represents the hazard function of Weibull distribution with shape $\alpha = 3$ and scale $\lambda = 0.007$ and the frailty term z_i is generated from a gamma distribution with a variance of 0.5. In addition to fitting the nonlinear model with $\log(X)$ as a covariate to these datasets, we also fit a shared frailty gamma model assuming a linear effect for X as a linear model. Then we can check whether the Z-residuals of the k groups are homogeneously distributed.

By examining whether the Z-residuals within the k groups are homogeneously distributed, we can assess the functional form of the covariates. The left panel of Figure 3 shows Z-residuals randomly scattered without any apparent differential group means or variances, indicating homogeneity. In contrast, the right panel of Figure 3 exhibits non-homogeneity, where the group means of Z-residuals differ significantly. To quantify the homogeneity of grouped Z-residuals, we employ an F-test in ANOVA to test the equality of means among the groups. This allows us to formally assess whether the observed trend deviates significantly from the expected horizontal line at 0, helping detect potential misspecifications in the covariate function.

4.4. A P-value Upper Bound for Assessing Replicated Z-residuals GOF Test P-values

Conducting statistical tests with Z-residuals can be challenging due to the randomness in the test p-values. When we fit a model, we can generate multiple sets of Z-residuals and obtain replicated test p-values. To address this randomness, we can use an upper bound for the p-values. Let's consider $p_1, \ldots, p_J$ as replicated Z-residual statistical test p-values obtained from J replicated samples, each derived from a fitted model using the same dataset. Based on the distribution of order statistics of correlated random

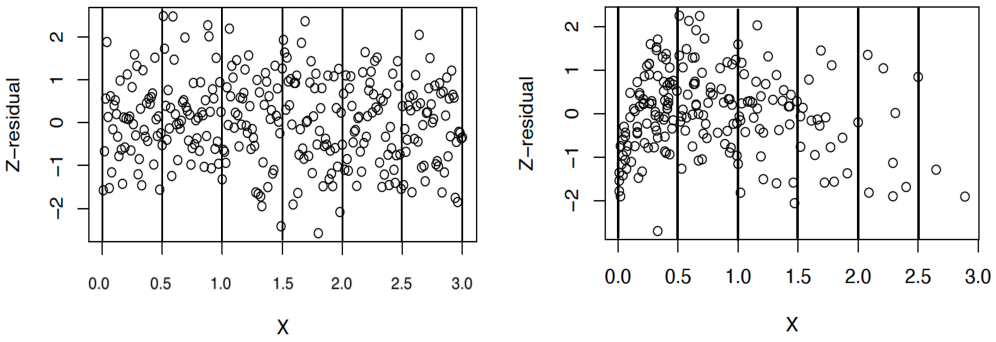

Figure 3. An illustrative plot demonstrating the non-homogeneity test with Z-residuals: Z-residuals divided by a covariate or linear predictor (LP) using equally spaced intervals, and the equality of means of grouped residuals is tested using an F-test in ANOVA.

variables [3, 37], we can derive the inequality for the rth order statistics $p_{(r)}$:

$$P(p_{(r)} < t) \leq \min\left(1, t\frac{J}{r}\right). \tag{8}$$

Using (8), we can obtain a p-value upper bound for the observed (simulated) rth statistics $p^{\text{obs}}_{(r)}$ is given by $\min\left(1, p^{\text{obs}}_{(r)}\frac{J}{r}\right)$. To avoid selecting a specific r, we calculate the minimal upper bound across $r = 1, \ldots, J$, denoted as $p_{\min}$:

$$p_{\min} = \min_{r=1,\ldots,J} \min\left(1, p^{\text{obs}}_{(r)}\frac{J}{r}\right). \tag{9}$$

The $p_{\min}$ provides a conservative measure for assessing model fit. When $p_{\min}$ is small, it suggests that the model can be improved to better fit the dataset. Considering the conservatism of $p_{\min}$, a rule of thumb for declaring model failure in practice should be much larger, say 0.25 as suggested by [48], rather than the conventional threshold of 0.05 for exact p-values.

5. Simulation Studies

In this section, we conducted simulation studies to evaluate the effectiveness of Z-residuals in checking the adequacy of the functional form of covariates in survival analysis. The simulation setup involved generating three covariates: $x^{(1)}_{ij}$ is from a Uniform[0, 1] distribution, $x^{(2)}_{ij}$ from a positive Normal distribution with mean 0 and standard deviation 1, and $x^{(3)}_{ij}$ from a Bernoulli distribution with a probability of success 0.25. The failure times t_{ij} were generated from a shared frailty model with a Weibull baseline hazard function. The hazard function is given by:

$$h_{ij}(t) = z_i \exp(\beta_1 x^{(1)}_{ij} + \beta_2 \log(x^{(2)}_{ij}) + \beta_3 x^{(3)}_{ij}) h_0(t), \tag{10}$$

where h_0 is the hazard function of Weibull with shape parameter $\alpha = 3$ and scale parameter $\lambda = 0.007$. The true survival time is generated by:

$$t_{ij}^{*} = \left\{ \frac{-\log(u_{ij})}{\lambda z_i \exp(x_{ij}^{(1)} - 2\log(x_{ij}^{(2)}) + 0.5x_{ij}^{(3)})} \right\}^{(1/\alpha)}, \tag{11}$$

where u_{ij} is simulated from a Uniform(0, 1) distribution, and the frailty term z_i is generated from a gamma distribution with a variance of 0.5. The censoring times C_{ij} were simulated from exponential distributions with rates γ, resulting in different censoring rates: 0%, 20%, 50%, and 80%. We considered 20 clusters with varying cluster sizes (n_i) ranging from 10 to 100. In our investigation, we generate survival times and censoring times from continuous distributions, so the likelihood of having tied event times (simultaneous occurrences of events) is generally lower compared to situations where the survival times are discrete. As a result, the Breslow method is used to estimate the baseline cumulative CHF.

For each combination of cluster size and censoring rate, we generated 1000 datasets and fitted the true model, which includes the covariate $\log(x_2)$ with a log-linear effect, to these datasets. We also considered fitting a wrong model with linear effect for x_2 to investigate the performance of different diagnostics methods. We evaluated various diagnostic methods, including graphical and numerical tests based on Z-residuals, to detect the adequacy of the functional form of covariates. The graphical methods included CHFs of CS residuals and quantile-quantile (QQ) plots of Z-residuals. The numerical tests involved dividing Z-residuals into groups by cutting the linear predictor or the covariate $\log(x_2)$ into equally-spaced intervals and testing the homogeneity of Z-residuals across these groups using ANOVA. The performance of these methods was assessed by estimating model rejection rates, which correspond to the proportion of test p-values less than the nominal level of 0.05, for each combination of cluster size and censoring rate.

The graphical methods used for assessing the overall GOF of the survival model and diagnosing the misspecification of the functional form of covariates are presented in Figure 4. We focus on a single simulated dataset with 20 clusters, each containing 40 observations, and a censoring percentage of approximately 50%. The first row of Figure 4 shows the CHFs of CS residuals for both the true model and the wrong model with linear covariate effects. The CHFs of CS residuals align well along the 45° straight line for both models, indicating that the CS residuals cannot effectively detect the model misspecification caused by the wrong model with linear covariate effects. The second row of Figure 4 displays the QQ plots of Z-residuals for the true and wrong models. The points in the QQ plots for Z-residuals align very well along straight lines, indicating that the distributions of the Z-residuals under both the true and wrong models are very close to a normal distribution. Thus, the QQ plots of Z-residuals do not provide clear evidence of misspecification in the wrong model either. The third and fourth rows of Figure 4 demonstrate the advantage of examining the scatterplots of Z-residuals against the linear predictor for diagnosing the misspecification of the functional form of covariates. Under the true model, the Z-residuals are mostly bounded between -3 and 3, following the standard normal distribution without a visible trend. The LOWESS curve in the scatterplot under the true model is very close to the horizontal line at 0, indicating a good fit. However, for the wrong model with linear covariate effects, a clear non-linear trend in the Z-residuals is observed in the scatterplot. In the fourth row, Z-residuals are divided into 10 groups by cutting the linear predictors into equally spaced intervals. The scatterplot and the boxplot indicate that the Z-residuals

are homogeneous across groups under the true model, but exhibit differential group means under the wrong model. This discrepancy suggests that the model with linear covariate effects does not fit well with the dataset, and there is a misspecification in the functional form of the covariate $\log(x_2)$. Additionally, the scatterplots and grouped boxplots of Z-residuals against $\log(x_2)$ are shown in the fifth and sixth rows of Figure 4. The Z-residuals of the true model are fairly homogeneous against $\log(x_2)$, indicating a good fit for the true model. On the contrary, for the wrong model, a clear non-linear pattern is observed in the scatterplots, and there are differential group means in the boxplots against $\log(x_2)$. These plots further support the conclusion that the model with linear covariate effects is inadequate for the dataset. In summary, the graphical methods presented in Figure 4 effectively assess the overall GOF and diagnose the misspecification of the functional form of covariates. The scatterplots of Z-residuals against the linear predictor and the covariate $\log(x_2)$ are particularly useful in identifying misspecification and providing insights into model improvement.

The performance of martingale and deviance residuals in assessing the functional form of x_2 is shown in Figure 5. Under the true model, the martingale residuals are mostly within the interval (-4, 1), while the deviance residuals are more symmetrically distributed and mainly fall within the interval (-3, 3). The LOWESS curves in the scatterplots of martingale and deviance residuals under the true model are very close to horizontal lines, with a slight downward tilt on the right due to the increased censoring for cases with large $\log(x_2)$. In contrast, under the wrong model, the LOWESS curves show more pronounced non-horizontal trends in the scatterplots of martingale and deviance residuals. This comparison demonstrates that the scatterplots of martingale and deviance residuals can distinguish between the true and wrong models and confirm that the true model is a better fit for the dataset. However, due to the lack of numerical measures, it is challenging to determine whether the observed non-horizontal trend is caused by chance or due to a misspecified functional form for the covariate. Decisions based on visual inspection can be subjective.

To complement the graphical assessment, numerical tests with Z-residuals can be used, as Z-residuals are approximately distributed as the standard normal under the true model. We compare a set of residual-based testing methods for detecting the inadequacy of fitted models. The overall GOF test methods are denoted by "R-T" with "R" denoting the residual name and "T" denoting the test method. For example, Z-SW is the test method used to assess the normality of Z-residuals with the Shapiro-Wilk test. Additionally, CZ-CSF is the method used to test the normality of censored Z-residuals, implemented with `gofTestCensored` in the R package `EnvStats`. For detecting misspecification in the covariate functional form, Z-residuals can be divided into groups by cutting the linear predictor or a covariate into equally-spaced intervals, as shown in the boxplots of Figure 4. We can then test the homogeneity of Z-residuals across the groups. Z-AOV-LP is the method used to apply ANOVA and test the equality of the means of Z-residuals against the groups formed with the linear predictor (LP). Similarly, Z-AOV-$\log(x_2)$ is the method used to test the equality of the means of Z-residuals against the groups formed with the covariate covariate $\log(x_2)$.

In our simulation studies, we generated 1000 datasets for each combination of cluster size and censoring rate, as described earlier. Using these datasets generated from the true model under each scenario, we estimated the model rejection rate of each test method by calculating the proportion of test p-values that are less than 0.05. The results of all the considered test methods are shown in Figures 6. The non-homogeneity test methods, Z-AOV-LP and Z-AOV-x_2, demonstrated excellent performance in detecting non-linear covariate effects with very high true-positive rates (model rejection

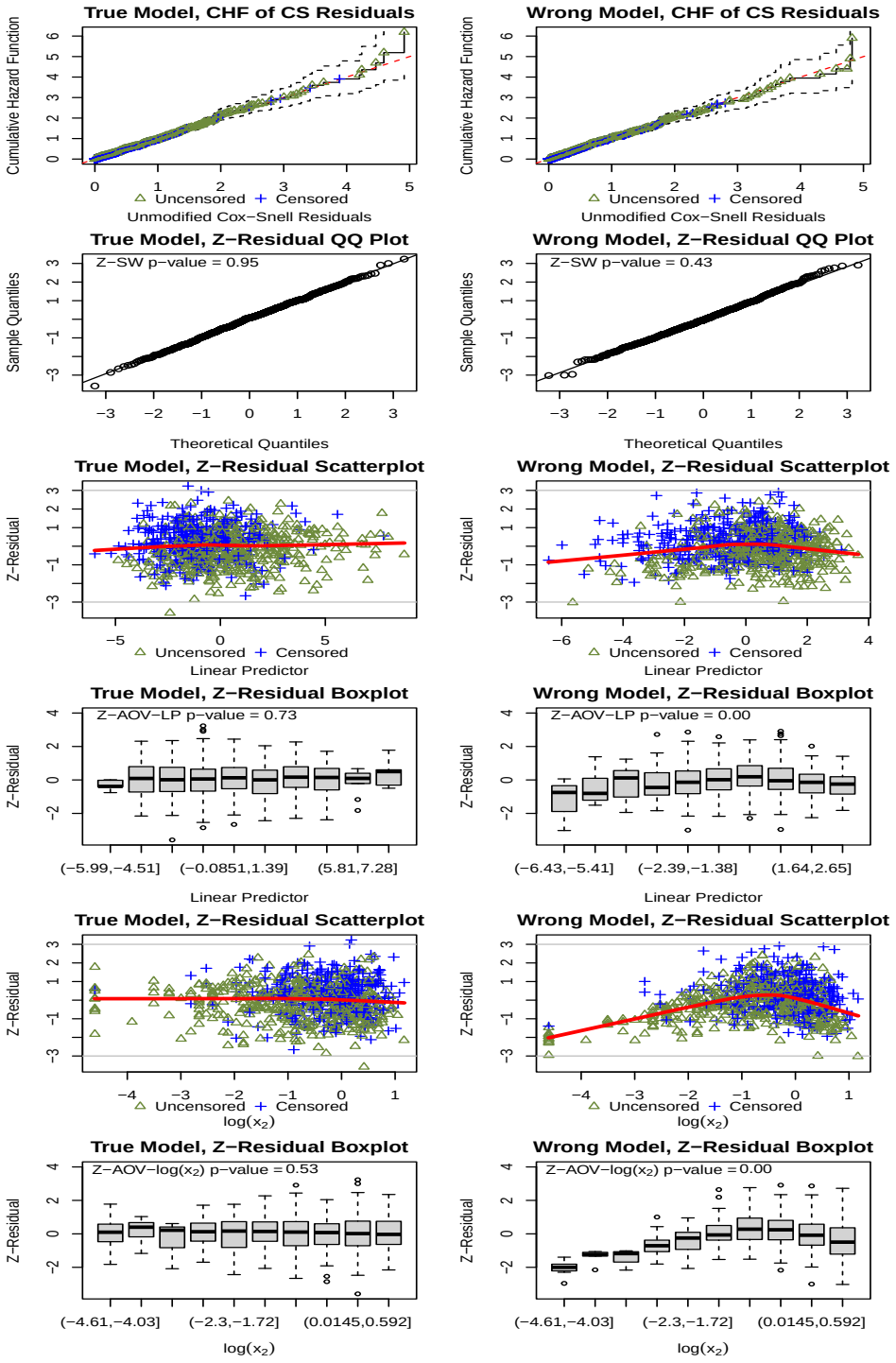

Figure 4. Performance of Z-residuals and Cox-Snell (CS) residuals as graphical tools for detecting the misspecification of the functional form of covariates. The dataset was generated with 20 clusters, each containing 40 observations, and a censoring rate of approximately 50%.

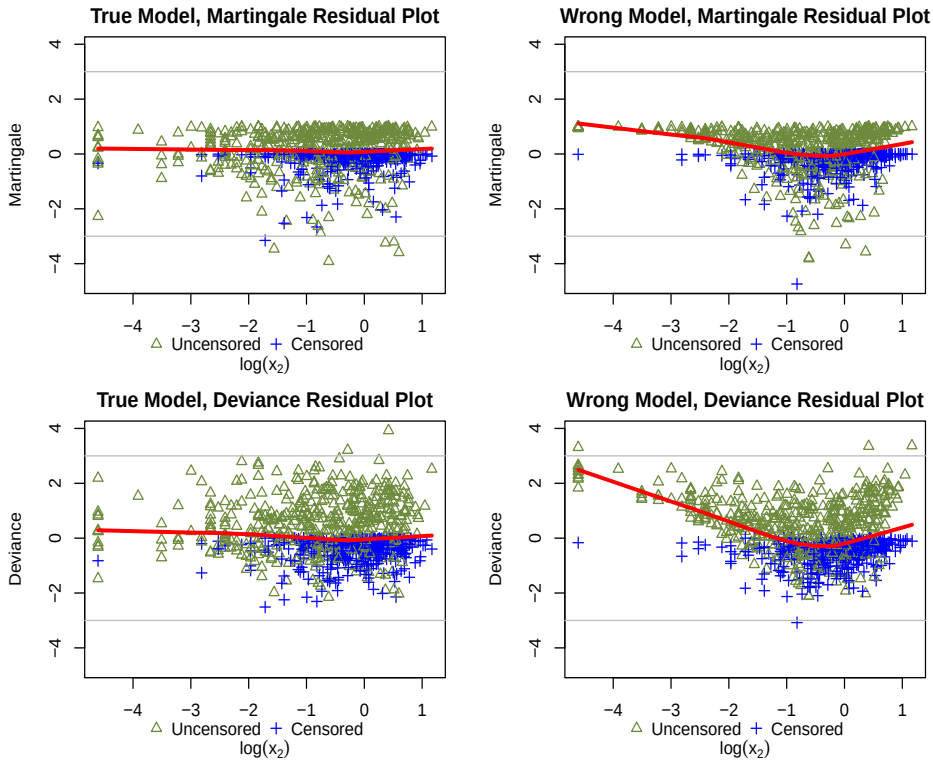

Figure 5. Performance of the martingale and deviance residuals as a graphical tool for checking the functional form of covariates. The dataset has a sample size n = 800 and a censoring rate $c \approx 50\%$.

rates under the wrong models) and low false-positive rates (model rejection rates under the true models). While the power of all the methods increased with cluster size, it is worth noting that the overall GOF tests had significantly smaller power compared to the Z-AOV-LP and Z-AOV-x_2 methods.

Among all the compared test methods, Z-AOV-x_2 performed the best for detecting the misspecified functional form for covariate x_2, achieving nearly 100% power. Our analysis revealed a notably liberal tendency of the Z-residual method under the true model conditions, as indicated by model rejection rates below the 5% nominal level. This suggests the method is less likely to incorrectly reject the true model, reducing Type I errors. Despite this liberal characteristic, the method maintained high diagnostic power, effectively identifying model misfits when present. This comparison highlights the advantage of testing the homogeneity of Z-residuals for checking the assumption of covariate functional form, in addition to the overall GOF tests, which do not inspect the relationship between residuals and covariates. Figure 6 also illustrates that as the censoring rate increases, the type I error rate for all methods remains at nominal levels at 5%. However, the power of all the residual diagnosis methods decreases due to the decreased number of non-censored observations and the randomness introduced in RSPs.

In Figure A.1 in the Appendix, we show the performances of the Z-KS and Dev-SW tests, presented separately from Figure 6 for improved clarity. The Z-KS test demonstrated low false-positive rates but also very low powers, indicating the conservatism of the KS test for testing the normality of Z-residuals. The Dev-SW method showed satisfactory performance when there was no censorship. However, when there were censored observations, the Dev-SW method exhibited very high (nearly 100%) model

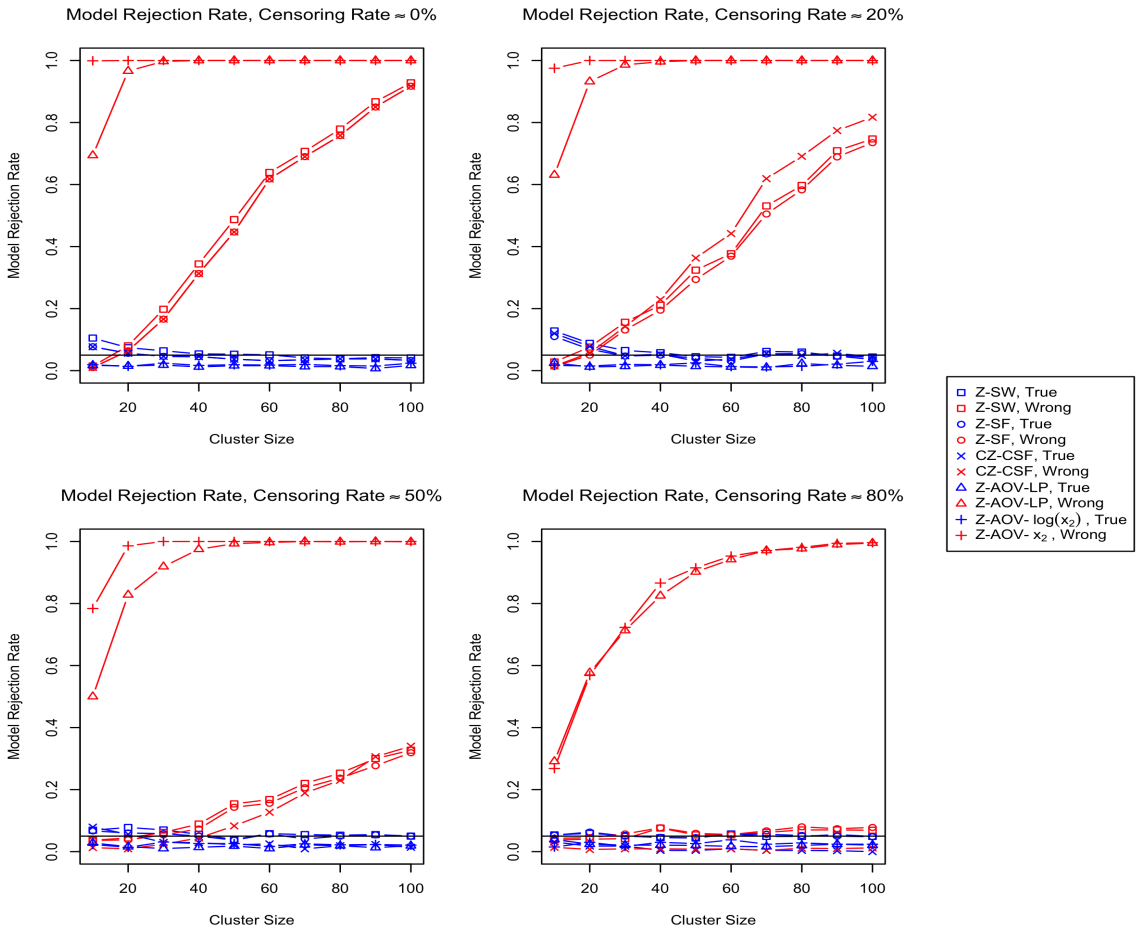

Figure 6. Model rejection rates of various statistical tests based on Z-residual for the simulation study with 20 clusters. A model is rejected when the test p-value is smaller than 5% (nominal level). Note that we use a random Z-residual test p-value rather than the $p_{\min}$. Detailed results corresponding to this figure can be found in Table A1 in the Appendix.

rejection rates under the correctly specified model. Hence, the high powers of Dev-SW do not indicate that it is a good test method in the presence of censored data.

Additionally, we extend our investigation to scenarios involving 10 and 30 clusters, each with varying cluster sizes n_i ranging from 10 to 100, alongside the original scenario with 20 clusters. This investigation aims to evaluate whether the performance of the Z-residual is affected by the number of clusters. Figures A.2 and A.3 in the appendix provide the results of the model rejection rate for each test method under the scenarios with 10 and 30 clusters. The findings generally align with those of the scenario involving 20 clusters. In general, as the number of clusters increases, the power of all the tests increases, and the type I error remains around the nominal level. The Z-AOV-x_2 method consistently outperforms the other methods. The performance improvement with increasing cluster numbers or sizes can be attributed to more accurate parameter estimates in the shared frailty model. With more clusters or larger cluster sizes, the model can better capture the variation between clusters, resulting in more reliable estimates of the random effects or frailty term. Consequently, the residuals may better reflect the true underlying variability in the data.

In our investigation, we also explored the impact of correlated covariates on the performance of the Z-residual. Covariates with correlation often arise in various fields, making it crucial to assess the robustness of our method under such conditions. Initially, our simulated dataset involves independent covariates under all scenarios. To specifically address scenarios with correlated covariates, we generate a dataset where the covariates x_1 and x_2 follow a multivariate normal distribution with a mean of 0, standard deviation of 2, and a correlation of 0.5. This introduces a level of correlation between the covariates, simulating a common scenario encountered in different studies. The other generating components of the dataset remain consistent with the previous simulation settings. To comprehensively evaluate the performance, we replicate this data-generating process 1000 times for the combination of 20 clusters with varying sizes and censoring rates, fitting both the true and wrong models. The true model includes the covariate $\log(x_2)$ with a log-linear effect, while the wrong model includes x_2 with a linear effect. Figures A.4 in the appendix present the results of the model rejection rate for the scenario involving correlated covariates. The findings suggest that the results are generally consistent with the independent covariate scenarios, but the model rejection rates of the Z-SW-LP are slightly lower when the wrong model is fitted to the dataset.

6. A Real Data Example

Section 6 presents an application of the proposed residual diagnostic tools based on Z-residuals to diagnose the functional form of covariates in a real dataset of acute myeloid leukemia patients. The dataset consists of 411 patients below the age of 60 from 24 administrative districts, recorded at the M.D. Anderson Cancer Center between 1980 and 1996. The dataset includes survival times for acute myeloid leukemia patients and several prognostic factors, such as age, sex, white blood cell count (wbc) at diagnosis, and the townsend score (tpi) indicating the affluence level of areas. The censoring rate in the dataset is 29.2%. The response variable of interest is the survival time in days, which is the time from entry to the study or death. In cancer research, white blood cell count is often considered an important marker of immune response and overall health. The preliminary study showed that the wbc is highly right-skewed. Logarithmic transformation is often used to reduce the impact of extremely large

values of the covariate on the response variable, such as the wbc variable in this application. However, using a logarithmic transformation may obscure the impact of extreme values of the covariate on the outcome variable.

Two shared frailty models are fitted to the data: one with the original wbc covariate and the other with log(wbc) as a replacement for wbc. These models are labelled as the wbc model and the lwbc model, respectively. Table 1 displays the estimated regression coefficients, their corresponding standard errors, and p-values for the covariate effects in both models. The results indicate that the estimated effect of wbc is statistically significant (p-value < 0.001), whereas the effect of log(wbc) is not significant (p-value $= 0.135$). This difference in p-values highlights that the statistical inference of the covariate effect may depend on the assumption of the functional form of the covariates.

Table 1. Parameter estimates of the shared gamma frailty model in the real data application. The left table represents the wbc model, and the right table corresponds to the lwbc model.

Covariates	Estimate	SE	P-value
Age	0.021	0.005	0.000
$SexMale$	0.215	0.118	0.068
wbc	0.005	0.001	0.000
tpi	0.023	0.016	0.140
$Frailty$			0.906

Covariates	Estimate	SE	P-value
Age	0.021	0.005	0.000
$SexMale$	0.216	0.118	0.069
$log(wbc)$	0.035	0.024	0.135
tpi	0.024	0.016	0.128
$Frailty$			0.906

The overall GOF tests and graphical checks with CS residuals and Z-residuals indicate that both the wbc and lwbc models provide adequate fits to the dataset. In the first row of Figure 7, the estimated CHFs of the CS residuals for both models closely align along the 45° diagonal line. Similarly, the QQ plots of Z-residuals (the second row of Figure 7) for both models show good alignment with the 45° diagonal line. The scatterplots of Z-residuals against the linear predictor show no visible trends, and the LOWESS lines are very close to the horizontal line at 0. The boxplots of Z-residuals grouped by cutting linear predictors into equal-spaced intervals (the fourth row of Figure 7) indicate approximately equal means and variances across groups. The Z-AOV-LP test also yields large p-values for both the wbc and lwbc models (0.59 and 0.78, respectively).

The above diagnostic results reveal no significant misspecification in either of these two models. However, further inspection of the Z-residuals against the covariate log(wbc) suggests that the functional form of the lwbc model may be misspecified. The scatterplots and comparative boxplots of the Z-residuals against log(wbc) are shown in the fifth and sixth rows of Figure 7. The LOWESS curve of the wbc model appears to align well with the horizontal line at 0, and the grouped Z-residuals of the wbc model seem to have approximately equal means and variances across groups. On the other hand, the diagnostic results for the lwbc model show a different pattern. There seems to be a non-linear trend in the LOWESS curve, and the grouped Z-residuals appear to have different means across groups. To assess the statistical significance of these observed trends, we use the Z-AOV-log(wbc) test to test the equality of means of the grouped Z-residuals for these two models. The resulting p-values are 0.16 and < 0.01, respectively, for the wbc and lwbc models, as shown in the boxplots. The very small p-value for the Z-AOV-log(wbc) test for the lwbc model strongly suggests that the log transformation of wbc is likely inappropriate for modelling the survival time.

The Z-residual test p-values quoted above contain randomness because of the randomization in generating Z-residuals. To ensure the robustness of the model diagnostic results, we generated 1000 replicated test p-values with 1000 sets of regenerated Z-

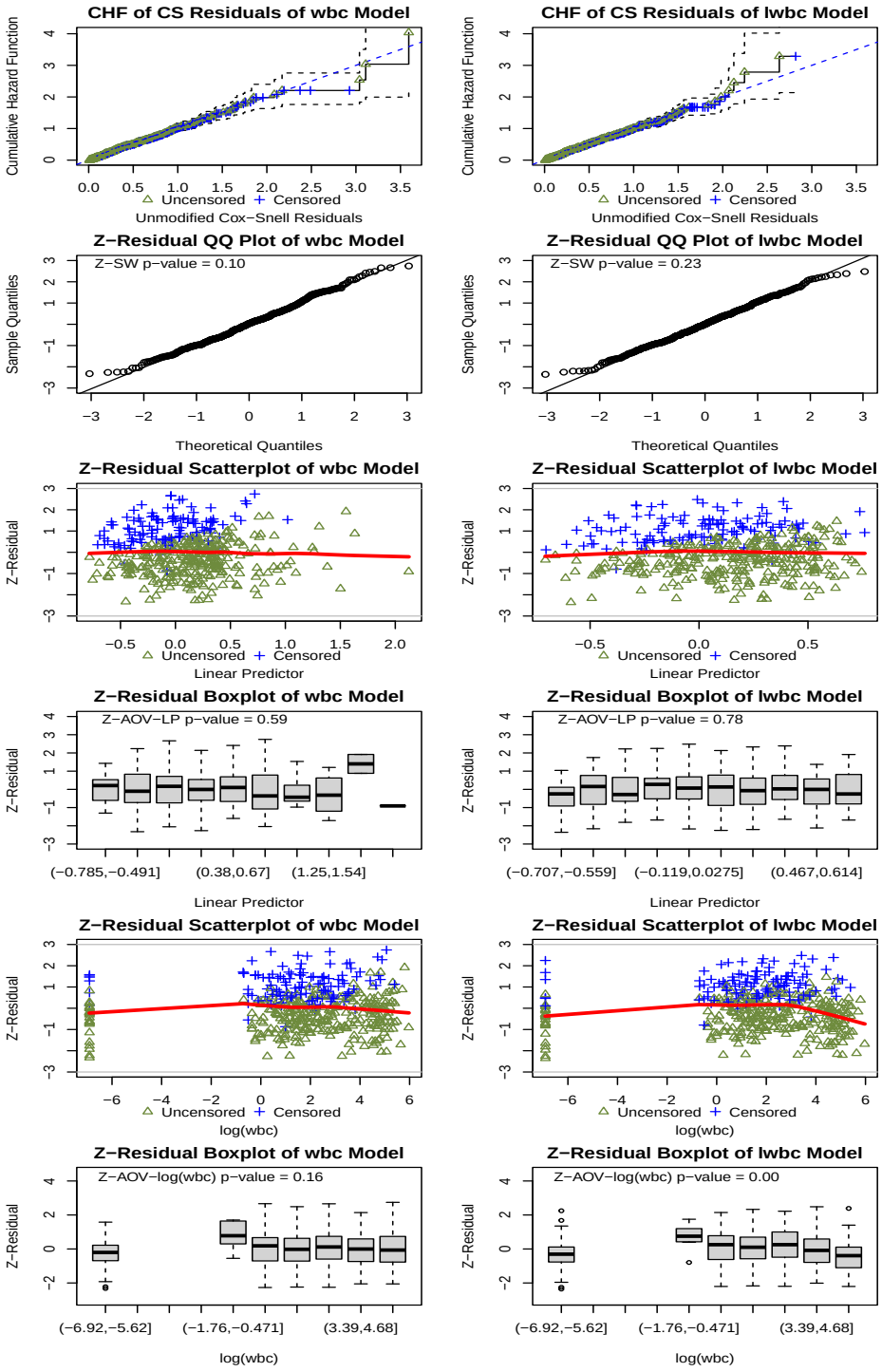

Figure 7. Diagnostics results for the wbc (left panels) and lwbc (right panels) models fitted to the survival data of acute myeloid leukemia patients.

Table 2. AIC, p-values or $p_{\min}$ values for the CZ-CSF test, $p_{\min}$ for Z-SW, Z-SF, Z-AOV-LP and Z-AOV-log(wbc) test for the wbc and lwbc models, respectively, for the acute myeloid leukemia data.

Model	AIC	CZ-CSF $p-$value	Z-SW $p_{\min}$	Z-SF $p_{\min}$	Z-AOV-LP $p_{\min}$	Z-AOV-log(wbc) $p_{\min}$
wbc model	3111.669	0.255	0.495	0.693	0.703	0.074
lwbc model	3132.105	0.305	0.579	0.781	0.978	$<$**0.00001**

residuals for each test method. Figure 8 displays the histograms of the 1000 replicated Z-residual test p-values for both the wbc and lwbc models. The red vertical lines in these histograms represent the upper bound summaries of these replicated p-values, denoted as, $p_{\min}$ (see Section 4.4 for details). These histograms reveal that the Z-SW, Z-SF, and Z-AOV-LP tests for both models have a substantial proportion of p-values greater than 0.05, leading to large $p_{\min}$ values. In contrast, the replicated Z-AOV-log(wbc) p-values for the lwbc model are nearly all smaller than 0.001. These consistently small Z-AOV-log(wbc) p-values provide strong evidence that the log transformation of wbc is inappropriate for modelling the survival time.

Table 2 presents all the $p_{\min}$ values (indicated with red lines in Figure 8) obtained from diagnosing the two models using Z-residual-based tests. Additionally, the table includes the non-random CZ-CSF test p-values for both models and the AIC values for model comparison. The CZ-CSF p-values for both models are greater than 5% (Table 2), indicating that the CZ-CSF test does not identify the inadequacy of the lwbc model either. Comparing the AIC values, the lwbc model with an AIC value of 3132.105 is much larger than the wbc model's AIC value of 3111.669. This conclusion is consistent with the model diagnostics results as given by the Z-AOV-log(wbc) test, which reveals that the lwbc model is inappropriate for modelling the survival time of this dataset by checking the homogeneity of Z-residuals against log(wbc). Although the AIC of the wbc model is smaller than that of the lwbc model, we also see that a large proportion of Z-AOV-log(wbc) p-values for the wbc model is quite small (e.g., $p_{\min}$ value of 0.074). This suggests that there is room for improvement in the wbc model to provide an even better fit for the survival time of this dataset.

7. Conclusions and Discussions

In this paper, we introduced an extension of the concept of randomized survival probability [27] to develop a novel residual diagnostic tool for assessing the covariate functional form in shared frailty models. The proposed Z-residuals offer valuable insights into model adequacy and provide both graphical and numerical methods to examine the fit of the model to the data. By plotting Z-residuals against covariates, we can visually inspect the presence of trends, allowing us to assess the appropriateness of the functional form assumptions. Moreover, we introduced a non-homogeneity test based on grouped Z-residuals, which helps us distinguish genuine misspecifications from chance variations. Our extensive simulation studies have convincingly demonstrated that the proposed non-homogeneity tests based on Z-residuals outperform traditional overall goodness-of-fit tests, such as CS-CSF, Z-SW, and Z-SF, especially in scenarios with complex covariate functional forms. Importantly, Z-residuals offer a powerful advantage in detecting specific model misspecifications that might be overlooked by traditional overall model diagnosis. The ability to visually inspect trends and apply non-homogeneity tests to grouped Z-residuals allows us to identify subtle

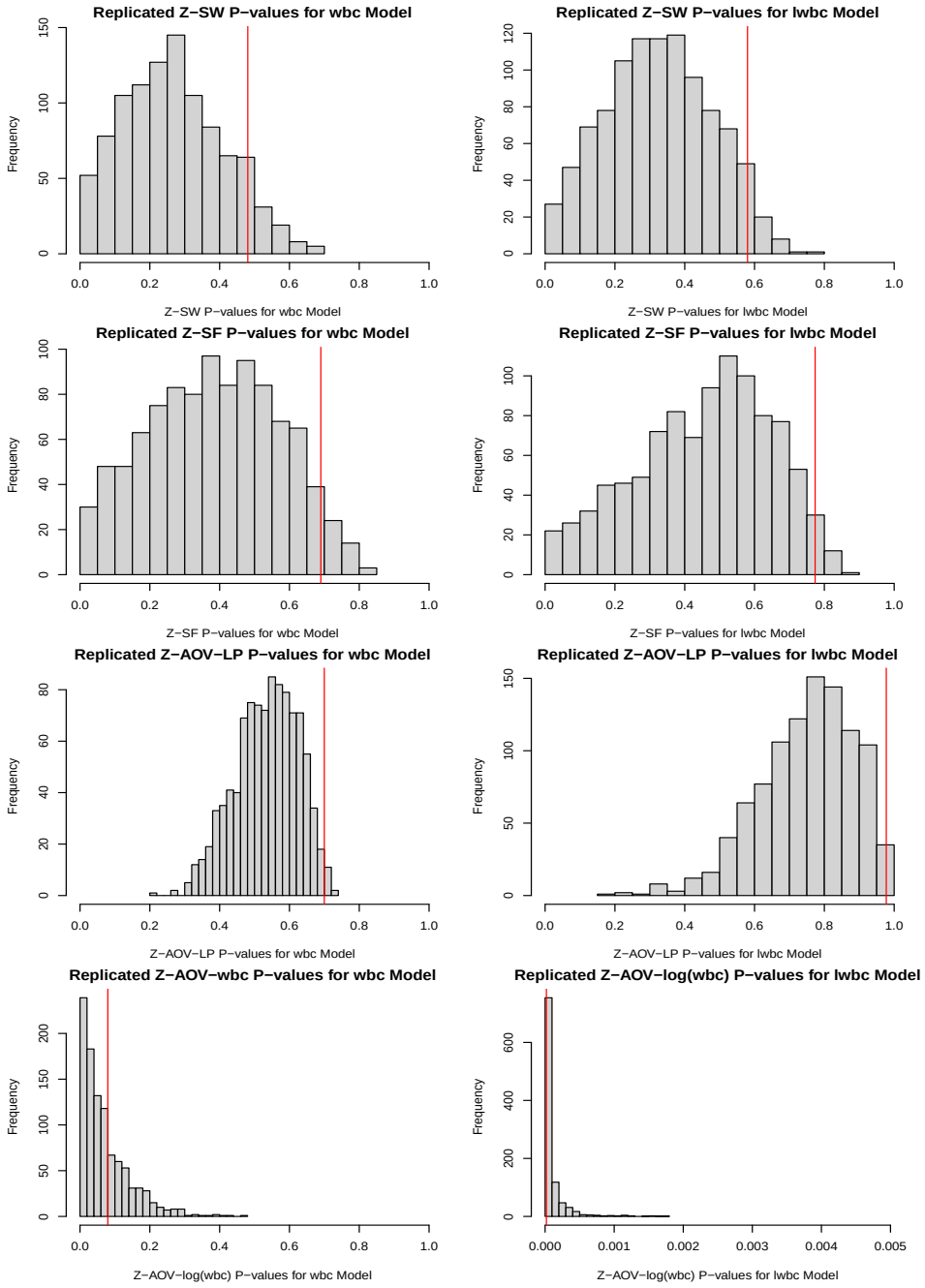

Figure 8. The histograms of 1000 replicated Z-SW, Z-SF, Z-AOV-LP and Z-AOV-log(wbc) p-values for the wbc model (left panels) and the lwbc model (right panels) fitted with the survival times of acute myeloid leukemia patients. The vertical red lines indicate $p_{\min}$ for 1000 replicated p-values. Note that the upper limit of the x-axis for Z-AOV-log(wbc) p-values for the lwbc model is 0.005, and 1 for others.

deviations from the expected fit, pinpointing precise areas of model inadequacy. This capability makes Z-residuals an invaluable tool for uncovering hidden relationships and providing insights into the functional form of covariates.

Furthermore, we have applied the Z-residual diagnostics to a real dataset of acute myeloid leukemia patients, wherein we discovered that a model with log-transformation is not appropriate for modelling survival time. This critical insight was not captured by other diagnostic methods, highlighting the significance of Z-residuals in real-world applications. The analysis of the white blood cell count variable, a commonly used marker in cancer research, serves as an essential illustration of the importance of carefully considering the choice of covariate transformations. While logarithmic transformation is often applied to mitigate the impact of highly skewed data, it may not always be appropriate, as demonstrated in this example. The presence of large values, such as white blood cell count, could be highly informative for modelling adverse health outcomes. Logarithmic transformation may obscure or mask this valuable information, leading to potential misinterpretation of the covariate's impact on survival time. To preserve the meaningful relationships between covariates and survival time accurately, alternative modelling techniques, such as using splines or allowing for non-linear effects, may be more suitable for capturing the complex relationship between white blood cell count and survival time. Taking these factors into account is crucial to ensure that the model captures the underlying patterns in the data and provides reliable insights for clinical decision-making.

In our study, we conducted an additional simulation study specifically aimed at evaluating the performance of Z-residuals in detecting frailty distribution misspecification. The results, presented in Figure A.5 indicate that the tests based on Z-residuals have limited power for detecting frailty distribution misspecifications, particularly in scenarios with large cluster sizes. This aligns with the robustness observed in mixed-effects models, as documented in previous research [38]. From a Bayesian perspective, where the frailty distribution serves as the prior for random effects (frailties), our findings support the established understanding that the prior's impact on parameter estimation diminishes with moderately large sample sizes. Therefore, Z-residuals, being derived from the tail probabilities of the distributions for observations, exhibit limited power to detect misspecifications in the prior distribution that do not affect parameter estimation. In consideration of these findings, we aim to further explore and develop tailored statistical tests based on Z-residuals specifically for diagnosing misspecified frailty distributions in our future research endeavours.

Looking ahead, our research opens up promising avenues for others' further development and enhancement of Z-residuals-based diagnostics in survival analysis. As we have acknowledged, one potential concern arises from the double use of data in both model parameter estimation and residual calculation, which may lead to conservative estimates of model fit and misspecification detection. To address this issue effectively, we propose the incorporation of cross-validation techniques. By employing cross-validatory Z-residuals, we can achieve more robust and unbiased assessments of model adequacy, improving the accuracy and reliability of our diagnostic framework. This cross-validation approach is particularly beneficial for datasets with limited sample sizes or high levels of censoring, where traditional methods may fall short in detecting subtle misspecifications. Furthermore, the extension of Z-residuals to accommodate time-dependent covariate effects and non-proportional hazards models presents an exciting avenue for future research. In practice, survival data often exhibit time-varying relationships between covariates and the event of interest, necessitating a more flexible diagnostic tool. A number of residuals have been proposed for evaluat-

ing the assumption of proportional hazards, such as the Schoenfeld [5, 39] and Scaled Schoenfeld [17], as well as cumulative sums of martingale residuals [29]. Expanding the Z-residual approach to diagnose the proportional hazards assumption and comparing its performance with existing methods would offer researchers a comprehensive toolkit for assessing model assumptions in various time-dependent scenarios. Moreover, investigating the potential of Z-residuals in the context of joint modelling approaches would be of great interest. Joint models, which simultaneously analyze longitudinal and survival data, have gained popularity in recent years due to their ability to capture complex disease processes. Integrating Z-residuals into joint modelling can offer new insights into the adequacy of joint models and the functional form of longitudinal covariates, further extending the applicability of Z-residuals in survival analysis.

Availability of Software and Datasets

We have developed an R package called `Z-residual`, which can be downloaded from https://github.com/tiw150/Zresidual. We have also provided a detailed demonstration of how to use this package for detecting the covariate functional form, available from this link: https://tiw150.github.io/Zresidual_demo.html.

Acknowledgement(s)

The authors gratefully acknowledge the support provided by the Natural Sciences and Engineering Research Council of Canada.

Disclosure statement

No Conflict of Interests.

Funding

This research was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] T.A. Balan and H. Putter, *A tutorial on frailty models*, Statistical Methods in Medical Research 29 (2020), pp. 3424–3454. Available at https://doi.org/10.1177/0962280220921889, PMID: 32466712.

[2] T.A. Balan and H. Putter, *frailtyEM : An R Package for estimating semiparametric shared frailty models*, Journal of statistical software 90 (2019), pp. 1–29.

[3] G. Caraux and O. Gascuel, *Bounds on Distribution Functions of Order Statistics for Dependent Variates*, Statistics & Probability Letters 14 (1992), pp. 103–105.

[4] D.G. Clayton, *A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence*, Biometrika 65 (1978), pp. 141–151.

[5] D. Collett, *Modelling Survival Data in Medical Research*, Chapman and Hall/CRC, 2015.

[6] D.R. Cox, *Regression Models and Life-Tables*, Journal of the Royal Statistical Society. Series B, Methodological 34 (1972), pp. 187–220.
[7] D.R. Cox and E.J. Snell, *A General Definition of Residuals*, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 30 (1968), pp. 248–275.
[8] F.N. David and N.L. Johnson, *The probability integral transformation when parameters are estimated from the sample*, Biometrika 35 (1948), pp. 182–190.
[9] A.C. Davison and A. Gigli, *Deviance Residuals and Normal Scores Plots*, Biometrika 76 (1989), pp. 211–221.
[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B, Methodological 39 (1977), pp. 1–38.
[11] F. Downton, *Discussion on Professor Cox's Paper*, Journal of the Royal Statistical Society. Series B, Methodological 34 (1972), pp. 202–220.
[12] L. Duchateau and L.a. Duchateau, *The frailty model*, Statistics for biology and health, Springer Verlag, New York, 2008.
[13] B. Efron, *The Efficiency of Cox's Likelihood Function for Censored Data*, Journal of the American Statistical Association 72 (1977), pp. 557–565.
[14] B. Efron, *Censored Data and the Bootstrap*, Journal of the American Statistical Association 76 (1981), pp. 312–319.
[15] E.H. Estey, Y. Shen, and P.F. Thall, *Effect of time to complete remission on subsequent survival and disease-free survival time in AML, RAEB-t, and RAEB*, Blood 95 (2000), pp. 72–77.
[16] C.P. Farrington, *Residuals for Proportional Hazards Models with Interval-Censored Survival Data*, Biometrics 56 (2000), pp. 473–482.
[17] P.M. Grambsch and T.M. Therneau, *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals*, Biometrika 81 (1994), pp. 515–526.
[18] D. Hanagal, *Modeling survival data using frailty models*, Statistical methods in Medical Research 24 (2015), pp. 936–936.
[19] R. Henderson, S. Shimakura, and D. Gorst, *Modeling Spatial Variation in Leukemia Survival Data*, Journal of the American Statistical Association 97 (2002), pp. 965–972.
[20] S.L. Hillis, *Residual Plots for the Censored Data Linear Regression Model*, Statistics in Medicine 14 (1995), pp. 2023–2036.
[21] P. Hougaard, *Frailty models for survival data*, Lifetime data analysis 1 (1995), pp. 255–273.
[22] P. Hougaard, *Analysis of Multivariate Survival Data*, Springer, 2000.
[23] E.L. Kaplan and P. Meier, *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association 53 (1958), pp. 457–481.
[24] A. Karagrigoriou, *Frailty Models in Survival Analysis*, Journal of Applied Statistics 38 (2011), pp. 2988–2989.
[25] S. Keleş and M.R. Segal, *Residual-Based Tree-Structured Survival Analysis*, Statistics in Medicine 21 (2002), pp. 313–326.
[26] M. Law and D. Jackson, *Residual Plots for Linear Regression Models with Censored Outcome Data: A Refined Method for Visualizing Residual Uncertainty*, Communications in Statistics - Simulation and Computation 46 (2017), pp. 3159–3171.
[27] L. Li, T. Wu, and C. Feng, *Model diagnostics for censored regression via randomized survival probabilities*, Statistics in Medicine 40 (2021), pp. 1482–1497.
[28] D.Y. Lin, *On the Breslow estimator*, Lifetime data analysis 13 (2007), pp. 471–480.
[29] D.Y. Lin, L.J. Wei, and Z. Ying, *Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals*, Biometrika 80 (1993), pp. 557–572.
[30] D.Y. Lin, L.J. Wei, and Z. Ying, *Accelerated failure time models for counting processes*, Biometrika 85 (1998), pp. 605–618.
[31] F.J. Massey, *The Kolmogorov-Smirnov Test for Goodness of Fit*, Journal of the American Statistical Association 46 (1951), pp. 68–78.
[32] P. McCullagh and J.A. Nelder, *Generalized Linear Models, Second Edition*, CRC Press,

1989 Aug.
[33] C.A. McGilchrist, *REML Estimation for Survival Models with Frailty*, Biometrics 49 (1993), p. 221.
[34] S.P. Millard, *EnvStats: An R Package for Environmental Statistics*, 2nd ed., Springer, New York, NY, 2013.
[35] A. Nardi and M. Schemper, *New Residuals for Cox Regression and Their Application to Outlier Screening*, Biometrics 55 (1999), pp. 523–529.
[36] Y. Peng and J.M.G. Taylor, *Residual-Based Model Diagnosis Methods for Mixture Cure Models*, Biometrics 73 (2017), pp. 495–505.
[37] T. Rychlik, *Stochastically Extremal Distributions of Order Statistics for Dependent Samples*, Statistics & probability letters 13 (1992), pp. 337–341.
[38] H. Schielzeth, N.J. Dingemanse, S. Nakagawa, D.F. Westneat, H. Allegue, C. Teplitsky, D. Réale, N.A. Dochtermann, L.Z. Garamszegi, Y.G. Araya-Ajoy, and C. Sutherland, *Robustness of linear mixed-effects models to violations of distributional assumptions*, Methods in ecology and evolution 11 (2020), pp. 1141–1152.
[39] D. Schoenfeld, *Partial residuals for the proportional hazards regression model*, Biometrika 69 (1982), pp. 239–241.
[40] B.E. Shepherd, C. Li, and Q. Liu, *Probability-scale residuals for continuous, discrete, and censored data*, Canadian journal of statistics 44 (2016), pp. 463–479.
[41] T.M. Therneau, *A Package for Survival Analysis in R* (2022). Available at https://CRAN.R-project.org/package=survival, R package version 3.3-1.
[42] T.M. Therneau and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer Science & Business Media, 2013.
[43] T.M. Therneau, P.M. Grambsch, and T.R. Fleming, *Martingale-Based Residuals for Survival Models*, Biometrika 77 (1990), pp. 147–160.
[44] J.W. Vaupel, K.G. Manton, and E. Stallard, *The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality*, Demography 16 (1979), pp. 439–454.
[45] T. Wu, C. Feng, and L. Li, *A Comparison of Parameter Estimation Methods for Shared Frailty Models*, arXiv.org (2023). Available at https://doi.org/10.48550/arXiv.2311.11543.
[46] T. Wu, C. Feng, and L. Li, *Cross-validatory z-residual for diagnosing shared frailty models*, arXiv.org (2023). Available at https://doi.org/10.48550/arXiv.2303.09616.
[47] T. Yamaguchi and Y. Ohashi, *Investigating Centre Effects in a Multi-Centre Clinical Trial of Superficial Bladder Cancer*, Stat Med 18 (1999), pp. 1961–1971.
[48] Y. Yuan and V.E. Johnson, *Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models*, Biometrics 68 (2012), pp. 156–164.

Appendix A. Additional Figures

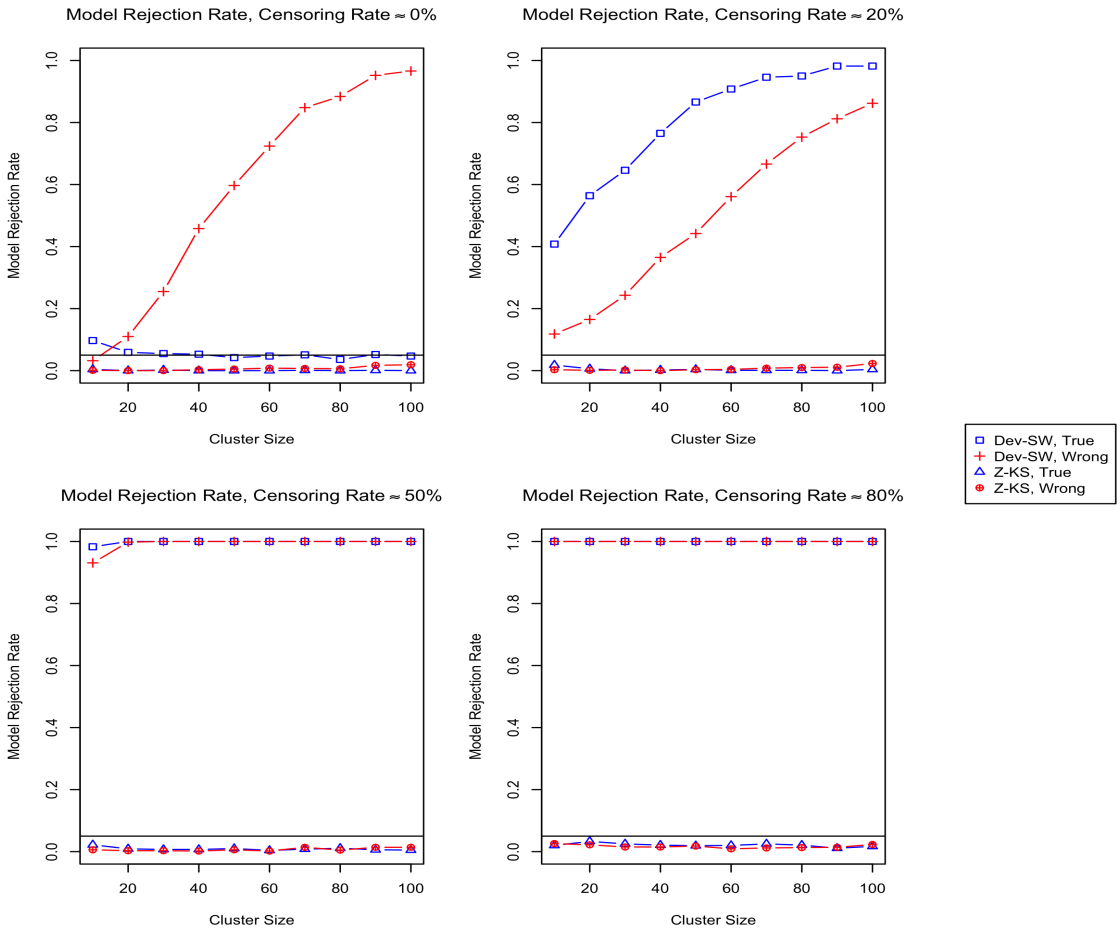

Figure A.1. Model rejection rate of the KS test applied to Z-residuals (Z-KS) and the SW test applied to deviance residuals (Dev-SW) for the simulation study in Section 4.3. A model is rejected when the test p-value is less than 5% (nominal level). The model rejection rates of Dev-SW tests are nearly 1 under the true and wrong models when the censoring rate is 50% and 80%, resulting in almost overlapped plots.

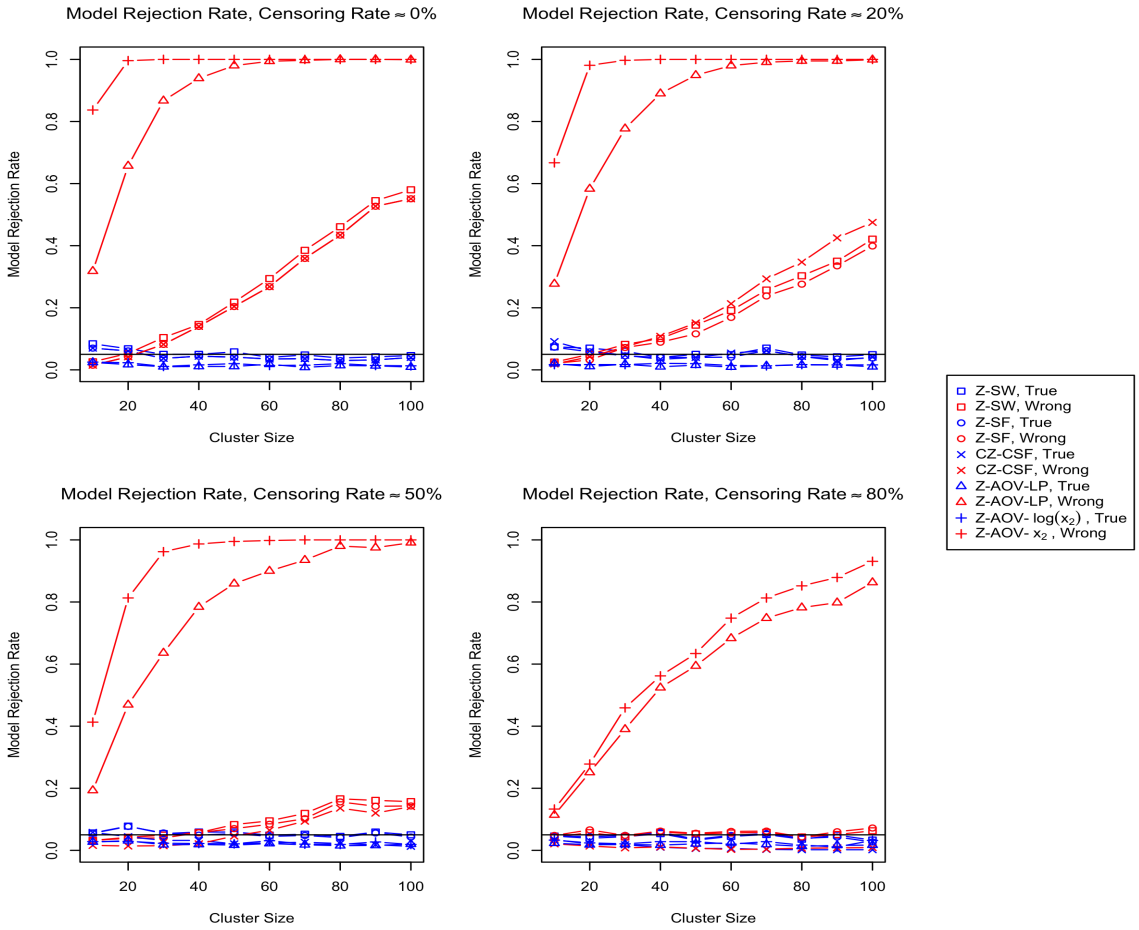

Figure A.2. Model rejection rates of various statistical tests based on Z-residuals for the simulation study with 10 clusters, as presented in Section 5. A model is rejected when the test p-value is less than 5% (nominal level).

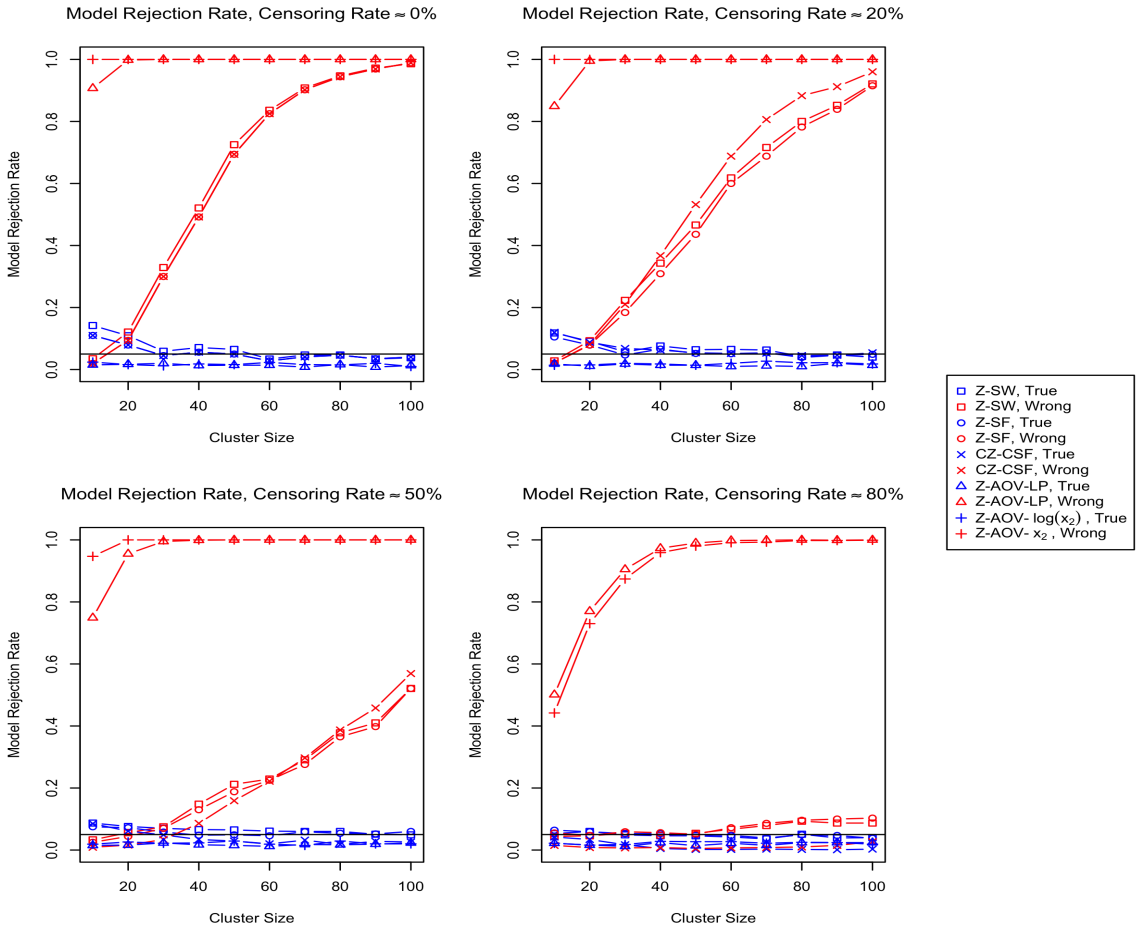

Figure A.3. Model rejection rates of various statistical tests based on Z-residuals for the simulation study with 30 clusters, as presented in Section 5. A model is rejected when the test p-value is less than 5% (nominal level).

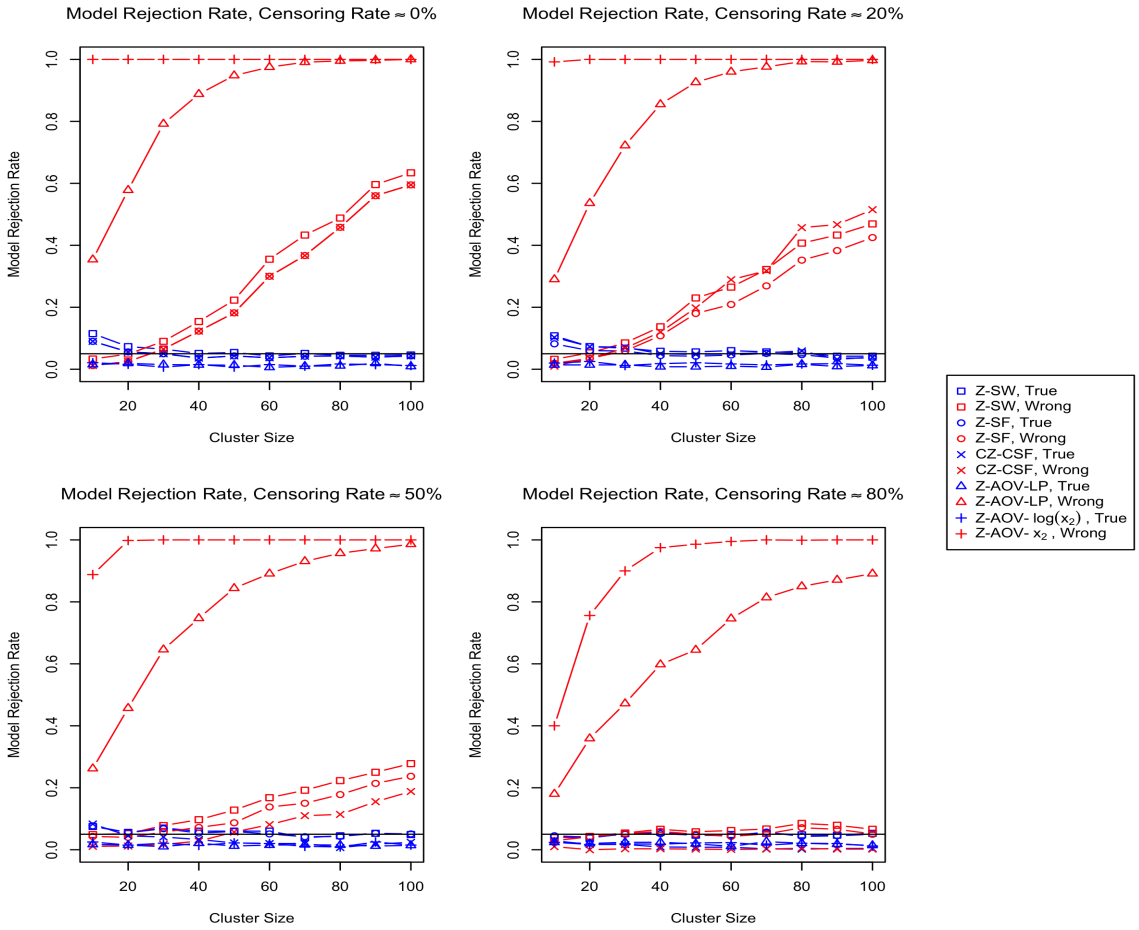

Figure A.4. Model rejection rates of various statistical tests based on Z-residuals for the simulation study of correlated covariates with 20 clusters, as presented in Section 5. A model is rejected when the test p-value is less than 5% (nominal level).

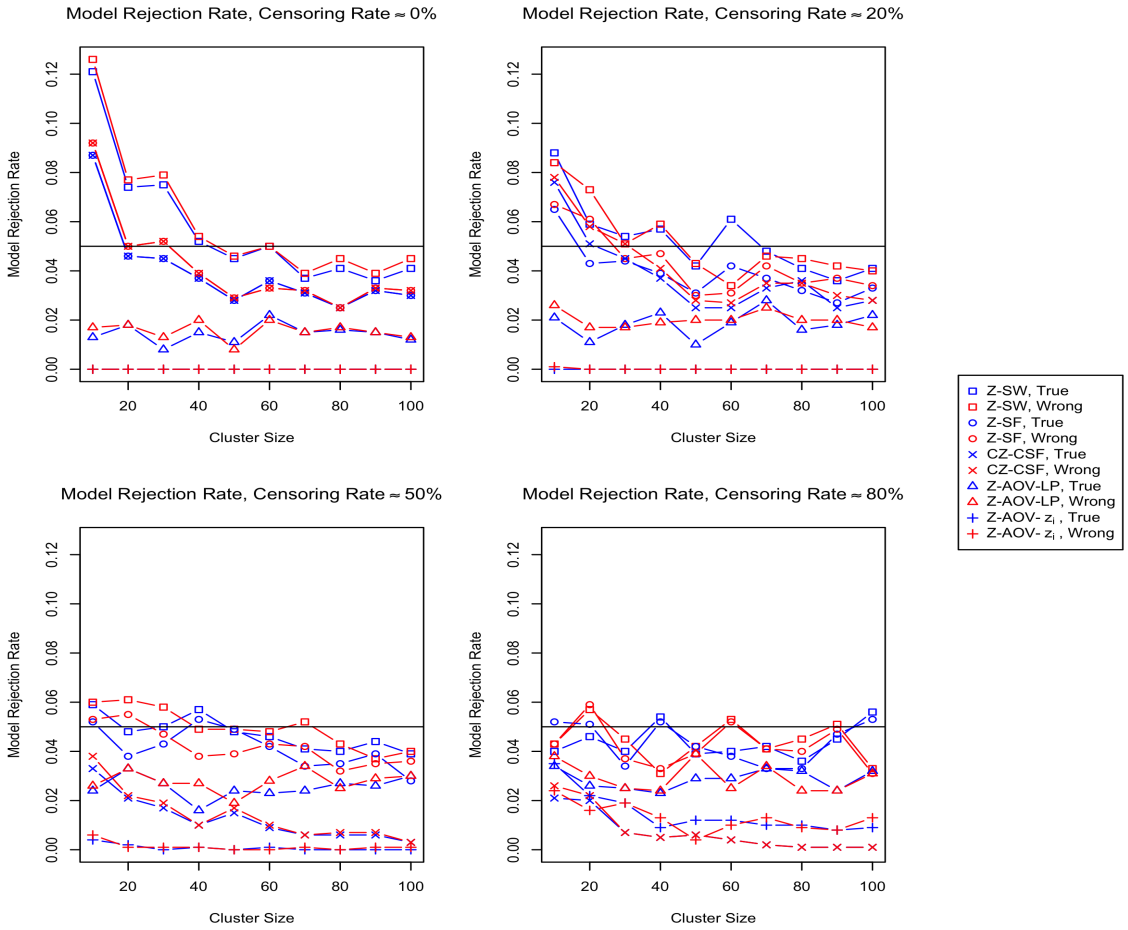

Figure A.5. Model rejection rates of various statistical tests based on Z-residual for the simulation study of frailty distribution misspecification with 20 clusters. A model is rejected when the test p-value is less than 5% (nominal level).

Table A1. Model rejection rates of various statistical tests based on Z-residual for the simulation study with 20 clusters, where c denotes the percentage of censoring and n denotes the total number of observations, which equals the number of clusters multiplied by the cluster size. A model is rejected when the test p-value is less than 5% (nominal level). The table corresponds to Figure 6.

Cluster Size	100c	n	Z-SW	Z-SW	Z-SF	Z-SF	CZ-CSF	CZ-CSF	Z-AOV-LP	Z-AOV-LP	Z-AOV-$\log(X_2)$	Z-AOV-X_2
			True	Wrong	True	Wrong	True	Wrong	True	Wrong	True	Wrong
10	0	200	0.105	0.015	0.077	0.010	0.077	0.010	0.017	0.694	0.019	0.999
20	0	400	0.074	0.080	0.056	0.064	0.056	0.064	0.015	0.966	0.014	1.000
30	0	600	0.064	0.198	0.046	0.166	0.046	0.166	0.018	0.997	0.024	1.000
40	0	800	0.054	0.344	0.045	0.313	0.045	0.313	0.012	1.000	0.017	1.000
50	0	1000	0.053	0.487	0.037	0.447	0.037	0.447	0.016	1.000	0.019	1.000
60	0	1200	0.051	0.639	0.032	0.618	0.032	0.618	0.016	1.000	0.018	1.000
70	0	1400	0.041	0.707	0.035	0.690	0.035	0.690	0.013	1.000	0.020	1.000
80	0	1600	0.038	0.779	0.038	0.758	0.038	0.758	0.013	1.000	0.016	1.000
90	0	1800	0.042	0.867	0.038	0.850	0.038	0.850	0.007	1.000	0.015	1.000
100	0	2000	0.041	0.928	0.033	0.917	0.033	0.917	0.017	1.000	0.024	1.000
10	20	200	0.128	0.028	0.110	0.014	0.120	0.015	0.025	0.631	0.017	0.975
20	20	400	0.088	0.076	0.069	0.050	0.078	0.057	0.012	0.932	0.015	1.000
30	20	600	0.065	0.156	0.047	0.131	0.048	0.142	0.014	0.986	0.020	1.000
40	20	800	0.058	0.211	0.050	0.195	0.053	0.229	0.018	0.996	0.019	1.000
50	20	1000	0.045	0.324	0.040	0.294	0.031	0.363	0.014	1.000	0.025	1.000
60	20	1200	0.043	0.377	0.031	0.369	0.037	0.442	0.012	1.000	0.013	1.000
70	20	1400	0.062	0.531	0.054	0.505	0.056	0.619	0.010	1.000	0.012	1.000
80	20	1600	0.060	0.597	0.055	0.583	0.049	0.691	0.023	1.000	0.014	1.000
90	20	1800	0.047	0.709	0.047	0.689	0.056	0.774	0.017	1.000	0.021	1.000
100	20	2000	0.044	0.747	0.036	0.735	0.044	0.817	0.014	1.000	0.030	1.000
10	50	200	0.069	0.037	0.067	0.036	0.079	0.013	0.028	0.500	0.023	0.784
20	50	400	0.078	0.044	0.060	0.038	0.056	0.009	0.016	0.828	0.014	0.986
30	50	600	0.070	0.062	0.058	0.054	0.030	0.024	0.010	0.919	0.031	1.000
40	50	800	0.056	0.089	0.050	0.073	0.029	0.044	0.014	0.975	0.026	1.000
50	50	1000	0.039	0.154	0.037	0.143	0.022	0.083	0.018	0.993	0.026	1.000
60	50	1200	0.058	0.168	0.057	0.156	0.025	0.127	0.010	0.997	0.017	1.000
70	50	1400	0.055	0.220	0.041	0.206	0.009	0.189	0.022	1.000	0.025	1.000
80	50	1600	0.053	0.253	0.050	0.237	0.019	0.230	0.019	0.999	0.022	1.000
90	50	1800	0.055	0.300	0.054	0.277	0.023	0.306	0.013	1.000	0.023	1.000
100	50	2000	0.050	0.327	0.050	0.319	0.014	0.340	0.019	0.999	0.021	1.000
10	80	200	0.054	0.038	0.053	0.043	0.031	0.014	0.039	0.291	0.015	0.268
20	80	400	0.059	0.040	0.063	0.045	0.018	0.007	0.024	0.577	0.030	0.568
30	80	600	0.052	0.042	0.050	0.057	0.018	0.009	0.021	0.713	0.017	0.723
40	80	800	0.045	0.076	0.046	0.077	0.004	0.009	0.021	0.825	0.030	0.866
50	80	1000	0.044	0.054	0.045	0.059	0.004	0.008	0.020	0.902	0.026	0.915
60	80	1200	0.057	0.054	0.057	0.055	0.009	0.009	0.017	0.942	0.038	0.953
70	80	1400	0.049	0.062	0.056	0.067	0.004	0.004	0.016	0.971	0.024	0.971
80	80	1600	0.049	0.070	0.052	0.080	0.004	0.011	0.020	0.977	0.028	0.981
90	80	1800	0.055	0.070	0.050	0.074	0.003	0.010	0.024	0.990	0.024	0.994
100	80	2000	0.049	0.069	0.048	0.078	0.000	0.012	0.021	0.995	0.024	0.996