Improved Confidence Intervals for Fixed Term Survival Probabilities in a Small Two-Arm Trial

Chen Qian
University of Louisville
Jianmin Pan
University of Louisville
Craig McClain
University of Louisville
Shesh Rai (shesh.rai@louisville.edu)
University of Louisville

Research Article

Keywords: Fixed-term confidence interval, Adjusted covariates, Kaplan-Meier, Cox regression

Posted Date: November 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-109278/v1

Improved Confidence Intervals for Fixed Term Survival Probabilities in a Small Two-Arm Trial

Chen Qian[a,b], Jianmin Pan[a], Craig J. McClain[c,d,e,f], Shesh N. Rai[a,b,e,f*]

[a] *Biostatistics and Bioinformatics Facility, James Graham Brown Cancer Center, University of Louisville, Louisville, Kentucky, 40202, USA*

[b] *Department of Biostatistics and Bioinformatics, University of Louisville, Louisville, Kentucky, 40202, USA*

[c] *Department of Medicine, University of Louisville, Louisville, Kentucky, 40202, USA*

[d] *Robley Rex Louisville VAMC, Louisville, Kentucky, 40206, USA*

[e] *University of Louisville Alcohol Research Center, University of Louisville, Louisville, Kentucky, 40202, USA*

[f] *University of Louisville Hepatobiology & Toxicology Center, University of Louisville, Louisville, Kentucky, 40202, USA*

[*] *Correspondence to Dr. Shesh N. Rai, Biostatistics and Bioinformatics Facility, James Graham Brown Cancer Center, University of Louisville, Louisville, Kentucky, 40202, USA*

[†] *E-mail: shesh.rai@louisville.edu*

Abstract

Background

The confidence interval for survival probability at a fixed time point provides valuable information on how the subject performs in terms of survival rate. However, in a two-arm trial when the sample size in each group is small or when the distribution of events that occurred within the group is skewed, the confidence interval might become very unstable, and thus may not provide accurate information for estimating survival rate. In addition, when there are other covariates available in the dataset, it is important to select those significant variables and include them in the model. On the other hand, researchers such as physicians who pay more attention to the final result often analyze the treatment group and control group separately, which may lead to inaccurate prediction.

Methods

In this study, two treatment groups are combined, and the group indicator variable is considered as a covariate and is included in the model for computation. Yuan and Rai's adjusted effective sample size methods are further extended along with Cox proportional hazard model, Weibull model, and log-logistic model to compute predicted fixed-term overall survival probabilities and corresponding confidence intervals with other covariates adjusted. Simulations are conducted to obtain coverage probability.

Results

In a single model, Wilson-Peto provides better confidence intervals than Kaplan-Meier, especially in the middle and later stages. In addition, AC-Peto produces better coverage probability at all time points. In a multivariate model, the log-logistic method provides both better confidence intervals and coverage probability than Cox regression model at all stages.

Conclusions

This paper provides a guideline on how one should correctly analyze survival data with the most appropriate method. Depending on the dataset, it is important to consider methods other than traditional Kaplan-Meier and Cox regression models when evaluating survival outcomes.

Keywords: Fixed-term confidence interval; Adjusted covariates; Kaplan-Meier; Cox regression.

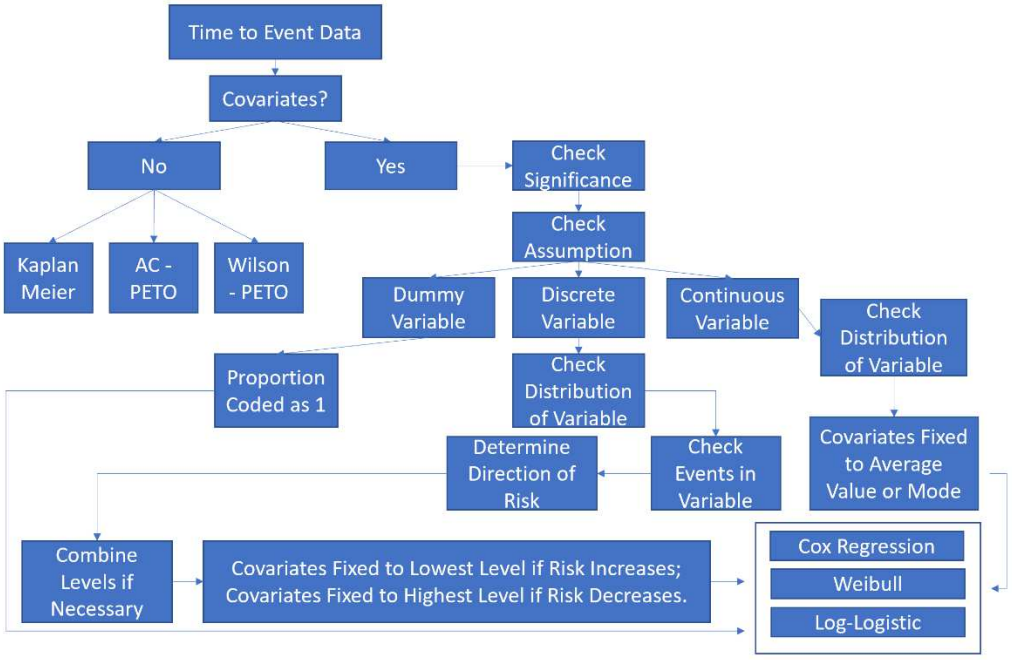
Background

When comparing multiple treatments, researchers often want to determine the effectiveness of a drug by looking at the p-value for the specific statistical test that has been conducted. Their bias is toward a small p-value so that they could make a successful conclusion on their study. However, as some articles have pointed out, in clinical research, solely looking at p-value is not enough to declare success. Alternatively, employing confidence intervals will provide solid evidence to support the conclusion [2]. The confidence interval can help researchers to decide whether one treatment is better than the other, which is extremely important for clinical decision making [3]. The calculation of confidence intervals has been widely adapted in comparing different treatments because it is more informative than a p-value [4]. The confidence interval is also highly important in survival analysis. Previous literature suggested the use of the confidence interval for a binomial proportion to calculate the interval for the survival function [5]. Multiple articles further examine how to estimate the best confidence interval with appropriate methods in order to better represent the survival data.

When there are several treatments involved in a study, researchers often look at the confidence intervals among each treatment separately. This allows researchers to make a prediction based on the observed range of the confidence interval of a certain treatment method. For example, when there are two treatment groups in a study, one could predict a certain patient's survival probability given the group to which that patient belongs. However, there will be a problem when the sample size is small or events are skewed. These conditions happen quite often in clinical researches due to the cost of trial and difficulty of patient follow up. Often, a trial will end with a small sample size. Thus, the estimated value which is derived from the data might not be the true value. Especially when the sample size is small, the true value might be quite different from the estimated value. The confidence interval provides a range that should contain the true value. But when the sample size is small, the standard error becomes large, and the confidence interval will become wider. As a result, the confidence interval will not produce a precise range to contain the true value. As Henderson and Keiding [6] noted, physicians face questions from patients for survival prediction after the diagnosis of a terminal disease, but they often cannot provide accurate survival time prediction since the variation is large among patients. Henderson further mentioned that the predictive interval could be a way to provide estimation since the actual survival might fall within the interval. Unfortunately, the predictive interval has rarely been used because the intervals are too wide. For example, a doctor can tell a patient's survival probability given a certain condition to be, for example, 50% in treatment A and 60% in treatment B. However, if the clinical trial information on which this prediction is based is from a trial with small sample sizes, that estimation is questionable.

In this paper, we are going to combine both two treatment groups and label the group indicator variable as a covariate. We include group covariate in all of our parametric and semi-parametric models. Our aim is to see if grouping has any impact on the model. We also want to see which method will provide the best predictive estimation with this improved confidence interval calculation method. Our overall aim is to provide a guideline on how basic survival data should be analyzed. Figure 1 below shows a schematic of the process.

Figure 1: Flowchart of the Data Analysis.



Data

The data used in this paper come from a randomized clinical trial conducted by the Radiation Therapy Oncology Group [7]. The dataset is publicly available, and therefore, neither ethical approval nor informed consent is needed for our study. The entire trial contains data from 15 sites with 16 participating institutions; however, in this paper, only the data on three sites with the six largest institutions will be used. At the beginning of this study, 193 patients were randomly assigned into two treatment groups. Group one (only radiation therapy) has 99 patients with 27 censored subjects. Group two (radiation therapy with a chemotherapeutic agent) has 94 patients with 26 censored subjects. Other variables including *sex*, *age*, *condition*, *T-staging*, and *N-staging*. Summary statistics can be found in Table 1.

Methods

Six methods are demonstrated in this paper, including Kaplan-Meier, Agresti-Coull with Peto's effective sample size adjusted method, Wilson with Peto's effective sample size adjusted method, Cox proportional hazard model, Weibull model, and log-logistic model. In each method, fixed-term overall survival probabilities and confidence intervals are calculated, then coverage probabilities for each method are obtained through simulation.

To have a good understanding of the data, *survdiff* in *R* is used to calculate whether or not there is any difference between the two treatment groups. It appears that according to the p-value (p=0.3), the two treatment groups are not significantly different. In this case, a closer look at the

confidence interval becomes necessary. Zhu et. al [8] evaluate several test procedures for survival functions comparison when data is interval-censored and the distribution of censoring is unequal. Similar approach could be tested for right-censored data in future research.

Kaplan-Meier

Kaplan-Meier curve can be easily produced with the help of *R*. Its confidence interval can also be obtained. Note that the default method for *R* in calculating confidence interval is the Greenwood (log) method, which can be treated as a Wald confidence interval, and has been proven not to be robust regardless the size of the sample [9]. The Kaplan-Meier estimate *S(t)* is

$$\hat{S}(t) = \prod_{j=1}^{k}(1 - \frac{d_j}{r_j}). \quad (1)$$

d_j and r_j are the number of deaths and the number of patients at risk at time t respectively. The R-code is as follows:

survdiff(Surv(Time,Status) ~ Group, data=oropharynx)
km=survfit(Surv(Time,Status)~Group,data= oropharynx)

Agresti-Coull-Peto

Brown et al. [9] recommend the Agresti-Coull interval when the sample size is greater than 40. It is a score interval and appears to be a better way to calculate the confidence interval. Moreover, Yuan and Rai further suggest that the combination of Agresti-Coull interval with Peto's adjusted effective sample size provides better coverage probability [1]. To construct the AC confidence interval, the formula can be written as:

$$\tilde{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + Z_{1-\alpha/2}}}, \quad (2)$$

where $\tilde{p} = \frac{M+{Z_{1-\alpha/2}}^2/2}{n+{Z_{1-\alpha/2}}^2}$ and $Z_{1-\alpha/2}$ is the critical value at 95% confidence level [10].

Here *M* is defined as the number of estimated events. Also, the sample size *n* needs to be adjusted by using Peto's effective sample size. *n* will be replaced by n_p [11]. n_p can be easily obtained by

$$n_p = \frac{r_t - d_t}{\hat{S}(t)}. \quad (3)$$

n_p is defined as the number of observations that remain at risk at time *t* divided by the survival probability at *t.* Replace *n* with n_p in equation 2 and it will generate the new Peto's adjusted confidence interval.

The *R* code is as follows:

```
kmout=summary(km)          #km from previous output
time=kmout$time            #death time
risk=kmout$n.risk          #number of risk
deathn=kmout$n.event       #number of events
surv=kmout$surv            #kaplan-meier estimate of survival function
sd=kmout$std.err           #standard error of each S(t) by greenwood formula
m=length(surv)
#calculate Estimated Effective Sample Size
np=(risk-deathn)/surv      #ESS from Peto
yp=np*surv
a=qnorm(1-0.05/2)
padjnp=(yp+0.5*a^2)/(np+a^2)      #adjusted p̂ for AC method using Peto's ESS
ci1=numeric(m)
ci2=numeric(m)
ci1=padjnp-a*sqrt( padjnp*(1-padjnp)/(np+a^2) )
ci2=padjnp+a*sqrt( padjnp*(1-padjnp)/(np+a^2) )
```

Wilson-Peto

Wilson method takes a similar approach. The variance is estimated differently. It proposes to provide a shorter confidence interval compare to AC method. It is calculated by using the following formula:

$$\tilde{p} \pm \frac{Z_{1-\alpha/2}}{1+Z^2_{1-\alpha/2}/n}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}+\frac{Z^2_{1-\alpha/2}}{4n^2}}. \quad (4)$$

Here, similarly, n will be replaced by n_p.

For Wilson-Peto's confidence interval, the *R* code is as follows:

```
ci1w<-padjnp-(a/(np+a^2))*sqrt(np*surv*(1-surv)+a^2/4)
ci2w<-padjnp+(a/(np+a^2))*sqrt(np*surv*(1-surv)+a^2/4)
```

Cox Regression Model

Since there are several covariates in the model, the above models might not best represent the data, and thus implementing other semi-parametric or parametric methods such as Cox proportional hazard model is necessary. The first step is to test whether those variables are significant. Univariate regression tests are used here, and only significant variables will be included in the multivariate regression. Next, it is important to make sure that the dataset does not violate the proportional hazard assumption. If the data does not violate the assumption, then the analysis will turn out to be good. If the data does violate the assumption, then the robust method should be used. There are three variables in the multivariate regression model including *T-staging*, *Condition*, and *Group* indicator. The purpose of this paper is to see whether making *Group* as a covariate will improve the predictive outcome. To look at the impact of *Group* on the model, the other covariates must be adjusted. Since *T-staging* and *Condition* are both categorical variables, it is essential to determine which level shall the covariates to be fixed at. Distributions

of the variables are checked to see whether they are evenly distributed. Next, the proportion of events under each level in each variable will be tested to see whether proportions are similar. Besides, based on the information of variable, the direction of risk can be determined. In other words, if a variable determines the size of tumor from level 1 to 5 in which 1 means the smallest, then one shall know that the direction of risk is increasing from 1 to 5. Note that if the variable is not evenly distributed and the proportion of event is also skewed, combining levels become necessary. Normally, if a level has the percentage of the number of subjects less than 10%, then combining levels will help. After combining levels, the covariate is fixed to the lowest level if risk increases, and to the highest level if risk decreases. *R* code as follow:

```
res.cox=coxph(Surv(Time,Status)~Group+factor(T)+factor(Condition),data= oropharynx)
res.zph=cox.zph(res.cox)                    #Check assumption
print(res.zph)
group_df <- with(oropharynx,
          data.frame(Group = c(1, 2)
                     T = c(3, 3),
                     Condition= c(1,1)
                    )
                )
fit=survfit(res.cox,newdata=group_df)
surv_summary(fit)
```

Predicted survival probability and confidence interval can be easily obtained from the summary output.

Weibull Model

Other than Cox regression, people usually look at parametric methods such as Weibull and log-logistic. Those methods are more often recommended when failure time distribution cannot be determined by a particular distribution family. Weibull distribution can model multiple types of data, regardless of the presence of skew or not. In addition, Weibull distribution can easily model hazard function for survival analysis. In *R*, *survreg* can be used and distribution can be specified as *weibull* to get scale and intercept values. To start, the cumulative distribution function of Weibull is

$$F(x) = P(X \leq x) = 1 - e^{-(1/\alpha)x^{\beta}}. \tag{5}$$

The survival probability at a fixed time point can be carried out as

$$\hat{S}(x) = P(X \geq x) = e^{-\left(\frac{1}{\alpha}\right)x^{\beta}}, \tag{6}$$

where α is the scale parameter and β is the shape parameter. The confidence interval can also be easily calculated. In *R* there is a package called *flexsurv*. Within it, the *flexsurvreg* command can easily model Weibull. The same approach that was used with Cox regression to fix variables at certain levels can also be used in Weibull model. The *R* code is as follows:

```
b=flexsurvreg(Surv(Time,Status)~factor(T)+Group+factor(Condition),data=oropharynx,dist="
weibull")
group_df <- with(oropharynx,
        data.frame(Group = c(1, 2),
                T = c(3, 3),
                Condition= c(1,1)
                )
            )
group_df
fit=summary(b,newdata=group_df,dist="weibull")
```

Log-Logistic Model

Alakus [12] developed a method to calculate confidence intervals for the log-logistic distribution survival function at any time point. To calculate survival probability, *survreg* can be used and the distribution can be specified as log-logistic to obtain scale and intercept values. The cumulative distribution function of Log-logistic is

$$F(x) = P(X \leq x) = \frac{(\lambda x)^{\theta}}{1 + (\lambda x)^{\theta}}. \quad (6)$$

The survival probability at a fixed time point can be carried out as

$$\hat{S}(x) = P(X \geq x) = \frac{1}{1 + (\lambda x)^{\theta}}, \quad (7)$$

where λ is the scale parameter and θ is the shape parameter. To calculate confidence intervals, the interval for score function needs to be formed first, and then it can be extended to the survival function. The score function can be defined as $R_i = log\theta + \theta log\lambda$. The estimated standard error can be defined as $se(\widehat{R_i}) = \{y_i^T Var(\theta) y_i\}^{1/2}$ where $y_i^T = [1\ 1]$ is the unit vector for the model and $Var(\theta)$ is the variance-covariance matrix that can be obtained from the program using *vcov* command. The interval can be shown as $R_i \pm Z_{1-\alpha/2} * se(R_i)$.

Therefore, the confidence interval for the survival function can be written as

$$CI_{low}: \theta[\theta + e^{R_{upp}} x_i^{\theta}]^{-1}, \quad (8)$$

$$CI_{upp}: \theta[\theta + e^{R_{low}} x_i^{\theta}]^{-1}. \quad (9)$$

Similarly, *flexsurvreg* can be used to obtain survival probability and confidence interval. Same approach that were used in Cox regression and Weibull to fix variable at certain level can also be used in log-logistic model. *R* code as follow:

```
c=flexsurvreg(Surv(Time, Status)~ factor(T)+Group+factor(Condition),data=
oropharynx,dist="llogis")
group_df <- with(oropharynx,
```

```
        data.frame(Group = c(1, 2),
                   T = c(3, 3),
                   Condition= c(1,1)
                   )
                 )
group_df
fit1=summary(c,newdata=group_df,dist="llogis")
```

For each of the above models, survival probability, as well as the confidence interval, are obtained at fixed time points such as 3-months, 6-months, 12-months, and 18-months. In addition, confidence interval length can be calculated by taking the difference between upper and lower bounds. Bootstrap sampling with 5000 iterations is used to obtain survival probability, confidence interval, and coverage probability.

Results

Of the 193 patients, approximately 27% are censored (Table 1). The proportion of events within each group is relatively similar. For all variables in the dataset, none has an association with treatment group, as p-values are all greater than 0.05. Methods without considering covariates were tested first. Note that the AC-Peto method and the Wilson-Peto method carry the same survival probability and are only different in confidence interval. By comparing both methods with the usual Kaplan-Meier method, it appears that Kaplan-Meier produces higher survival probabilities at all time points in both treatment groups (Table 2). In group 1, Kaplan-Meier produces a relatively shorter confidence interval than the other two methods at an early stage such as 3-months (Table 3). At 6-months, Wilson produces the shortest interval among all three methods. Toward the later stages, both the AC-Peto and the Wilson-Peto methods produce better intervals than Kaplan-Meier, and Wilson-Peto is relatively better than AC-Peto. In the later stage, AC-Peto and Wilson-Peto improve the confidence interval by about 2% from Kaplan-Meier. Similarly, coverage probabilities also appear to be improved by a small percent for AC-Peto and Wilson-Peto (Table 4). In this case, AC-Peto has the highest coverage among all methods. Group 2 seems to carry a similar result to group 1. Kaplan-Meier has the shortest confidence interval at the early stage, and Wilson-Peto makes the greatest improvement at a later stage (Table 3). In terms of coverage probability, AC-Peto has better coverage than Kaplan-Meier but is close to Wilson-Peto in the early stage, whereas, and in the later time, AC-Peto has the best coverage among all methods (Table 4).

To better predict survival outcome, significant covariates must be taken into consideration. Univariate Cox regression analysis showed that only variable *Condition* ($p<0.01$) and *T-Staging* ($p=0.01$) are significant (Table 5). Therefore, variables that will be included in the multivariate regression are limited to *Condition*, *T-Staging*, and *Group*. Proportional hazards assumption analysis showed that the model does not violate the assumption (Global $p=0.052$) which indicates that the Cox regression model would be a good fit for analyzing those variables (Table 6). Since *Condition* and *T-Staging* are categorical variables, it is necessary to take a closer look at their distribution. Under *T-Staging*, the number of subjects at level 1 is only 4.6% of overall subjects. Typically, combing levels become necessary when a certain group falls below 10% of the total. Before combining levels, it is extremely important to check the distribution of events

among all levels. In this case, the ratios between the number of events and number of subjects are relatively close among all levels (Table 7). Therefore, combining level 1 and level 2 becomes reasonable. In the variable *Condition*, the distribution of subjects is skewed wherein level 1 has around 73% of the total, but level 3 and level 4 have only 3% and 0.5%, respectively (Table 7). Taking a closer look at the distribution of events, due to the small sample size of level 4, the proportion will be either 100% or 0% in this case. This situation does not provide much valuable information, and based on the definition of this variable, people with a higher level of the condition tend to have a higher risk. It is therefore reasonable to see that the proportion increases from level 1 to level 3. In this study, level 3 and level 4 are combined for computation. In group 1, log-logistic has the highest survival probabilities at all time points (Table 8). Weibull has relatively higher survival probabilities than Cox in the later stage but *vice versa* in the early stage. In terms of confidence intervals, at 3-months, the log-logistic interval is around 15% shorter than Cox, and about 21% shorter at 6-months (Table 9). In the later stage, Weibull has roughly 20% shorter intervals compared to Cox intervals. On the other hand, Weibull has the least coverage probabilities at all tested time points (Table 11). Log-logistic has slightly better coverage than Cox at 3-months, 6-months, and 12-months, but not at 18-months. In group 2, Weibull and log-logistic have higher survival probabilities at most time points (Table 8). Confidence intervals follow the same pattern as group 1 (Table 9). In terms of coverage probability, log-logistic has the best improvement at all time points (Table 11). Weibull has better coverage than Cox at earlier stages but becomes worse at later stages.

Further comparisons were made between semi/parametric models and AC/Wilson-Peto methods. As seen in Table 10, in both groups semi-parametric and parametric models produce shorter confidence intervals than AC/Wilson-Peto methods in earlier stages, but in the long term, basic models tend to perform better. In terms of coverage, semi-parametric and parametric models produce better coverage than basic models only at 3-months in group 1 (Table 12). Survival curves for Kaplan-Meier, AC-Peto, and Cox regression can be found in Figure 2. All methods compared to Kaplan-Meier can be found in Figure 3 and Figure 4. In most cases, semi-parametric and parametric models produce shorter confidence intervals in the early stage, but the pattern does not hold for later stages. Similarly, coverage is higher at early stages for semi-parametric and parametric methods but becomes worse in the long term.

Discussion

This paper illustrates the group effect with other covariates adjusted in survival calculation. This method can also be expanded to three or more groups. It can further be expanded to determine if making group as a covariate will benefit large sample size as well. The method provides more important and more accurate information to researchers as well as clinicians when making a survival prediction. Note that when distributions of subjects and events are skewed in certain variables, it is important to determine the best way to combine levels within the variable. There are many ways to combine levels, such as making it into two blocks. The best way always depends on the dataset.

In this paper, the predictive survival probability is used rather than directly obtaining the result from data analysis. The reason for doing so is that for a fixed term estimation, it is more

accurate. For example, one event occurred at 9-months, and the next event occurred at 13-months, if we want to find fixed term survival rate at 12-months (1-year), we know that it is the same between 9-month and 13-month because it is a stepwise function. But the interval, in this case, is very wide. Suppose there is no event at 12-months, then the survival function is zero, but the size of the risk set is not. In order to make sense of the data, we need to calculate the estimated event at 12-months, and then find its corresponding survival probability and confidence interval.

In this paper, only 3-month, 6-month, 12-month, and 18-month survival are tested. The reason for doing so is that 1-year survival is a normal clinical indicator for many terminal illnesses, thus, it is the most important term we want to look at. Dong et al. [13] recommend use generalized inference approach to calculate confidence interval when normality is satisfied, and use bootstrap percentile approach when assumption being violated as well as when the probability of detecting early disease stage is large and sample sizes are small. That approach should be considered in future research.

Based on the result above, in a single model, Wilson-Peto provides better confidence interval than Kaplan-Meier at 12-months and later. In addition, AC-Peto produces better coverage probability at all time points. In a multivariate model, the log-logistic method provides both better confidence intervals and coverage probability than Cox regression model at all stages. Comparing all methods together, long term confidence intervals all become very wide and lose coverage, and therefore, the confidence interval is better when n is small.

Fay et. al. proposed a non-parametric method called beta product confidence procedure for right-censored data with independent censoring that should provide better coverage than the basic Kaplan-Meier method [14]. Further research could compare this method with methods that are discussed in this paper. Lee et. al also mentioned that the average covariate method might have some disadvantage, and instead, the corrected group prognostic curve approach was recommended [15]. This could also be tested in the future.

In summary, we have examined six methods for predicting overall survival probabilities and confidence intervals. Coverage probabilities for each method are obtained through simulation. In this paper, we combined both two treatment groups and labeled the group indicator variable as a covariate. We included group covariate in all of our parametric and semi-parametric models. Our aim was to see if grouping has any impact on the model. We also wanted to see which method will provide the best predictive estimation with this improved confidence interval calculation method. Our overall aim is to provide a guideline on how basic survival data should be analyzed.

Conclusion

This paper provides detailed guidance on how to deal with survival data. Predicted overall survival probability and confidence intervals are calculated by making the group a covariate and looking at group effect with other covariates adjusted. It also compares different methods of whether or not covariate is present. In a single model, it proves that Wilson-Peto provides better confidence intervals than Kaplan-Meier, especially in the middle and later stages. Moreover, AC-Peto produces better coverage probability at all time points. In a multivariate model, the log-

logistic method provides both better confidence intervals and coverage probability than Cox regression model at all stages.

List of Abbreviations

AC: Agresti-Coull

Declarations

Ethics approval and consent to participate

The dataset is publicly available, and therefore, neither ethical approval nor informed consent is needed for this study.

Consent to publish

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

C. Qian was supported by the National Institute of Health grant 5P50 AA024337 to principal investigator Dr. Craig J. McClain and the University of Louisville Fellowship. S. N. Rai was partly supported with Wendell Cherry Chair in Clinical Trial Research Fund.

Author's contributions

Conception: SNR, CQ, JP

Data analysis: CQ

Original Draft: SNR, CQ, JP

Critical Input: CJM

Review & Editing: All Authors

Manuscript Revision: All Authors

Approval of the Final Version: All Authors

Acknowledgements

Not applicable

References

[1] Yuan X, Rai SN. Confidence Intervals for Survival Probabilities: A Comparison Study. *Communications in Statistics - Simulation and Computation.* 2011;40(7):978–991.

[2] Jiroutek MR, Turner JR. In Praise of Confidence Intervals: Much More Informative than P Values Alone. *The Journal of Clinical Hypertension.* 2016;18(9):955–957.

[3] Pandis N. Confidence Intervals rather than P Values. *American Journal of Orthodontics & Dentofacial Orthopedics*. 2013;143(2):293–294.

[4] du Prel JB, et al. Confidence Interval or P-Value? *Dtsch Arztebl Int.* 2009;106(19):335–339.

[5] Rothman KJ. Estimation of Confidence Limits for the Cumulative Probability of Survival in Life Table Analysis. *Journal of Chronic Diseases*. 1978;31:557–560.

[6] Henderson R, Keidling N. Individual survival time prediction using statistical models. *Journal of Medical Ethics.* 2005;31:703-706.

[7] Kalbfleisch JD, Prentice RL. Data Set II: Clinical Trial in the Treatment of Carcinoma of the Oropharynx. In: *The Statistical Analysis of Failure Time Data Second Edition*; Hoboken, New Jersey. John Wiley & Sons, Inc.; 2002. p. 379–384; Wiley Series in Probability and Statistics.

[8] Zhu C, et al. A Nonparametric Test for Interval-Censored Failure Time Data with Unequal Censoring. *Communications in Statistics: Theory and Methods.* 2008; 37:1895-1904.

[9] Brown LD, et al. Interval Estimation for a Binomial Proportion (with discussion). *Statistical Science.* 2001;16:101–133.

[10] Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician.* 1998;52:119–126.

[11] Peto R, et al. Design and Analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer.* 1977; 35:1–39.

[12] Alakus K, Erilli NA. Confidence Intervals Estimation for Survival Function in Log-Logistic Distribution and Proportional Odds Regression Based on Censored Survival Time Data. *Biometrics & Biostatistics.* 2011;2(3):116.

[13] Dong T, et al. Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups. *Statistics in Medicine*. 2011; 30(30):3532-3545.

[14] Fay MP, et al. Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics.* 2013;14(4):723-736.

[15] Lee J, et al. Covariance adjustments of survival curves based on Cox's proportional hazards regression model. *Computer Applications in the Biosciences.* 1992;8(1):23-27.

Appendix

Table 1. Summary Statistics for Oropharynx Data.

Variables	Total (N=193)	Treatment 1 (N=99)	2 (N=94)	P Value
Age				0.72
Mean ± SE	60.19 ± 0.79	60.47 ± 1.18	59.89 ± 1.05	
Sex				0.98
Male (%)	148 (76.7)	76 (76.8)	72 (76.6)	
Female (%)	45 (23.3)	23 (23.2)	22 (23.4)	
Condition				0.09†
1 (%)	143 (74.1)	79 (79.8)	64 (68.1)	
2 (%)	43 (22.3)	19 (19.2)	24 (25.5)	
3 (%)	6 (3.1)	1 (1.0)	5 (5.3)	
4 (%)	1 (0.5)	0 (0.0)	1 (1.1)	
T-Staging				0.72†
1 (%)	9 (4.7)	3 (3.0)	6 (6.4)	
2 (%)	26 (13.5)	14 (14.1)	12 (12.8)	
3 (%)	92 (47.7)	49 (49.5)	43 (45.7)	
4 (%)	66 (34.2)	33 (33.3)	33 (35.1)	
N-Staging				0.92
0 (%)	38 (19.7)	21 (21.2)	17 (18.1)	
1 (%)	28 (14.5)	15 (15.2)	13 (13.8)	
2 (%)	37 (19.2)	19 (19.2)	18 (19.1)	
3 (%)	90 (46.6)	44 (44.4)	46 (48.9)	
Status				0.95
Censored (%)	53 (27.5)	27 (27.3)	26 (27.7)	
Dead (%)	140 (72.5)	72 (72.7)	68 (72.3)	

†Exact test

Table 2. Survival Probability and Confidence Interval at Fixed Time Points.

Group 1		3 Months	6 Months	12 Months	18 Months
	S(t)	0.979	0.847	0.632	0.477
Kaplan-Meier	CI Lower	0.955	0.779	0.544	0.388
	CI Upper	0.999	0.921	0.734	0.587
	S(t)	0.964	0.834	0.627	0.478
AC – Peto	CI Lower	0.931	0.763	0.533	0.381
	CI Upper	0.997	0.906	0.720	0.575
	S(t)	0.964	0.834	0.627	0.478
Wilson – Peto	CI Lower	0.935	0.764	0.533	0.381
	CI Upper	0.993	0.905	0.719	0.575
Group 2					
	S(t)	0.947	0.819	0.524	0.357
Kaplan-Meier	CI Lower	0.904	0.745	0.432	0.272
	CI Upper	0.990	0.899	0.636	0.470
	S(t)	0.929	0.807	0.523	0.363
AC - Peto	CI Lower	0.880	0.729	0.423	0.265
	CI Upper	0.979	0.884	0.623	0.462
	S(t)	0.929	0.807	0.523	0.363
Wilson - Peto	CI Lower	0.883	0.730	0.423	0.265
	CI Upper	0.976	0.883	0.623	0.460

Table 3. Confidence Interval Length Comparison.

Group 1	3 Months	6 Months	12 Months	18 Months
Kaplan-Meier	0.044	0.141	0.191	0.199
AC - Peto	0.066	0.143	0.187	0.195
Wilson - Peto	0.058	0.141	0.187	0.195
AC vs KM (%)	49.19	1.57	-1.95	-2.02
Wilson vs KM (%)	31.85	-0.15	-2.11	-2.04
Group 2				
Kaplan-Meier	0.086	0.154	0.204	0.199
AC - Peto	0.099	0.155	0.199	0.197
Wilson - Peto	0.093	0.153	0.199	0.197
AC vs KM (%)	15.50	0.58	-1.99	-0.75
Wilson vs KM (%)	8.17	-0.74	-2.02	-0.96

Table 4. Coverage Probability Comparison.

Group 1	3 Months	6 Months	12 Months	18 Months
Kaplan-Meier (%)	87.02	93.12	94.66	94.68
AC - Peto (%)	87.02	94.94	94.96	95.34
Wilson - Peto (%)	87.02	93.80	94.96	95.30
AC vs KM (%)	0	1.95	0.32	0.70
Wilson vs KM (%)	0	0.73	0.32	0.65
Group 2				
Kaplan-Meier (%)	87.62	93.10	95.16	95.02
AC - Peto (%)	95.96	95.12	95.48	95.70
Wilson - Peto (%)	95.96	95.12	95.44	95.60
AC vs KM (%)	9.52	2.17	0.33	0.72
Wilson vs KM (%)	9.52	2.17	0.29	0.61

Table 5. Univariate Cox Regression Test.

Covariates	Beta	HR (95% CI)	Wald Test	P Value
Sex	0.17	1.2 (0.8-1.8)	0.72	0.39
Age	0.0042	1 (0.99-1)	0.27	0.60
Condition	0.9	2.4 (1.9-3.2)	40	<0.01
Site	-0.04	0.96 (0.84-1.1)	0.34	0.56
T-Staging	0.31	1.4 (1.1-1.7)	7.1	0.01
N-Staging	0.13	1.1 (0.98-1.3)	3.1	0.08

Table 6. Test the Proportional Hazards Assumption of a Cox Regression.

Covariates	Rho	Chi-Square	P-Value
Group	-0.0338	0.164	0.69
T-Staging at Level 2	0.1197	2	0.16
T-Staging at Level 3	0.1132	1.82	0.18
T-Staging at Level 4	0.0571	0.467	0.49
Condition at Level 2	-0.2454	8.13	<0.01
Condition at Level 3	-0.0613	0.526	0.47
Condition at Level 4	0.2492	1.42e-07	0.99
Global	NA	1.4	0.052

Table 7. Summary Statistics for Variable *T*-Staging and *Condition*.

T-Staging	1	2	3	4
Number of Subjects	9	26	92	66
Number of Events	6	16	64	54
Proportion	67%	62%	70%	82%

Condition	1	2	3	4
Number of Subjects	143	43	6	1
Number of Events	96	38	6	0
Proportion	67%	88%	100%	0%

Table 8. Survival Probability and Confidence Interval at Fixed Time Points.

Group 1		3 Months	6 Months	12 Months	18 Months
Cox Regression	S(t)	0.981	0.903	0.715	0.568
	CI Lower	0.965	0.853	0.606	0.436
	CI Upper	0.997	0.957	0.845	0.743
	S(t)	0.954	0.894	0.773	0.652
Weibull	CI Lower	0.922	0.833	0.666	0.512
	CI Upper	0.975	0.937	0.853	0.764
	S(t)	0.984	0.943	0.821	0.678
Log-Logistic	CI Lower	0.966	0.891	0.704	0.524
	CI Upper	0.994	0.973	0.902	0.803
Group 2					
Cox Regression	S(t)	0.980	0.897	0.699	0.546
	CI Lower	0.962	0.844	0.587	0.412
	CI Upper	0.997	0.953	0.832	0.725
	S(t)	0.952	0.889	0.763	0.637
Weibull	CI Lower	0.918	0.825	0.651	0.493
	CI Upper	0.974	0.934	0.847	0.754
	S(t)	0.980	0.929	0.784	0.624
Log-Logistic	CI Lower	0.959	0.869	0.656	0.467
	CI Upper	0.992	0.965	0.876	0.759

Table 9. Confidence Interval Length Comparison.

Group 1	3 Months	6 Months	12 Months	18 Months
Cox Regression	0.032	0.104	0.239	0.307
Weibull	0.053	0.104	0.188	0.252
Log-Logistic	0.028	0.082	0.198	0.279
Weibull vs Cox (%)	62.76	-0.19	-21.37	-18.01
Log-logistic vs Cox (%)	-14.83	-21.00	-17.05	-9.00
Group 2				
Cox Regression	0.034	0.109	0.245	0.313
Weibull	0.055	0.109	0.196	0.261
Log-Logistic	0.033	0.096	0.219	0.292
Weibull vs Cox (%)	60.31	0.13	-19.95	-16.54
Log-logistic vs Cox (%)	-5.39	-12.04	-10.44	-6.66

Table 10. Confidence Interval Length Comparison with AC-Peto and Wilson-Peto (Change in Percentage).

Group 1	3 Months	6 Months	12 Months	18 Months
Cox vs Wilson	-44.59	-25.92	27.83	57.51
Cox vs AC	-51.03	-27.17	27.64	57.47
Weibull vs Wilson	-9.81	-26.05	0.52	29.13
Weibull vs AC	-20.29	-27.31	0.36	29.10
Log-logistic vs Wilson	-52.81	-41.48	6.05	43.32
Log-logistic vs AC	-58.29	-42.47	5.88	43.29
Group 2				
Cox vs Wilson	-63.10	-28.82	22.67	59.14
Cox vs AC	-65.44	-29.76	22.64	58.81
Weibull vs Wilson	-40.84	-28.73	-1.80	32.81
Weibull vs AC	-44.60	-29.66	-1.83	32.53
Log-logistic vs Wilson	-65.09	-37.40	9.86	48.53
Log-logistic vs AC	-67.31	-38.22	9.83	48.22

Table 11. Coverage Probability Comparison.

Group 1	3 Months	6 Months	12 Months	18 Months
Cox (%)	91.14	92.70	93.80	95.16
Weibull (%)	90.28	92.62	92.92	93.30
Log-Logistic (%)	92.06	94.30	94.22	94.08
Weibull vs Cox (%)	-0.94	-0.09	-0.94	-1.95
Log-logistic vs Cox (%)	1.01	1.73	0.45	-1.13
Group 2				
Cox Regression (%)	91.04	93.64	94.58	95.12
Weibull (%)	92.12	94.02	94.14	93.96
Log-Logistic (%)	93.94	96.38	95.88	95.74
Weibull vs Cox (%)	1.19	0.41	-0.47	-1.22
Log-logistic vs Cox (%)	3.19	2.93	1.37	0.65

Table 12. Coverage Probability Comparison with AC-Peto and Wilson-Peto (Change in Percentage).

Group 1	3 Months	6 Months	12 Months	18 Months
Cox vs Wilson	4.73	-1.17	-1.22	-0.15
Cox vs AC	4.73	-2.36	-1.22	-0.19
Weibull vs Wilson	3.75	-1.26	-2.15	-2.09
Weibull vs AC	3.75	-2.44	-2.15	-2.14
Log-logistic vs Wilson	5.79	0.53	-0.78	-1.28
Log-logistic vs AC	5.79	-0.67	-0.78	-1.32
Group 2				
Cox vs Wilson	-5.13	-1.56	-0.90	-0.50
Cox vs AC	-5.13	-1.56	-0.94	-0.61
Weibull vs Wilson	-4.00	-1.16	-1.36	-1.72
Weibull vs AC	-4.00	-1.16	-1.40	-1.81
Log-logistic vs Wilson	-2.11	1.32	0.46	0.15
Log-logistic vs AC	-2.11	1.32	0.42	0.04

Figure 2: Survival curve comparison between Kaplan-Meier, AC-Peto, and Cox Regression.

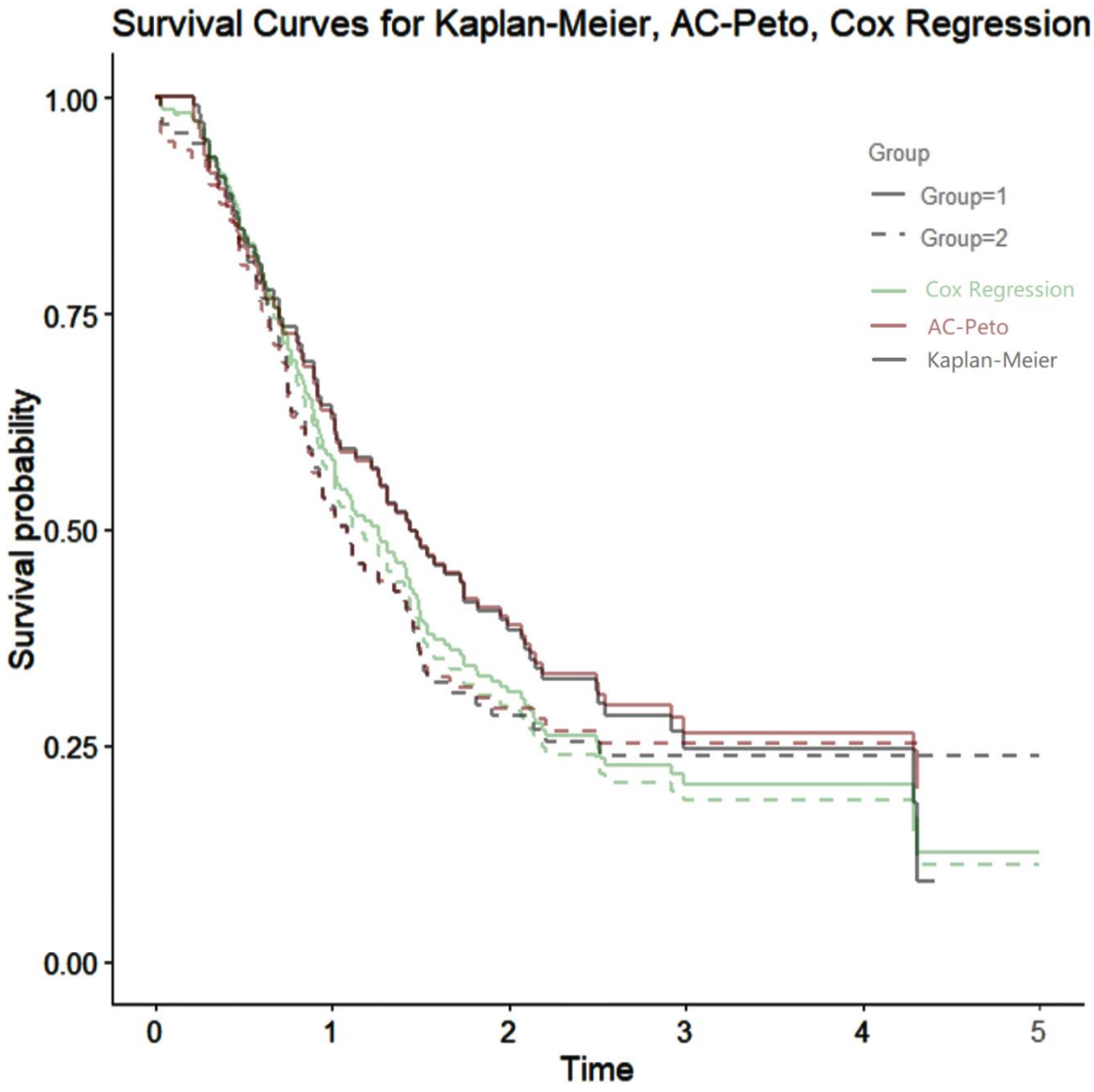


Figure 3: Comparison of confidence interval length changes to the Kaplan-Meier method.

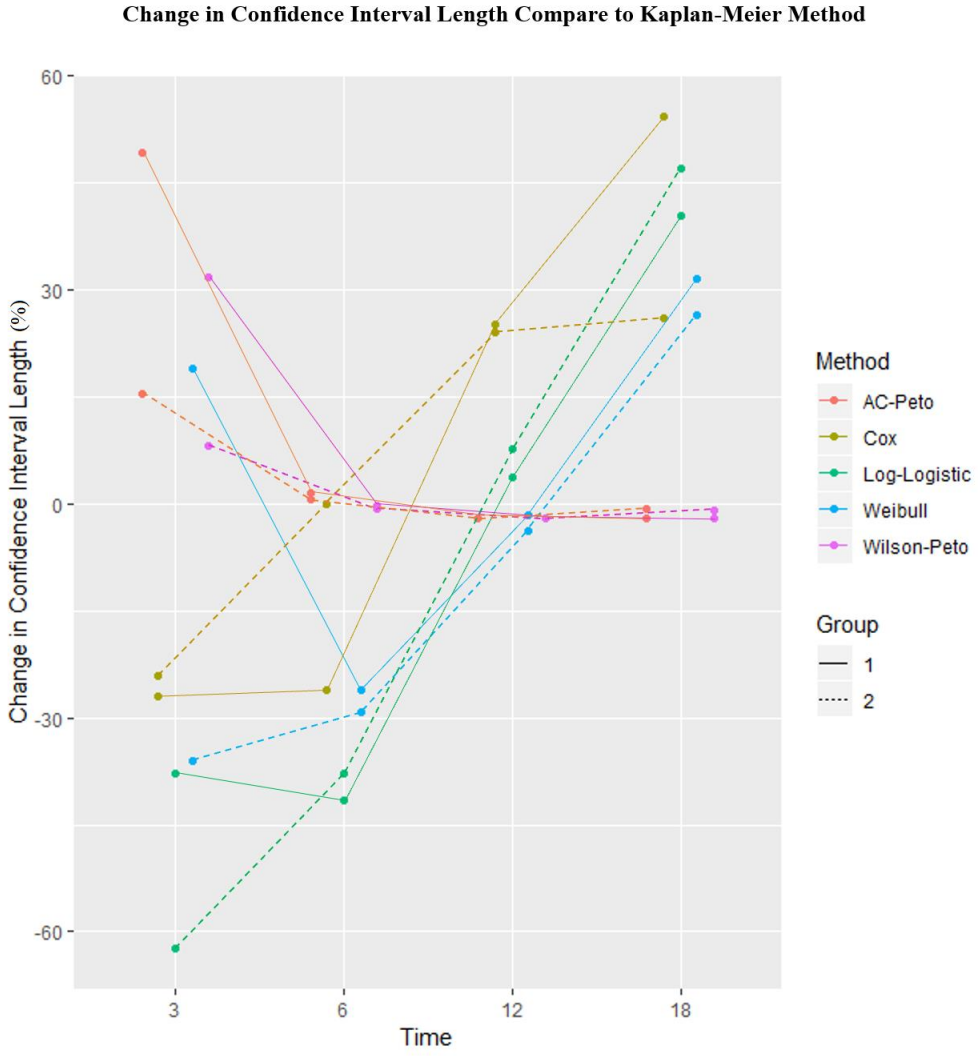
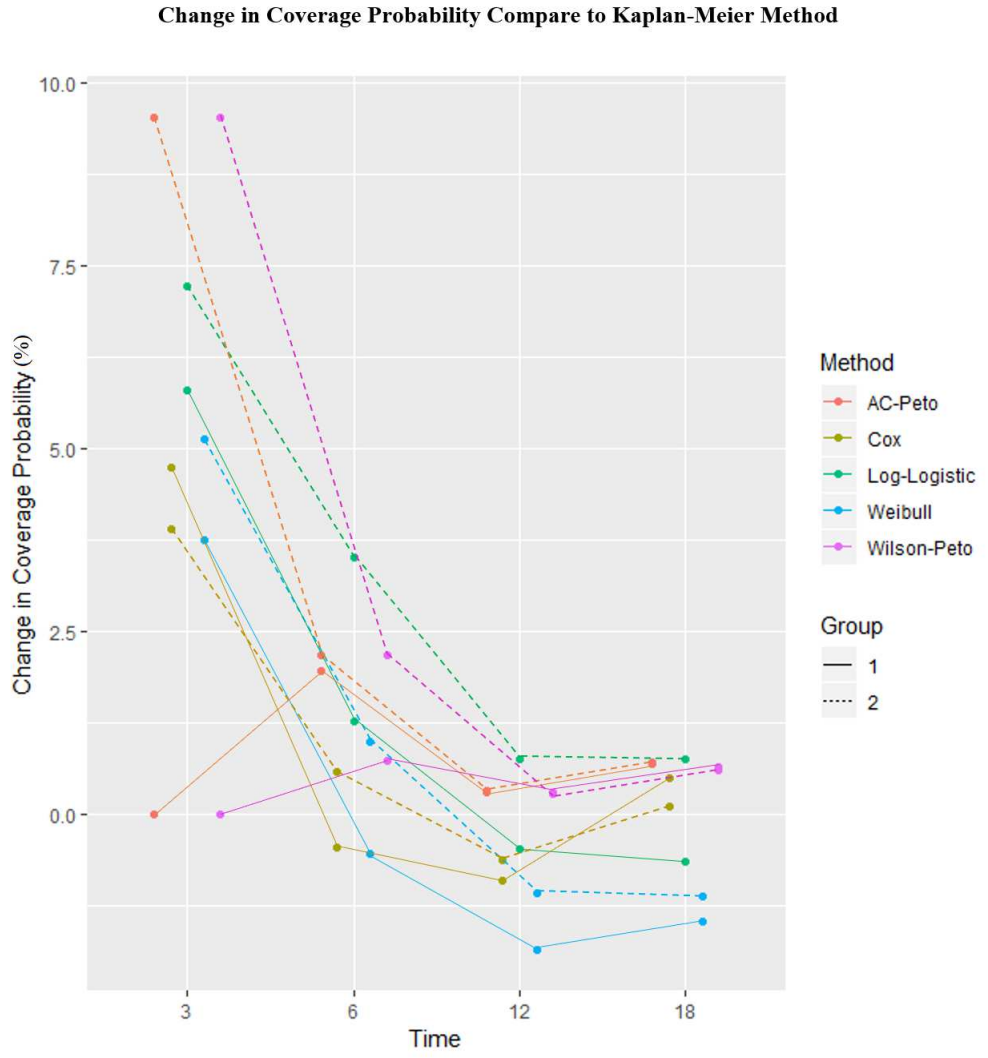


Figure 4: Comparison of coverage probability changes to the Kaplan-Meier method.



Figures

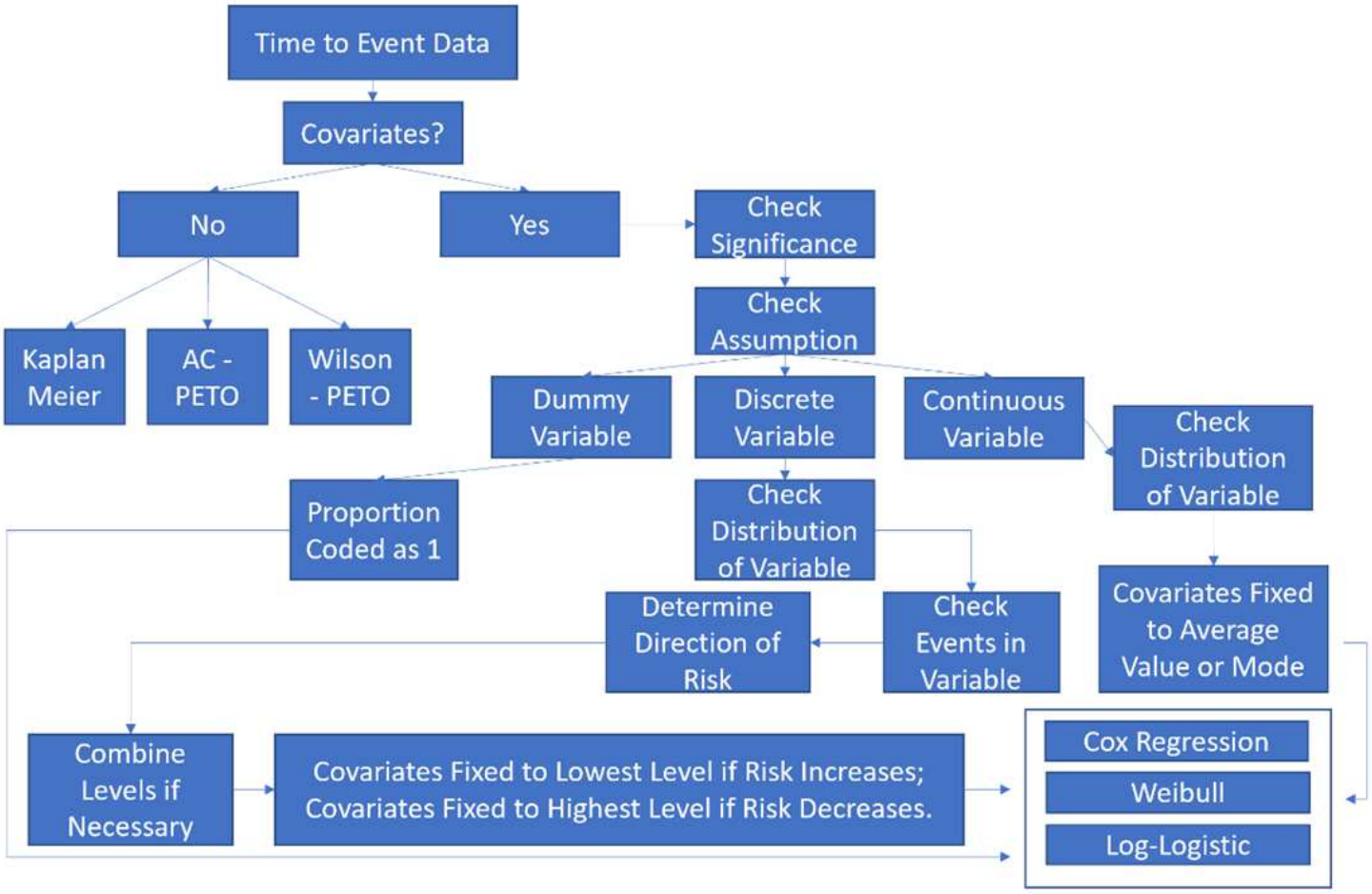


Figure 1

Flowchart of the Data Analysis.

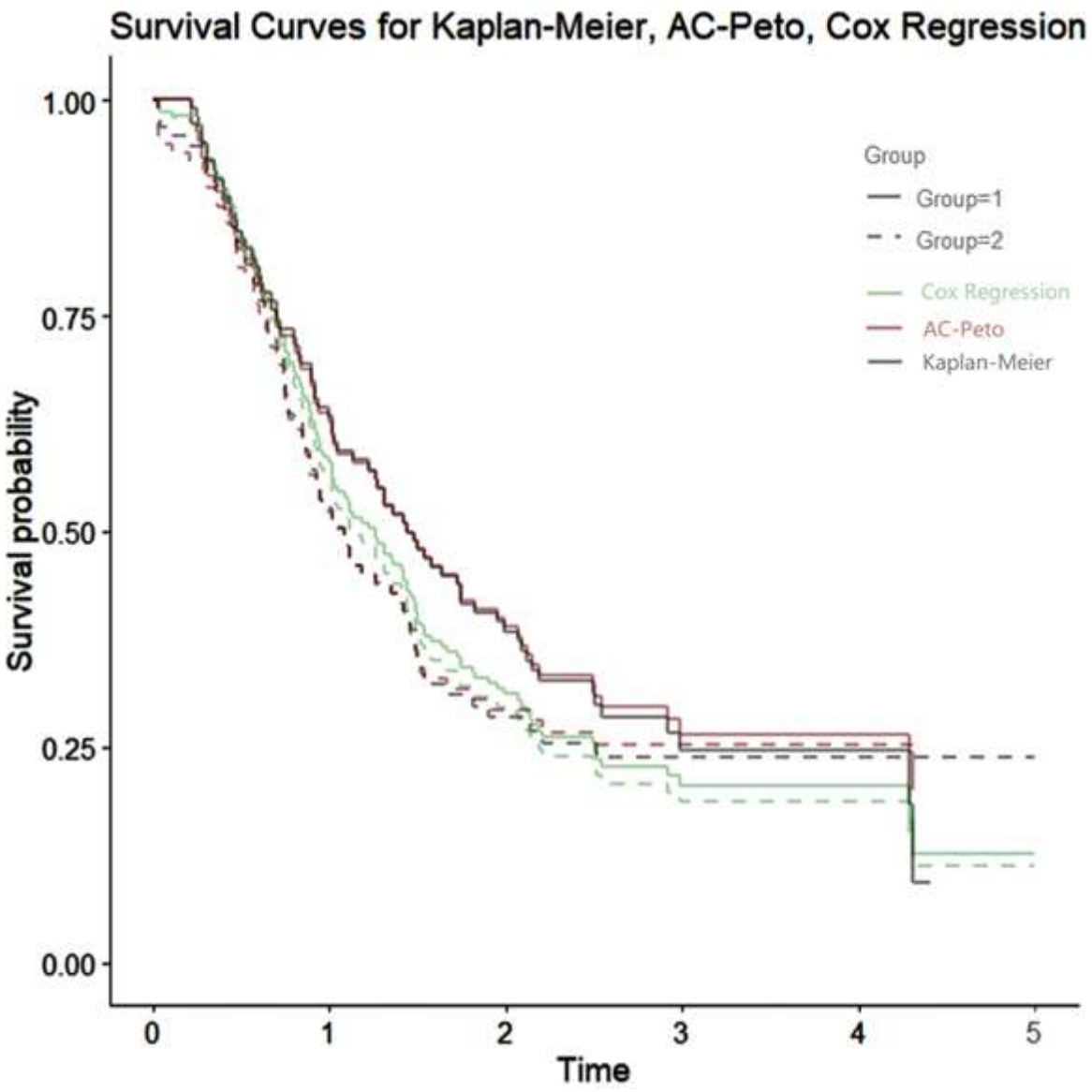


Figure 2

Survival curve comparison between Kaplan-Meier, AC-Peto, and Cox Regression.

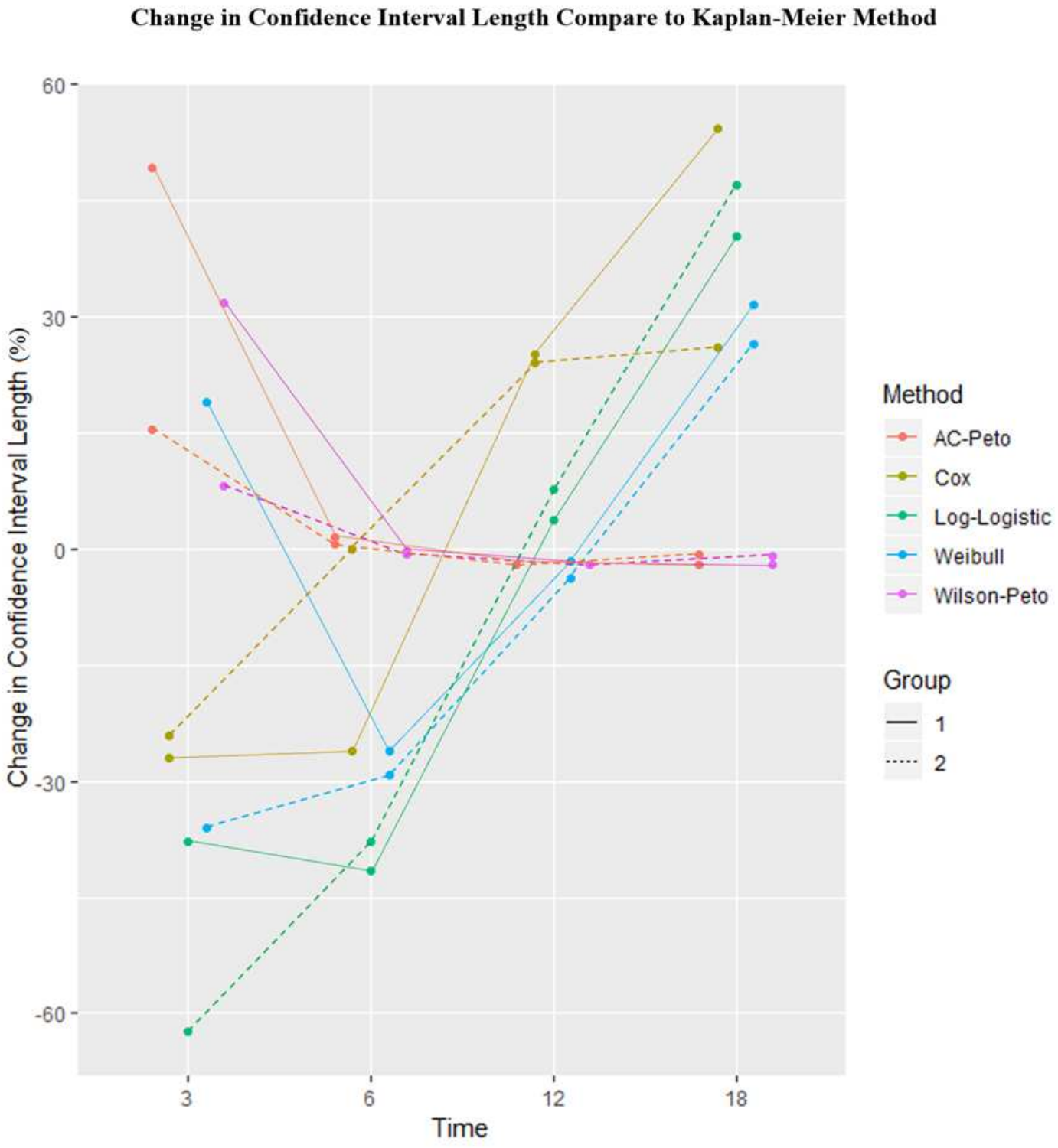


Figure 3

Comparison of confidence interval length changes to the Kaplan-Meier method.

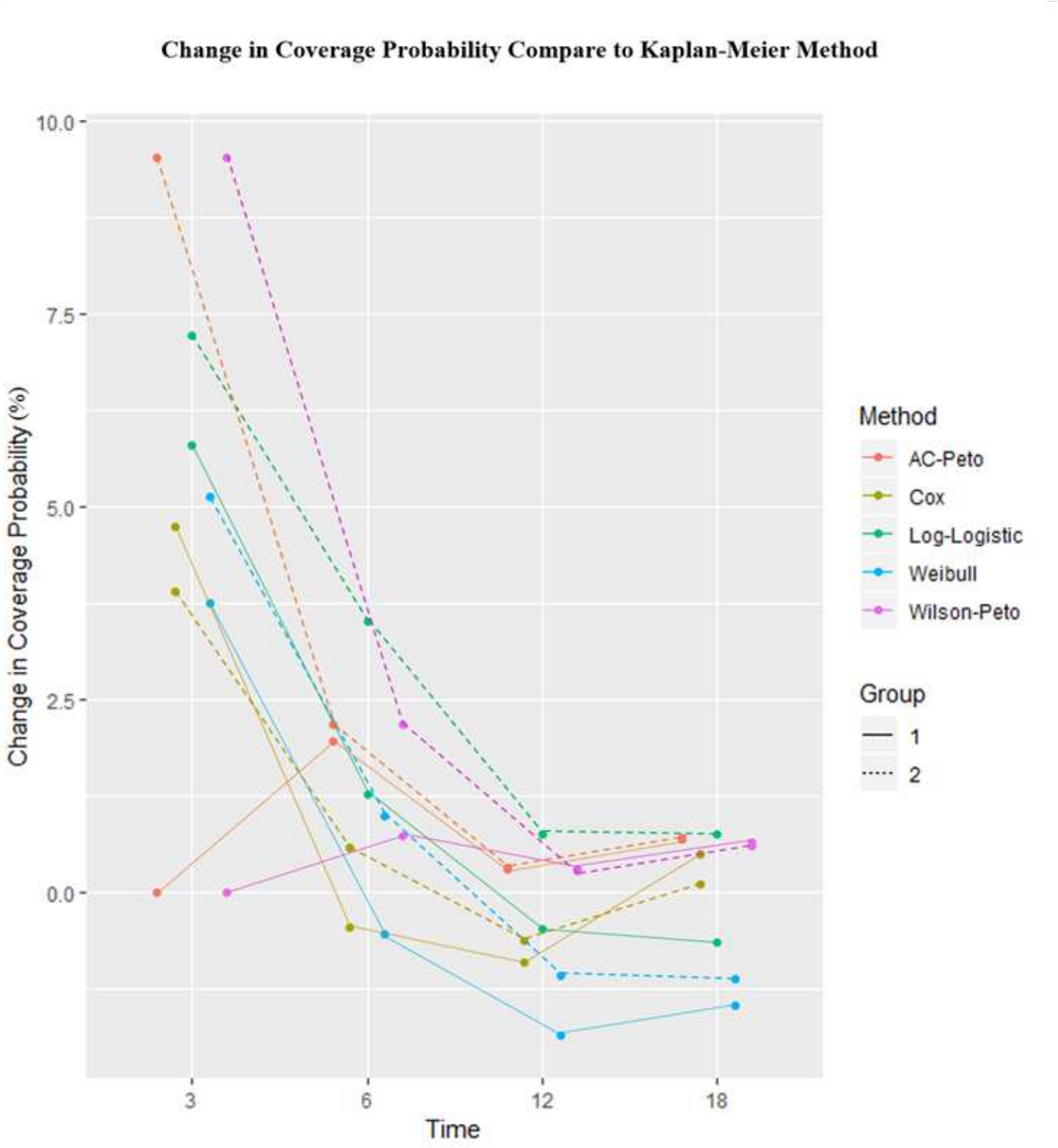


Figure 4

Comparison of coverage probability changes to the Kaplan-Meier method.