Statistical Inference Concerning Minimal Mortality of Patients with Congenital Heart Defects after Surgical Repair

By

Wenhao Mao

Master of Science in Public Health

Biostatistics and Bioinformatics

Eugene Huang, Ph.D.

(Thesis Advisor)

Limin Peng, Ph.D.

(Reader)

Statistical Inference Concerning Minimal Mortality of Patients with Congenital Heart Defects after Surgical Repair

By

Wenhao Mao

B.S.

Peking University

2019

Thesis Advisor: Eugene Huang, Ph.D.

Reader: Limin Peng, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree of
Master of Science in Public Health in
Biostatistics and Bioinformatics
2021

Abstract

Statistical Inference Concerning Minimal Mortality of Patients with Congenital Heart Defects after Surgical Repair

By Wenhao Mao

Background: Congenital heart defect (CHD) is a defect in the heart's structure and function due to abnormal heart development before birth. Survival rate after pediatric cardiac surgery has been improved substantially over the last 2-3 decades. Five-year mortality rate reflects the survival experience of the surviving patient sub-population. As the estimated 5-year mortality function had a sharp decline right after the surgery and then rose gradually, the minimal mortality, which is the minimum point of the mortality function and its timing, may characterize the pattern of the 5-year mortality function. However, statistical inference for the timing and minimum mortality is not a standard problem and warrants investigation.
Methods: In this project, we used the Pediatric Cardiac Care Consortium (PCCC) data, a U.S.-based, multicenter registry of pediatric cardiac surgery, as linked to the National Death Index. We tackled two problems. First, we used subsampling method and divide-and-conquer method to construct 95% confidence intervals for timing of the minimum mortality, as we conjectured that the estimator had a cubic root convergence rate. Our goal was to evaluate the feasibility of these confidence intervals. Second, we developed a bias correction procedure for the standard nonparametric minimal mortality estimator, which had a downward bias.
Results: For subsampling method, from simulations with given parameters, the performance of new confidence interval we developed is the best among our methods. When it comes to simulations mimicking the real dataset, it still had a good performance on coverage when the block size is fixed.

The divide-and-conquer method achieved excellent results on simulated data with parameters as obtained by fitting two datasets and performed badly those by fitting single ventricle dataset. It seemed not to be an enough reliable method for inference.

For bias correction for the point estimator of the minimal mortality, our method can reduce bias and mean squared error (MSE). Also, the confidence intervals were close to the nominal level.

Conclusion: Both subsampling and divide-and-conquer methods are state-of-the-art methods for the challenging statistical problem with cube root asymptotics, and they have different assumptions and requirements. Through extensive simulation studies, we found that, unfortunately, neither approach had consistently reliable performance. Further investigation is warranted.

Also, we recommend our bias correction method, which helped us achieve the estimator with smaller bias and MSE.

Statistical Inference Concerning Minimal Mortality of Patients with Congenital Heart Defects after Surgical Repair

By

Wenhao Mao

B.S.

Peking University

2019

Thesis Advisor: Eugene Huang, Ph.D.

Reader: Limin Peng, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree of
Master of Science in Public Health in
Biostatistics and Bioinformatics
2021

Acknowledgement

First and foremost, I would like to thank the Biostatistics Department at Emory for their constant support in all areas of my education. I would especially like to thank my thesis advisor Eugene Huang, Ph.D. for his guidance. He inspired my interest in survival analysis, helped me choose the thesis topic, and gave some suggestions for the proof and algorithms. His understanding of the cubic root asymptotics and the reference books listed gave me a deeper understanding of my model. He also read the drafts of my thesis carefully and gave me advice, for which I would like to express my sincere gratitude. I would also like to thank my thesis reader Limin Peng, Ph.D. for her careful read and suggestions for my thesis.

I would like to thank the authors of my reference documents, thank them for providing me with a large number of documents so that I have a basis for reference in the process of writing the argument.

I would like to thank the developers of R and C. They allow me to perform calculations and simulations better.

Lastly, I would like to thank my family for their understanding and support so that I can complete my thesis writing exclusively. They talked with me through every situation whether I got panicked, frustrated, or celebratory.

Contents

1. Introduction

Congenital heart defect (CHD) is a defect in the structure and function of the heart due to abnormal heart development before birth (Mendis et al., 2011). In 2017, the global prevalence of congenital heart defect at birth was estimated to be nearly 1.8 cases per 100 live births, a 4.2% increase since 1990 (Zimmerman et al., 2020). Survival after pediatric cardiac surgery was improved substantially even for patients with complex defects (Vinocur et al., 2013; Kempny et al., 2017; Jacobs et al., 2018). However, for the high socio-demographic index countries, the relative importance of congenital heart disease as a cause of child mortality was rapidly increasing, as evidenced by the increase in the proportion of deaths due to congenital heart disease from 1990 to 2017 (Zimmerman et al., 2020). Some data on long-term outcomes after congenital heart surgery have been reported from some European countries with national health systems (Erikssen et al., 2015; Nieminen et al., 2001; Raissadati et al., 2015; Larsen et al., 2017). In the last few years, investigators linked the Pediatric Cardiac Care Consortium (PCCC) data, a large U.S.-based registry of interventions for CHD that has collected patient-level data since 1982, to the National Death Index (NDI), creating a cohort describing the long-term mortality of patients with repaired CHD more thoroughly than prior studies (Spector et al., 2016). In 2018, the long-term survival of patients who were operated on for congenital heart defects was evaluated, which also showed survival had improved over time but still lagged behind the general population (Spector et al., 2018).

We began to investigate the mortality rate over a 5-year window. The 5-year mortality rate is the death rate during the following 5-year period. It can help us better understand the survival of patients after surgery. 5-year mortality function is a function of time t, which represents 5-

year mortality rate at time t. For the CHD dataset, the shape of the nonparametric-estimated 5-year mortality function had a sharp decline at the beginning and then rose gradually. We defined the minimum point of 5-year mortality function as the minimal mortality. The minimal mortality and its timing may give us the pattern of the 5-year mortality function. Nevertheless, research for the estimated 5-year minimum mortality rate after congenital heart surgery in the United States was lacking. How to estimate the minimum mortality became an essential issue.

In this project, we wanted to solve two problems. First, we wanted to evaluate the feasibility of confidence interval estimation for the timing of 5-year minimum mortality. We used Kaplan–Meier estimator to estimate the survival function, from which a 5-year mortality function could be subsequently obtained. The timing and value of the minimum 5-year mortality are then estimated via minimization of the mortality function. As the estimator for timing minimizes an approximate mean process, which is mean of bounded variables, we conjecture that the estimator had a cubic root convergence rate and its limiting distribution is not normal and difficult to estimate on the basis of asymptotic theory (Kim & Pollard, 1990). The bootstrap method did not consistently estimate the asymptotic distribution for estimators with cube root convergence (Ou et al., 2016; Abrevaya et al., 2005), so the bootstrap method was invalid in this situation. The subsampling method is a well-known method to construct confidence intervals for cubic root asymptotics (Politis et al., 1999). Apart from it, many other methods can also be applied for cube root asymptotics. Cattaneo, Jansson, and Nagasawa (2020) proposed a generic and easy-to-implement bootstrap-based distributional approximation, in which they added a generical estimation into every bootstrap sample and so they could adjust the bootstrap method applicable in the context of cube root asymptotics. Lee and Pun (2006) extended the scope of m out of n bootstrap applications to a general class of nonstandard M-estimation problems which included cube root asymptotics. Some papers showed that the divide-and-conquer principle worked well in non-standard problems where rates of convergence are typically slower than $\sqrt{n}$ under some conditions, but this method was often used in the analysis of massive data sets (Shi et al., 2018; Banerjee et al., 2019). These methods are asymptotically justified, but their finite-sample performance, in particular for our application to the PCCC-NDI cohort, is unclear. Our goal is to provide an evaluation for some of these methods for confidence interval estimation on the CHD dataset. In the simulation, we used mixture Weibull distribution to mimic the survival distribution.

Second, the minimum mortality was informative. It might give us more information about the pattern of the 5-year mortality rate function. So, we also constructed the point estimator and 95% confidence interval of it. The traditional nonparametric estimator had a downward bias, so we used extrapolation to correct this bias. Also, we used simulated data to evaluate the feasibility of its confidence interval estimation.

2. Materials and Method

2.1 Data Sources

Recently, investigators linked the Pediatric Cardiac Care Consortium data to the National Death Index, creating a cohort describing patients' long-term mortality with repaired CHD, which had longer follow-up time than prior studies. Using this linkage, we analyzed the 5-year mortality in a large cohort of patients who survived their first CHD operation (1982 to 2003) with NDI follow-up through 2014.

2.1.1 Pediatric Cardiac Care Consortium data

The Pediatric Heart Diseases Data Registry Core provided access to rich registry data from electrophysiologic, surgical and catheter-based studies and interventions for multiple pediatric heart diseases. The data represented over 300,000 event outcomes and had been collected from over 140,000 patients since 1982. Investigators queried the PCCC registry for patients who were U.S. residents and underwent congenital heart surgery (CHS) in a U.S. center between January 1, 1982, and April 15, 2003. All patients <21 years of age were included except isolated ductal ligation in pre-term infants weighing <2.5 kg because of the significant morbidity associated with prematurity rather than the CHD itself.

2.1.2 Death Ascertainment

NDI was a centralized database of death record information compiled from state vital statistics offices. Investigators ascertained death from the PCCC and by matching to NDI records through December 31, 2014. Records submitted to the NDI included first name, middle initial (when available), surname, date of birth, sex, state of last known residence, and state of birth (imputed from state of residence for those <1 year of age at first surgery). Investigators could not link patients without these identifiers and excluded them from the datasets (Spector et al., 2018).

2.1.3 Three Datasets

This project mainly used three datasets, a simple Tetralogy of Fallot dataset, a mild congenital heart defect dataset, and a single ventricle dataset. Simple Tetralogy of Fallot dataset contained observations from 3283 patients with simple Tetralogy of Fallot. Single ventricle dataset consisted of observations from 3818 patients with single ventricle defects and the other dataset included observations from 14861 patients with mild congenital heart defects. We regarded the follow-up time to death, which was a continuous variable, as the outcome, and every patient also had a censoring indicator (1: dead; 0: censored) .

2.2 Software and Nominal Level

All data were analyzed in R version 4.0.3 and C language version 17. The nominal coverage probability of confidence interval is set at 0.95.

2.3 Statistical Model

2.3.1 Data Setting

Suppose T_1, T_2,…, T_n are i.i.d. nonnegative random variables (failure time) with common continuous distribution function F with corresponding survival function S(·) = 1-F(·). Suppose C_1, C_2,…, C_n are i.i.d. nonnegative random variables (censoring time) with common distribution function G. Assume that failure time and censoring time are independent. In our setting of survival analysis data with random right censorship, we observed the bivariate sample (X_1, δ_1), …, (X_n, δ_n), where $\mathrm{X}_i = \min\{\mathrm{T}_i, \mathrm{C}_i\}$, $\delta_i = \mathrm{I}\{\mathrm{T}_i \leq \mathrm{C}_i\}$ with $\mathrm{I}\{\cdot\}$ denoting the indicator function on a set.

Let M(·) be the 5-year mortality function, which was

$$\mathrm{M(t)}{=}1-\frac{S(t+5)}{S(t)}. \qquad (1)$$

We were interested in the 5-year minimal mortality:

$$\min_t M(t),$$

and its timing:

$$\operatorname*{argmin}_t M(t).$$

2.3.2 Model

We can use Kaplan–Meier estimator to estimate the survival function,

$$\hat{S}(\mathrm{t})=\prod_{i:\, t_i \le t}(1-\frac{d_i}{n_i}), \qquad (2)$$

with t_i a time when at least one event happened, d_i the number of events that happened at time t_i, and n_i the individuals known to have survived up to time t_i.

From this, we can subsequently obtain a nonparametric estimator of the 5-year mortality function,

$$\widehat{M}(\mathrm{t})=1-\frac{\hat{S}(t+5)}{\hat{S}(t)}. \qquad (3)$$

The timing of the minimal 5-year mortality was then estimated by the minimizer of the mortality function.

As Kaplan–Meier estimator can be represented as a mean process, with a remainder of order $n^{-3/4}(\mathrm{logn})^{-3/4}$ (Lo, S. H., & Singh, K., 1986) and its estimated 5-year mortality function was 1 minus the ratio of the estimated survival function. Thus we expected that the estimated mortality function should also be approximately a mean process. Therefore, as the estimator for timing of the minimal 5-year mortality minimizes such a process, we conjectured that the estimator had a cubic root convergence rate, as studied in Kim & Pollard (1990). We used the subsampling method and divide-and-conquer method to construct the 95% confidence interval for the timing.

So, we obtained the point estimators and 95% confidence intervals for minimal mortality and its timing. In this project, we had two goals. The first goal was to evaluate the feasibility of confidence interval estimation for the timing. The second goal was to improve the point estimator of the minimal mortality as the traditional nonparametric estimator had a downward bias.

2.4 Inference for the timing – subsampling method

We first considered using subsampling method to construct the confidence interval since the bootstrap didn't consistently estimate the asymptotic distribution for estimators with cube-root convergence (Abrevaya et al., 2005). The subsampling method described below was from Politis et al. (1999). They also showed subsampling can produce consistent estimated sampling distributions under extremely weak assumptions even when the bootstrap failed and it can be used to obtain confidence intervals for parameter estimates.

2.4.1 Asymmetric Confidence Interval

To obtain confidence intervals for the minimizer of (3), $\hat{t}$, we produced subsamples K_1, $K_2, \ldots$, K_{N_n}, where K_j's are the N_n=$\binom{n}{b}$ distinct subsets of {(X_i, δ_i), i=1, ⋯, n} of block size b. Let t_0 denote the true timing and $\hat{t}_{b,j}$ denote the estimated minimizer of (3) using the K_jth dataset.

Define

$L_{n,b}(x)$=${N_n}^{-1}\sum_{j=1}^{N_n} I\,[b^{1/3}(\hat{t}_{b,j} - \hat{t}) \le x]$, and $c_{n,b}(\gamma)$= inf $\{x{:}\, L_{n,b}(x) \ \ge\ \gamma\}$.

From Theorem 2.2.1 of Politis et al. (1999), for any 0 < γ < 1,

$$\mathrm{P}\,[n^{1/3}(\hat{t} - t_0) \ \le\ c_{n,b}(\gamma)] \rightarrow \gamma,$$

under the condition that b → ∞ as n → ∞ and b/n → 0, it followed that for any 0 < α < 1,

$$\mathrm{P}\,[c_{n,b}(\tfrac{\alpha}{2}) \ <\ n^{1/3}(\hat{t} - t_0) \ \le\ c_{n,b}(1-\tfrac{\alpha}{2})] \rightarrow 1 - \alpha.$$

Thus an asymptotic 1 – α level confidence interval for t_0 can be constructed with

$$[\hat{t} - n^{-1/3}c_{n,b}(1-\tfrac{\alpha}{2}), \hat{t} - n^{-1/3}c_{n,b}(\tfrac{\alpha}{2})]. \qquad (4)$$

To respect the range of the estimand, which was a positive half line, we did a log transformation first and then transformed it back.

Define

$L'_{n,b}(x)$=$N_n{}^{-1}\sum_{j=1}^{N_n} I\,[b^{1/3}(log(\hat{t}_{b,j}) - log(\hat{t})) \leq x]$, and $c'_{n,b}(\gamma)$= inf $\{x\colon L'_{n,b}(x) \geq \gamma\}$.

Similarly, for any 0 < γ < 1,

$$\mathrm{P}\,[n^{1/3}(\,log(\hat{t}) - log(t_0))] \leq c'_{n,b}(\gamma)] \rightarrow \gamma,$$

under the condition that b → ∞ as n → ∞ and b/n → 0. It followed that for any 0 < α < 1,

$$\mathrm{P}\,[c'_{n,b}(\tfrac{\alpha}{2}) < n^{1/3}(\,log(\hat{t}) - log(t_0)) \leq c'_{n,b}(1-\tfrac{\alpha}{2})] \rightarrow 1 - \alpha$$

thus an asymptotic 1 − α level confidence interval for t_0 can be constructed with

$$[exp\{log(\hat{t}) - n^{-1/3}c'_{n,b}(1-\tfrac{\alpha}{2})\}, exp\{log(\hat{t}) - n^{-1/3}c'_{n,b}(\tfrac{\alpha}{2})\}]. \qquad (5)$$

2.4.2 Symmetric Confidence Interval

Symmetric confidence intervals can be obtained by modifying the above approach slightly.

Define

$\tilde{L}_{n,b}(x)$=$N_n{}^{-1}\sum_{j=1}^{N_n} I\,[b^{1/3}|\hat{t}_{b,j} - \hat{t}| \leq x]$, and $\tilde{c}_{n,b}(\gamma)$= inf $\{x\colon \tilde{L}_{n,b}(x) \geq \gamma\}$.

Again, if b → ∞ as n → ∞ and b/n → 0, a symmetric confidence interval for t_0 can be constructed as

$$[\hat{t} - n^{-1/3}\tilde{c}_{n,b}(1-\alpha), \hat{t} + n^{-1/3}\tilde{c}_{n,b}(1-\alpha)]. \qquad (6)$$

To respect the range of the estimand, we also had symmetric confidence intervals after a log transformation.

Define

$\tilde{L}'_{n,b}(x)$=$N_n{}^{-1}\sum_{j=1}^{N_n} I\,[b^{1/3}|log(\hat{t}_{b,j}) - log(\hat{t})| \leq x]$, and $\tilde{c}'_{n,b}(\gamma)$= inf $\{x\colon \tilde{L}'_{n,b}(x) \geq \gamma\}$.

Again, if b → ∞ as n → ∞ and b/n → 0, a confidence interval for t_0 can be constructed as

$$[exp\{log(\hat{t}) - n^{-1/3}\tilde{c}'_{n,b}(1-\alpha)\}, exp\{log(\hat{t}) + n^{-1/3}\tilde{c}'_{n,b}(1-\alpha)\}]. \qquad (7)$$

2.4.3 Confidence Interval we developed

In the spirit of bootstrap percentile confidence interval, we can take the empirical 100($\frac{\alpha}{2}$) and

$100(1-\frac{\alpha}{2})$ percentiles from the subsampling replicates as the left and right end points, and have another two asymptotic $1-\alpha$ level confidence intervals

$$[\hat{t} + n^{-1/3}c_{n,b}(\frac{\alpha}{2}), \hat{t} + n^{-1/3}c_{n,b}(1-\frac{\alpha}{2})] \tag{8}$$

and

$$[exp\{log(\hat{t}) + n^{-1/3}c'_{n,b}(\frac{\alpha}{2})\}, exp\{log(\hat{t}) + n^{-1/3}c'_{n,b}(1-\frac{\alpha}{2})\}]. \tag{9}$$

(4)-(9) are 6 different asymptotic confidence intervals from subsampling method.

2.4.4 Algorithm

To avoid large scale computation issues, a stochastic approximation was employed where B randomly chosen datasets from {1, 2, ⋯ , Nn} were used in the above calculation, which was mentioned in Politis et al. (1999). Furthermore, the block size was chosen and the algorithm for choosing block size is described below:

1. Fix a selection of reasonable block sizes b between limits $\mathrm{b_{low}}$ and $\mathrm{b_{up}}$.
2. Draw M bootstrap samples from the actual dataset.
3. For each bootstrap sample, construct a subsampling symmetric confidence interval with asymptotic coverage 1 - α for each block size b. Let $R_{m,b}$ be one if $\hat{t}$ was within the m-th interval based on block size b and zero otherwise.
4. Compute $\hat{h}(b) = M^{-1}\sum_{m=1}^{M} R_{m,b}$.
5. Find the value $\tilde{b}$ that minimizes $|\hat{h}(b) - \alpha|$ and use $\tilde{b}$ as the block size when constructing confidence interval for the original data.

2.5 Inference for the timing – divide-and-conquer method

2.5.1 Pooled Estimator Based on Mean

Banerjee et al. (2019) showed that the divide-and-conquer principle worked well in nonstandard problems where convergence rates were typically slower than $\sqrt{n}$ under some conditions, though this method was often used in the analysis of massive data sets. So, we can also use the divide-and-conquer method to construct the confidence interval for the estimator with cubic root convergence rate. The algorithm is described below:

1. Divide the set of samples (X_1, δ_1), …, $(\mathrm{X}_\mathrm{n}, \delta_\mathrm{n})$into m disjoint subsets S_1, …, S_m of (approximately) equal size.
2. For each j = 1, …, m, compute the estimated minimizer of $\hat{t}_j$ (3) based on the S_jth dataset
3. Average together these estimators to obtain the final "pooled" estimator:

$$\bar{t} = \frac{1}{m} \sum_{j=1}^{m} \hat{t}_j.$$

From Remark 2.2 of Banerjee et al., (2019), we can construct an approximate $(1 - \alpha)$ CI for t_0

$$[\bar{t} - \frac{\hat{\sigma}}{r_n\sqrt{m}} t_{\alpha/2,m-1},\ \bar{t} - \frac{\hat{\sigma}}{r_n\sqrt{m}} t_{\alpha/2,m-1}]$$

where $t_{\alpha/2,m-1}$ denotes the $\frac{\alpha}{2}$th quantile of the t -distribution with m−1 degrees of freedom, $r_n = [\frac{n}{m}]^{1/3}$, $\hat{\sigma}^2 = \frac{{r_n}^2}{m-1} \sum_{j=1}^{m} (\hat{t}_j - \bar{t})^2$.

To respect the range of the estimand, we also did a log transformation first and then transformed it back. Let

$$\overline{lt} = \frac{1}{m} \sum_{j=1}^{m} log(\hat{t}_j).$$

The corresponding approximate $(1 - \alpha)$ CI for t_0 is:

$$[exp(\overline{lt} - \frac{\hat{\sigma}\prime}{r_n\sqrt{m}} t_{\alpha/2,m-1}),\ exp(\overline{lt} - \frac{\hat{\sigma}\prime}{r_n\sqrt{m}} t_{\alpha/2,m-1})], \qquad (10)$$

where $\hat{\sigma}'^2 = \frac{{r_n}^2}{m-1} \sum_{j=1}^{m} [log(\hat{t}_{j)} - log(\bar{t})]^2$

2.5.2 Pooled Estimator Based on Median

As we found that the empirical distribution of the subsample estimator was skewed, we came up with using the median as the "pooled" estimator, which might be more robust to the skewed distribution. So we had another choice for the final “pooled” estimator and that was

$$\tilde{t} = median\{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_m\}$$

From this, we can construct an approximate $(1 - \alpha)$ exact CI for t_0 based on Clopper–Pearson interval (Clopper & Pearson, 1934):

$$[\hat{t}_{(L)}, \hat{t}_{(U)}], \tag{11}$$

where L= sup {k: $2^{-m} \sum_{i=0}^{k-1} \binom{m}{i} \leq \frac{\alpha}{2}$} and U= inf {k: $2^{-m} \sum_{i=k-1}^{m} \binom{m}{i} \leq \frac{\alpha}{2}$}.

2.6 Minimal Mortality

2.6.1 Bias of the point estimator

As we can achieve point estimator $\hat{t}$ for timing of the minimal 5-year mortality from (3), we plugged it in and get a point estimator for minimal mortality $\widehat{M}(\hat{t})$.

$$\widehat{M}(\hat{t}) - M(t_0) = \{ \widehat{M}(t_0) - M(t_0)\} + \{\widehat{M}(\hat{t}) - \widehat{M}(t_0)\}$$

For the first term, as $\widehat{M}$(t) had the similar property of $\hat{S}$(t) and it was an approximate mean process, it was asymptotically normal, of order $O_p(n^{-1/2})$ (Lo & Singh, 1986),

$$\widehat{M}(t_0) - M(t_0) = n^{-1} \sum_{i=1}^{n} \xi(\mathrm{X_i}, \delta_\mathrm{i}, t_0) + r_n(t),$$

where $\mathrm{E}[\xi(\mathrm{X_i}, \delta_\mathrm{i}, t_0)]=0$, $r_n(t)=O_p(n^{-3/4}(\mathrm{logn})^{-3/4})$.

So, $\mathrm{E}(\widehat{M}(t_0) - M(t_0)) = O_p(n^{-3/4}(\mathrm{logn})^{-3/4})$.

For the second term,

$$\widehat{M}(\hat{t}) - \widehat{M}(t_0) = \min_t \widehat{M}(t) - \widehat{M}(t_0),$$

where $n^{2/3}\{\widehat{M}(\hat{t}) - \widehat{M}(t_0)\}$ coverages weakly to a Gaussian process (Kim & Pollard, 1990). Therefore, $n^{2/3}\{\widehat{M}(\hat{t}) - \widehat{M}(t_0)\}$ coverages to a distribution, and the second-order asymptotic bias was of order $n^{-2/3}$, which meant $\mathrm{E}[\widehat{M}(\hat{t}) - \widehat{M}(t_0)] = O(n^{-2/3})$.

So, $\mathrm{E}[\widehat{M}(\hat{t}) - \widehat{M}(t_0)]$ was main influence in $\mathrm{E}[\widehat{M}(\hat{t}) - M(t_0)]$.

$$\mathrm{E}[\widehat{M}(\hat{t}) - M(t_0)] = O(n^{-2/3}).$$

Also, as $\hat{t} = \underset{t}{\operatorname{argmin}} \widehat{M}(\mathrm{t})$, $\widehat{M}(\hat{t}) - \widehat{M}(t_0) \leq 0$, so $\mathrm{E}[\widehat{M}(\hat{t}) - \widehat{M}(t_0)] \leq 0$.

Therefore, the estimator had a downward bias and we wanted to do bias correction, which actually was a second term bias correction.

2.6.2 Algorithms

As $\mathrm{E}[\widehat{M}(\hat{t}) - M(t_0)] = O(n^{-2/3})$, we can use this linear relationship to do extrapolation, which was a method of bias correction. The algorithm 2.6.1 is described below:

1. Draw M random subsamples with size (0.25+0.05i)n from the actual dataset, i = 1, 2, …, 14.
2. From every subsample, we can get the estimated minimal mortality $\widehat{M}[\hat{t}_{(0.25+0.05\mathrm{i})\mathrm{n,j}}]$, j=1, …, M.
3. Let $\widehat{M}[\hat{t}\big((0.25 + 0.05\mathrm{i})\mathrm{n}\big)] = \frac{1}{M}\sum_{j=1}^{M} \widehat{M}[\hat{t}_{(0.25+0.05\mathrm{i})\mathrm{n,j}}]$.
4. Fit a simple linear regression from {$((0.25 + 0.05\mathrm{i})\mathrm{n})^{-2/3}$, ($\widehat{M}[\hat{t}\big((0.25 + 0.05\mathrm{i})\mathrm{n}\big)]$) and point estimator}.
5. Get $\lim_{x \to \infty} \widehat{M}[\hat{t}(\mathrm{x})]$ from this simple linear regression.

$\lim_{x \to \infty} \widehat{M}[\hat{t}(\mathrm{x})]$ was our adjusted estimator for minimal mortality. Also, we can change subsample size in Step 1 and get different adjusted estimator.

Also, we wanted to construct 1-α confidence interval for $\widehat{M}(\hat{t})$.

$$\widehat{M}(\mathrm{t}) = \frac{\hat{S}(t) - \hat{S}(t+5)}{\hat{S}(t)} \text{ and } \hat{S}(\mathrm{t}) = \prod_{i:\, t_i \leq t} (1 - \frac{d_i}{n_i})$$

Let $\widehat{M}(\hat{t}) = \widehat{M_c}(5)$, where $1 - \widehat{M_c}(\mathrm{t}) = \frac{\hat{S}(\hat{t}+t)}{\hat{S}(\hat{t})} = \prod_{i:\, \hat{t} \leq t_i \leq t + \hat{t}} (1 - \frac{d_i}{n_i})$

Actually, $1 - \widehat{M_c}(t)$ is a Kaplan–Meier estimator based on those people whose observed time > $\hat{t}$ and their adjusted observed time become their original observed time - $\hat{t}$.

The Greenwood's formula was used to achieve the variance of $\frac{\hat{S}(\hat{t}+t)}{\hat{S}(\hat{t})}$ (Greenwood 1926):

$$\widehat{Var}\{1-\widehat{M_c}(\mathrm{t})\} = \{1-\widehat{M_c}(\mathrm{t})\}^2 \sum_{\hat{t}\le t_k \le t+\hat{t}} (\frac{d_k}{n_k(n_k-d_k)})$$

As $\widehat{M}(\hat{t}) = \widehat{M_c}(5)$, $\widehat{Var}\{\widehat{\mathrm{M}}(\hat{\mathrm{t}})\} = \widehat{Var}\{1-\widehat{M_c}(5)\} = \{1-\widehat{M_c}(5)\}^2 \sum_{\hat{t}\le t_k \le 5+\hat{t}} (\frac{d_k}{n_k(n_k-d_k)})$.

As $0 \le \widehat{\mathrm{M}}(\hat{\mathrm{t}}) \le 1$, so we wanted to do a logit transformation first and then transform it back.

By delta method, we can get $\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\}= \widehat{Var}\{\widehat{\mathrm{M}}(\hat{\mathrm{t}})\}/ [\widehat{\mathrm{M}}(\hat{\mathrm{t}})\left(1-\widehat{\mathrm{M}}(\hat{\mathrm{t}})\right)]^2$.

An asymptotic 1 – α level confidence interval for $M(t_0)$ can be constructed with

$$[\frac{\exp(logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))-\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}{1+\exp(logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))-\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}, \frac{\exp(logit\left(\widehat{\mathrm{M}}(\hat{\mathrm{t}})\right)+\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}{1+\exp(logit\left(\widehat{\mathrm{M}}(\hat{\mathrm{t}})\right)+\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}].$$

where denotes the $(1-\frac{\alpha}{2})$th quantile of the standard normal distribution.

Also, we had another algorithm 2.6.2:

1. Draw M random subsamples with size 0.5n from the actual dataset.

2. From every subsample, we can get the estimated minimal mortality $\widehat{M}[\hat{t}_{0.5\mathrm{n,j}}]$, j=1, …, M.

3. Let $\widehat{M}[\hat{t}(0.5\mathrm{n})]= \frac{1}{M}\sum_{j=1}^{M}\widehat{M}[\hat{t}_{0.5\mathrm{n,j}}]$.

4. As $\mathrm{E}[\widehat{M}(\hat{t})-M(t_0)]=O_p(n^{-2/3})$, Let $\mathrm{E}[\widehat{M}(\hat{t})-M(t_0)]=\alpha+\beta\cdot n^{-2/3}$. By using $\hat{t}(0.5\mathrm{n})$ and point estimator, we estimate the slope $\hat{\beta} = \{\mathrm{M}[\hat{t}(0.5\mathrm{n})]\text{- }\mathrm{M}(\hat{t})\}/[(0.5\mathrm{n})^{-2/3}-n^{-2/3}]$.

   Then we can estimate $\hat{\alpha}=\widehat{M}(\hat{t}) - n^{-2/3}\cdot\{\mathrm{M}[\hat{t}(0.5\mathrm{n})]\text{- }\mathrm{M}(\hat{t})\}/[(0.5\mathrm{n})^{-2/3}-n^{-2/3}]$.

5. Then we can get $\lim_{x\to\infty}\widehat{M}[\hat{t}(\mathrm{x})] = \hat{\alpha}$.

Then we can have the same inference on this estimator.

An asymptotic 1 – α level confidence interval for $M(t_0)$ can be constructed with

$$[\frac{\exp(logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))-\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}{1+\exp(logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))-\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}, \frac{\exp(logit\left(\widehat{\mathrm{M}}(\hat{\mathrm{t}})\right)+\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}{1+\exp(logit\left(\widehat{\mathrm{M}}(\hat{\mathrm{t}})\right)+\mathrm{Z}_{1-\alpha/2}\cdot\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\})}].$$

where denotes the $(1-\frac{\alpha}{2})$th quantile of the standard normal distribution, and we can calculate $\widehat{Var}\{logit(\widehat{\mathrm{M}}(\hat{\mathrm{t}}))\}$ with the same form as before.

3. Simulation

3.1 simulation models

We used mixture Weibull distribution to mimic the survival distribution, combining two Weibull distribution with different parameters.

$$f(x) = w \cdot \frac{a_1}{b_1}(\frac{x}{b_1})^{(a_1-1)}e^{-(\frac{x}{b_1})^{a_1}} + (1-w) \cdot \frac{a_2}{b_2}(\frac{x}{b_2})^{(a_2-1)}e^{-(\frac{x}{b_2})^{a_2}}$$

The shape of the nonparametric-estimated 5-year mortality function had a sharp decline at the beginning and then rose gradually. However, we could not get such a 5-year mortality function from the exponential distribution or Weibull distribution, which was often used in the survival analysis, because 5-year mortality rate functions from them were monotone. The mixture Weibull distribution had more parameters and gave us a more flexible choice. Therefore, the mixture Weibull distribution was a reasonable assumption.

First, we used fixed and reasonable parameters to simulate a dataset without censoring, which gave us insight into our methods' performance.

Second, we used mixture Weibull distribution to fit the real dataset by weibullRMM_SEM function in mixtools package in R and got the estimated parameters for the distribution. To fit the data better, we artificially censored some data set at a fixed time point and used empirical censoring distribution to mimic censoring distribution. With the mixture Weibull distribution with different parameters and empirical censoring distribution, we could evaluate confidence intervals through their coverages and average lengths.

3.2 Evaluation of inference for the timing – subsampling method

As we can achieve 6 different asymptotic confidence intervals from subsampling method. In this section, we use different name to call these 6 confidence intervals. Asymmetric standard

CI refers to the confidence interval (4) in the method part. Asymmetric log CI refers to the confidence interval (5) in the method part. Symmetric standard CI refers to the confidence interval (6) in the method part. Symmetric log CI refers to the confidence interval (7) in the method part. New standard CI refers to the confidence interval (8) in the method part. New log CI refers to the confidence interval (9) in the method part.

3.2.1 Artificial Data with Given Parameters

In this simulation, we set a_1= 1, b_1= 2, a_2 = 5, b_2 = 75, w = 0.2, in which there is no censoring. We chose these values simply because they could give us a similar pattern of the 5-year mortality rate. We used the procedures in the method part to choose the block size. Also, we set the sample size as 200 and run 1000 iterations.

From Table 1, we could see the coverages of asymmetric standard CI and asymmetric log CI were much smaller than the nominal level. Also, the coverages of symmetric standard CI, symmetric log CI, and New standard CI were smaller than the nominal level. The coverage of new log CI was very close to the nominal level. However, these six confidence intervals' average lengths were large, and the average length of new log CI was the smallest.

Table 1. The performances of 95% confidence intervals in the simulation with given parameters with sample size 200

Method	Coverage	Average Length of CI
Asymmetric Standard CI	0.610	5.64
Asymmetric log CI	0.580	7.83
Symmetric Standard CI	0.847	7.41
Symmetric log CI	0.909	7.86
New Standard CI	0.831	6.16
New log CI	0.948	5.41

Then we set the sample size as 1000 and run 1000 iterations with the same parameters.

From Table 2, we could see the coverages of asymmetric standard CI and asymmetric log CI were much smaller than the nominal level, but they were better than those in the previous simulation. The coverage of new standard CI was smaller than the nominal level. The coverages of symmetric standard CI and symmetric log CI were close to the nominal level. The coverage of new log CI was a little larger than the nominal level. These six confidence intervals' average lengths became shorter than those in the previous simulation, and the average length of new log CI was the smallest.

Table 2. The performances of 95% confidence intervals in the simulation with given parameters with sample size 1000

Method	Coverage	Average Length of CI
Asymmetric Standard CI	0.760	4.91
Asymmetric log CI	0.804	4.31
Symmetric Standard CI	0.965	6.24
Symmetric log CI	0.958	4.18
New Standard CI	0.864	4.86
New log CI	0.985	3.80

Then we set the sample size as 4000 and run 500 iterations with the same parameters.

Table 3. The performances of 95% confidence intervals in the simulation with given parameters with sample size 4000

Method	Coverage	Average Length of CI
Asymmetric Standard CI	0.819	3.13
Asymmetric log CI	0.861	2.73
Symmetric Standard CI	0.979	3.81
Symmetric log CI	0.972	2.67
New Standard CI	0.889	3.10
New log CI	0.993	2.53

From Table 3, we could see the coverages of asymmetric standard CI, asymmetric log CI, and new standard CI were still smaller than the nominal level, but they were better than those in the previous simulation. The coverages of symmetric standard CI and symmetric log CI were close to the nominal level. However, they were both a little larger than the nominal level. The coverage of new log CI was larger than the nominal level. These six confidence intervals' average lengths became shorter than those in the previous simulation, and the average length of new log CI was the smallest.

In summary, the average length of new log CI was always the smallest, and its coverage was close to the nominal level, although it was a little larger than the nominal level when the sample size was large. As the conservative confidence interval is acceptable and it had the smallest length, its performance is the best. The performances of symmetric standard CI and symmetric log CI were also good, and their coverage goes to the nominal level as the sample size is larger. However, the average length of symmetric standard CI was a little large.

3.2.2 Artificial Data with Estimated Parameters after Fitting the Real Dataset – Simple Tetralogy of Fallot Dataset

For simple Tetralogy of Fallot dataset, from Figure 1 in Appendix, we could see the curves of parametric estimated mortality function and nonparametric estimated mortality function are very similar to each other in (0,15), although the parametric estimated timing is a little bit different from the nonparametric one. It was an all-right fit and could give us some insight into the performance of subsampling method on simple Tetralogy of Fallot dataset. We set the sample size as 4000 because the size of the original dataset is 3283 and we ran 500 iterations. For the censoring rate, the real censoring rate is 95.31%, and the average censoring rate of the artificial data is 94.07 %, which were very close to each other.

First, we fixed the block size b as $n^{0.8}$, where n is the sample size.

From Table 4, we could see the coverages of five confidence intervals except new log CI were all much smaller than the nominal level. The coverage of new log CI was close to the nominal level. Also, the average length of new log CI was the smallest. However, these six confidence intervals' average lengths are very large.

Table 4. The performances of 95% confidence intervals in the simulation with estimated parameters from simple Tetralogy of Fallot dataset with fixed block size $n^{0.8}$

Method	Coverage	Average Length of CI
Asymmetric Standard CI	0.548	9.65
Asymmetric log CI	0.512	12.63
Symmetric Standard CI	0.795	12.31
Symmetric log CI	0.847	12.73
New Standard CI	0.827	9.65
New log CI	0.943	9.36

Second, we used the procedures in the method part to choose the block size.

Table 5. The performances of 95% confidence intervals in the simulation with estimated parameters from simple Tetralogy of Fallot dataset after choosing the block size

Method	Coverage	Average Length of CI
Asymmetric Standard CI	0.377	4.20
Asymmetric log CI	0.465	14.59
Symmetric Standard CI	0.527	5.53
Symmetric log CI	0.837	14.81
New Standard CI	0.563	4.42
New log CI	0.782	7.06

From Table 5, we could see the coverages of all confidence intervals were much smaller than the nominal level. However, six confidence intervals' average lengths are large. It meant the performances of these six confidence intervals were not good.

3.2.3 Summary for Subsampling Method

We gave six types of confidence intervals by the subsampling method. From simulation with given parameters, we could see Symmetric log CI and New log CI performed well, and their coverages were close to the nominal level, especially new log CI's. Although some of their coverages were slightly larger than the nominal level when the sample size was large, the conservative confidence interval is acceptable. When it comes to the simulation with estimated parameters after fitting the real dataset, only new log CI had a good performance on coverage when the block size is fixed. All lengths of these confidence intervals were large, which meant it was hard to get precise inference for the timing by subsampling method. Also, these six confidence intervals all became worse after we chose the block size. It was not a reliable method for real dataset at present.

3.3 Evaluation of inference for the timing – divide-and-conquer method

The divide-and-conquer method often performs well on massive data. So we only simulated for the mild congenital heart defect dataset, whose sample size was nearly 15000. In this section, we regarded the confidence interval (10) in the method part as mean confidence interval and regarded the confidence interval (11) in the method part as median confidence interval.

3.3.1 Artificial Data with Estimated Parameters after Fitting the Real Dataset - Mild Congenital Heart Defect Dataset

For mild congenital heart defect dataset, from Figure 2 in Appendix, we could see the curves of parametric estimated mortality function and nonparametric estimated mortality function were very similar to each other in (0,15), and the parametric estimated timing was also close to the nonparametric one. It was a good fit and can give us some insights into the performance of subsampling method on mild congenital heart defect dataset. For the censoring rate, the real censoring rate was 97.90%, and the average censoring rate of the artificial data was 98.03%, which were very close to each other. We artificially censored this data set at year 20.

We set the sample size as 16000 and split it into 32 or 16 disjoint subsets. We ran 1000 iterations.

Table 6. The performances of divide-and-conquer method in the simulation with estimated parameters from mild congenital heart defect dataset with sample size 16000

Method	The number of Subsets: m	Coverage	Average Length of CI
mean CI	16	0.897	5.44
	32	0.928	3.64
median CI*	16*	0.978	9.27
	32**	0.965	5.91

* Its exact confidence level is 0.951.

** Its exact confidence level is 0.9649.

From Table 6, we could see the coverage of mean CI, whose subsets' number was 16, was a little smaller than the nominal level. The coverages of the other three were close to the nominal level. Among these three, the average length of mean CI, whose subsets' number was 32, was the smallest.

Also, we set the sample size as 4000 and split it into 20 disjoint subsets. We run 1000 iterations.

Table 7. The performances of divide-and-conquer method in the simulation with estimated parameters from mild congenital heart defect dataset with sample size 4000

Method	The number of Subsets: m	Coverage	Average Length of CI
mean CI	20	0.928	3.91
median CI	20*	0.987	6.31

* Its exact confidence level is 0.9586.

From Table 7, we could see the coverages of mean CI and median CI were close to the nominal level. As the conservative coverage was more acceptable, median CI performed better. However, the average length of mean CI was smaller than that of median CI.

3.3.2 Artificial Data with Estimated Parameters after Fitting the Real Dataset – Simple Tetralogy of Fallot Dataset

We used the mixture Weibull distribution to fit simple Tetralogy of Fallot dataset as by the same way in 3.2.2. As this method can only deal with massive data, we set the sample size as 16000 and split it into 32 or 16 disjoint subsets. We ran 1000 iterations.

Table 8. The performances of divide-and-conquer method in the simulation with estimated parameters from simple Tetralogy of Fallot dataset with sample size 16000

Method	The number of Subsets: m	Coverage	Average Length of CI
mean CI	16	0.917	5.54
	32	0.939	4.11
median CI*	16	0.979	9.61
	32	0.980	7.18

* Its exact confidence level is 0.951.

** Its exact confidence level is 0.9649.

From Table 8, we could see the coverages of all methods were also close to the nominal level, especially mean CI whose subsets' number was 32. Actually, median CI whose subsets' number also perform excellently as its exact confidence level was 0.9649.

Also, we set the sample size as 4000 and split it into 20 disjoint subsets. We ran 1000 iterations.

Table 9. The performances of divide-and-conquer method in the simulation with estimated parameters from simple Tetralogy of Fallot dataset with sample size 4000

Method	The number of Subsets: m	Coverage	Average Length of CI
mean CI	20	0.794	4.83
median CI	20*	0.969	7.91

* Its exact confidence level is 0.9586.

From Table 9, we could see the coverage of median CI was close to the nominal level. However, the coverage of mean CI was terrible.

3.3.3 Artificial Data with Estimated Parameters after Fitting the Real Dataset – Single Ventricle Dataset

For single ventricle dataset, from Figure 3 in Appendix, we could see the curves of parametric estimated mortality function and nonparametric estimated mortality function are very similar to each other although the parametric estimated timing is a little bit different from the nonparametric one. It is an all-right fit and can give us some insight on the performance of subsampling method on single ventricle dataset. As this method can only deal with massive data, we set the sample size as 16000 and split it into 32 disjoint subsets. We ran 1000 iterations.

Table 10. The performances of divide-and-conquer method in the simulation with estimated parameters from single ventricle dataset with sample size 16000

Method	The number of Subsets: m	Coverage	Average Length of CI
mean CI	32	0.259	3.24
median CI*	32	0.805	5.93

* Its exact confidence level is 0.951.

** Its exact confidence level is 0.9649.

From Table 10, we could see the coverages of all methods were much smaller than the nominal level, which meant this method is not reliable on single ventricle dataset.

3.3.3 Summary for Divide-and-Conquer Method

Confidence intervals based on the divide-and-conquer method seemed to have a good performance on the first two datasets, especially the method based on the median whose coverages were always very close to the nominal level. It meant our inference might be very accurate. Surprisingly, the divide-and-conquer method's performances were still outstanding when the sample size dropped from 16000 to 4000 for mild congenital heart defect dataset. However, I'm still not very confident of this method, as it performed pretty badly on single ventricle dataset. In summary, divide-and-conquer method seemed not to be a reliable method for inference.

3.4 Estimator and inference for the Minimal Mortality

In this section, we regarded the estimator and confidence interval from the algorithm 2.6.1 in the method part as method 1 and the estimator and confidence interval from the algorithm 2.6.2 in the method part as method 2. We also gave a Traditional method to compare. In that method, we just used the point estimator as our estimator and used the similar procedures in the method part to get the inference.

3.4.1 Artificial Data with Given Parameters

In this simulation, we set a_1= 0.2, b_1= 0.4, a_2 = 3, b_2 = 25, w = 0.5 with the same reason as that in 3.2.1. Also, we set the sample size as 400 and run 500 iterations.

From Table 11, we could see that our methods both performed better than the traditional one. They had smaller standardized bias, smaller standardized RMSE, better coverage, and smaller average length. The performances of method 1 and method 2 were similar, and method 1 seemed to be a little better as it had smaller standardized RMSE. However, method 2 needs less computation.

Table 11. The performances of bias correction in the simulation with fixed parameters with sample size 400

True value	Method	Standardized bias*	Standardized RMSE*	Coverage	Average Length
0.0752	Traditional	-0.290	0.369	0.976	0.0922
	Method 1	-0.058	0.305	0.964	0.0885
	Method 2	-0.045	0.339	0.944	0.0887

* As the true value is very small, so we use standardized bias and standardized RMSE to evaluate the performance of our estimators. Standardized bias is the difference from the true value divided by that true value and Standardized RMSE is RMSE divided by that true value.

3.4.2 Artificial Data with Estimated Parameters after Fitting the Real Dataset – Simple Tetralogy of Fallot Dataset

We used the mixture Weibull distribution to fit simple Tetralogy of Fallot dataset as by the same way in 3.2.2. We set the sample size as 4000 and ran 1000 iterations.

From Table 12, we could see that our methods still both performed better than the traditional one, and the most pleasant thing was the significant progress in coverage in both methods. They had smaller standardized bias, smaller standardized RMSE, better coverage, and similar average length. The performances of method 1 and method 2 were similar, and method 1 seemed to be a little better as it had smaller standardized RMSE. However, method 2 needs less computation.

Table 12. The performances of bias correction in the simulation with estimated parameters from simple Tetralogy of Fallot dataset with sample size 4000

True value	Method	Standardized bias*	Standardized RMSE*	Coverage	Average Length
0.00687	Traditional	-0.2418	0.273	0.869	0.0047
	Method 1	-0.0550	0.170	0.954	0.0047
	Method 2	-0.0338	0.196	0.930	0.0047

** As the true value is very small, so we use standardized bias and standardized RMSE to evaluate the performance of our estimators. Standardized bias is the difference from the true value divided by that true value and Standardized RMSE is RMSE divided by that true value.

3.4.3 Summary

From our two simulations, we can see both of our methods made bias and MSE smaller, which is the goal we want to reach. Also, the confidence intervals for these two methods were still close to the nominal level. Therefore, both methods could get better estimation for minimal mortality. Method 1 is more reliable for its smaller RMSE. Our bias correction was successful.

4. Application

4.1 Subsampling method on Simple Tetralogy of Fallot Dataset

We applied subsampling method to simple tetralogy of Fallot dataset (N=3283).

Table 13. The 95% confidence intervals for timing on simple tetralogy of fallot dataset by subsampling method with fixed block size $n^{0.8}$

Method	95% CI	The Length of CI
Asymmetric Standard CI	(0, 5.59)	5.59
Asymmetric log CI	(1.45,7.90)	6.45
Symmetric Standard CI	(0, 14.79)	14.79
Symmetric log CI	(1.48, 11.01)	9.53
New Standard CI	(2.43, 15.34)	12.91
New log CI	(2.26, 11.26)	9.00

From Table 13, we could see the confidence intervals for timing with fixed block size $n^{0.8}$.

Table 14. The 95% confidence intervals for timing on simple tetralogy of fallot dataset by subsampling method after choosing the block size

Method	Block Size	95% CI	The Length of CI
Asymmetric Standard CI	57	(0.86, 4.94)	4.08
Asymmetric log CI	57	(2.82, 6.52)	3.70
Symmetric Standard CI	57	(1.15, 6.94)	5.79
Symmetric log CI	57	(2.64, 6.19)	3.55
New Standard CI	57	(3.08, 7.23)	4.15
New log CI	57	(2.02, 5.77)	3.74

From Table 14, we could see the confidence intervals for timing after choosing the block size. After choosing the block size, we achieved narrower confidence intervals.

However, after simulations, I was not very confident about whether they are reliable on real data, although the coverage of new log CI with fixed block size was close to the nominal level. There are two reasons. First, new log CI became more and more conservative as the sample size became larger from the simulation with given parameters, which is weird. As constructing this confidence interval need assumptions, we should be more careful. Second, symmetric log CI and new log CI were based on the same foundation, and symmetric log CI also performed well in the simulation with given parameters. However, symmetric log CI performed badly in the simulation with estimated parameters, and the reason why this happens wasn't clear.

4.2 Divide-and-conquer method on Mild Congenital Heart Defect Dataset

Divide-and-conquer method was applied to mild congenital heart defect dataset.

From Table 15, we can see the confidence intervals for timing by divide-and-conquer method. When the number of subsets is 32, the confidence interval had all-right length, and its length was similar to that from simulation. However, when the number of subsets is 16, the length of the confidence interval was too large, and they were different from those from the simulation.

Therefore, for the real dataset, maybe we should choose the number of subsets more carefully. Also, as we mentioned before, considering the bad performance on single ventricle dataset, it wasn't an enough reliable method.

Table 15. The confidence intervals for timing on mild congenital heart defect dataset by divide-and-conquer method

Method	The number of Subsets: m	95% CI	The Length of CI
mean CI	16	(4.16, 12.66)	8.50
	32	(3.49, 7.10)	3.61
median CI	16	(3.26, 22.25)	18.99
	32	(2.88, 8.32)	5.44

4.3 Bias correction on Simple Tetralogy of Fallot Dataset

We also applied bias correction to simple tetralogy of Fallot dataset. As method 1 is more reliable, we only used method 1 on real data. From Table 16, we could see that our bias correction indeed work for original point estimator and confidence interval.

Table 16. The point estimator and confidence intervals for minimal mortality on simple tetralogy of fallot dataset

Method	Point Estimator	95% CI	The Length of CI
Traditional	0.00529	(0.00328, 0.00851)	0.00523
Method 1	0.00677	(0.00466, 0.00982)	0.00515

5. Discussion and conclusion

In this project, we first evaluated the inference for timing by subsampling method and divide-and-conquer method on CHD dataset.

For subsampling method, from simulations with given parameters, we used subsampling method with block size choosing method and some confidence intervals can achieve the coverage that was close to the nominal level and narrow length, which was what we wanted. When it comes to the simulation with estimated parameters after fitting the real dataset, only new log CI had a good performance on coverage when the block size is fixed. However, the method after choosing block size was unsatisfactory, with bad coverages from the simulation with estimated parameters. Apparently, results from simulations with given parameters and those from simulations with estimated parameters are different. We found that after choosing the block size, the smaller block size was more likely to be chosen, which may cause the inference from simulations with estimated parameters to be more unstable. This could be caused by the lower bound of the block size choices is too low. We need to try to get a more appropriate lower bound of the block size choices. In addition, we increased the sample size to 16000, and the coverages for fixed block sizes were closer to the nominal level. However, the sample size of simple tetralogy of Fallot dataset was only 3283. So, maybe the next step would be to try the method on the simulations for mild congenital heart defect dataset, whose sample size was 14861. At present, this method wasn't reliable.

For divide-and-conquer method, it performed well from simulations on the first two datasets, which can help us achieve confidence intervals with good coverage and narrow length. Median CI was more reliable as its coverage seemed to stay close to the nominal level from simulations. However, it performed pretty badly on simulated data from single ventricle dataset. It seemed not to be an enough credible method for inference. We need more studies on this method. Also, the result from the real dataset also showed the number of subsets might be a significant decision. When the number of subsets was 16, the lengths of both confidence intervals were too large on the real dataset, and they were different from those from the simulation. The reason needed further study.

Second, we used extrapolation to do bias correction for the estimator for the minimal mortality.

We could see both of our methods made bias and MSE smaller from our simulation, which was the goal we wanted to reach. The confidence intervals for these two methods were still close to the nominal level. So, our bias correction was very successful. Also, for the real dataset, we could see our estimators both had an increase on point estimator, which may improve downward bias.

In this project, the strength was that we used nonparametric methods to construct confidence intervals for timing, and they didn't need any assumptions on survival distribution. It was a design that permits wide use. The limitation was that we tried to evaluate whether a few CI's may perform well in the analysis of PCCC data, by assuming that mixture Weibull distribution provides a good fit.

In the future, several further studies can be done. First, we can prove why the estimated timing had a cubic root convergence rate in the future.

Second, we will investigate whether the block size choosing method can be improved. As we can see from the real dataset, block size choosing indeed helped us achieve narrower confidence intervals, but from simulation, the cost was lowering its coverage. However, simulation with given parameters showed the confidence intervals after choosing block size can also have coverage which is close to the nominal level. But how to connect these simulations and real datasets and build a confidence interval with good coverage and narrow length is an exciting question.

Besides, we can search for better distribution to mimic the CHD dataset or a better way to fit the parameters of mixture Weibull distribution. Because although our curve of parametric estimated mortality function was very similar to the nonparametric estimated mortality function, the parametric estimated timing was a little bit different from the nonparametric one, and it was always larger by using weibullRMM_SEM to fit the parameters.

Reference

1. Abrevaya, J., & Huang, J. (2005). On the bootstrap of the maximum score estimator. Econometrica, 73(4), 1175-1204.
2. Banerjee, M., Durot, C., & Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. Annals of Statistics, 47(2), 720-757.
3. Bordes, L., & Chauveau, D. (2016). Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. Computational Statistics, 31(4), 1513-1538.
4. Cattaneo, M. D., Jansson, M., & Nagasawa, K. (2020). Bootstrap-Based Inference for Cube Root Asymptotics. Econometrica, 88(5), 2203-2219.
5. Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26(4), 404-413.
6. Erikssen, G., Liestøl, K., Seem, E., Birkeland, S., Saatvedt, K. J., Hoel, T. N., ... & Lindberg, H. L. (2015). Achievements in congenital heart defect surgery: a prospective, 40-year study of 7038 patients. Circulation, 131(4), 337-346.
7. Greenwood, M. (1926). The natural duration of cancer (report on public health and medical subjects no 33). London: Stationery Office.
8. Jacobs, J. P., Mayer Jr, J. E., Pasquali, S. K., Hill, K. D., Overman, D. M., Louis, J. D. S., ... & Jacobs, M. L. (2018). The society of thoracic surgeons congenital heart surgery database: 2018 update on outcomes and quality. The Annals of thoracic surgery, 105(3), 680-689.
9. Kempny, A., Dimopoulos, K., Uebing, A., Diller, G. P., Rosendahl, U., Belitsis, G., ... & Wort, S. J. (2017). Outcome of cardiac surgery in patients with congenital heart disease in England between 1997 and 2015. PLoS One, 12(6), e0178963.
10. Kim, J., & Pollard, D. (1990). Cube root asymptotics. The Annals of Statistics, 18(1), 191-219.
11. Larsen, S. H., Olsen, M., Emmertsen, K., & Hjortdal, V. E. (2017). Interventional treatment of patients with congenital heart disease: nationwide Danish experience over 39 years. Journal of the American College of Cardiology, 69(22), 2725-2732.
12. Lee, S. M. S., & Pun, M. C. (2006). On m out of n bootstrapping for nonstandard M-estimation with nuisance parameters. Journal of the American Statistical Association, 101(475), 1185-1197.
13. Lo, S. H., & Singh, K. (1986). The product-limit estimator and the bootstrap: some asymptotic representations. Probability Theory and Related Fields, 71(3), 455-465.
14. Mendis, S., Puska, P., Norrving, B., & World Health Organization. (2011). Global atlas on cardiovascular disease prevention and control. World Health Organization.
15. Nieminen, H. P., Jokinen, E. V., & Sairanen, H. I. (2001). Late results of pediatric cardiac surgery in Finland: a population-based study with 96% follow-up. Circulation, 104(5), 570-575.
16. Ou, F. S., Zeng, D., & Cai, J. (2016). Quantile regression models for current status data. Journal of statistical planning and inference, 178, 112-127.
17. Politis, D. N., Romano, J. P., & Wolf, M. (1999). Subsampling. Springer Science & Business Media.
18. Raissadati, A., Nieminen, H., Jokinen, E., & Sairanen, H. (2015). Progress in late results

among pediatric cardiac surgery patients: a population-based 6-decade study with 98% follow-up. Circulation, 131(4), 347-353.

19. Shi, C., Lu, W., & Song, R. (2018). A massive data framework for m-estimators with cubic-rate. Journal of the American Statistical Association, 113(524), 1698-1709.
20. Spector, L. G., Menk, J. S., Vinocur, J. M., Oster, M. E., Harvey, B. A., St. Louis, J. D., ... & Kochilas, L. K. (2016). In-hospital vital status and heart transplants after intervention for congenital heart disease in the Pediatric Cardiac Care Consortium: completeness of ascertainment using the National Death Index and United Network for Organ Sharing Datasets. Journal of the American Heart Association, 5(8), e003783.
21. Spector, L. G., Menk, J. S., Knight, J. H., McCracken, C., Thomas, A. S., Vinocur, J. M., ... & Kochilas, L. (2018). Trends in long-term mortality after congenital heart surgery. Journal of the American College of Cardiology, 71(21), 2434-2446.
22. Vinocur, J. M., Menk, J. S., Connett, J., Moller, J. H., & Kochilas, L. K. (2013). Surgical volume and center effects on early mortality after pediatric cardiac surgery: 25-year North American experience from a multi-institutional registry. Pediatric cardiology, 34(5), 1226-1236.
23. Zimmerman, M. S., Smith, A. G. C., Sable, C. A., Echko, M. M., Wilner, L. B., Olsen, H. E., ... & Kassebaum, N. J. (2020). Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Child & Adolescent Health, 4(3), 185-200.

Appendix

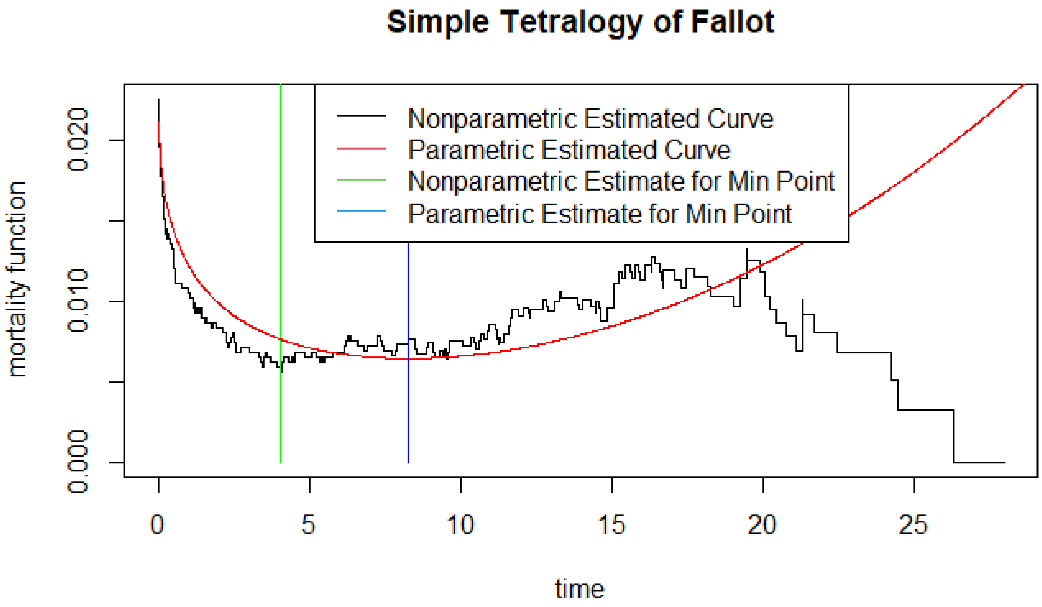

Figure 1: Parametric estimated mortality function and Nonparametric estimated mortality function for Simple Tetralogy of Fallot Dataset

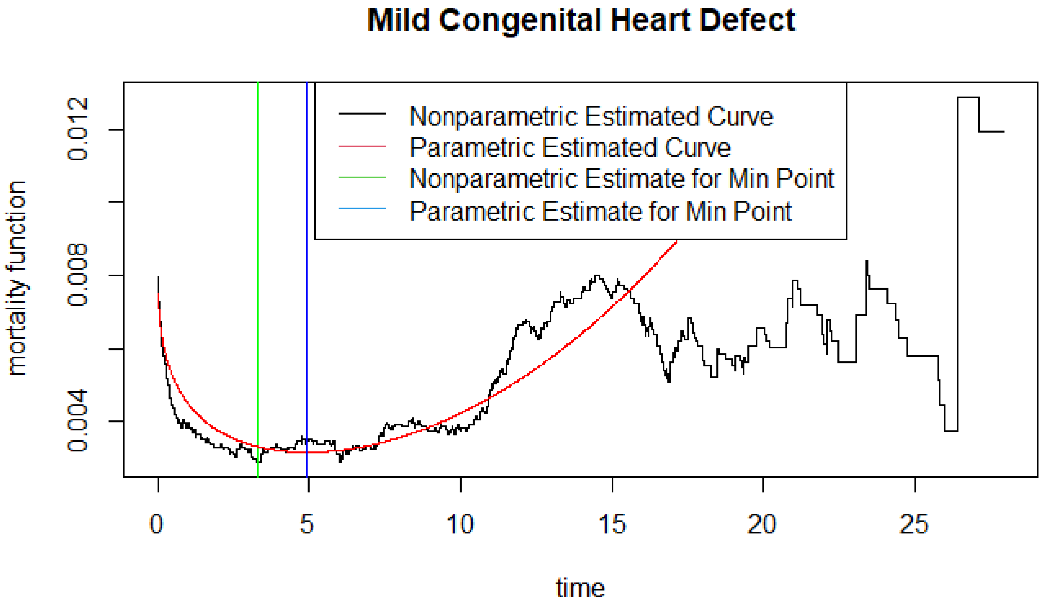

Figure 2: Parametric estimated mortality function and Nonparametric estimated mortality function for Mild Congenital Heart Defect Dataset

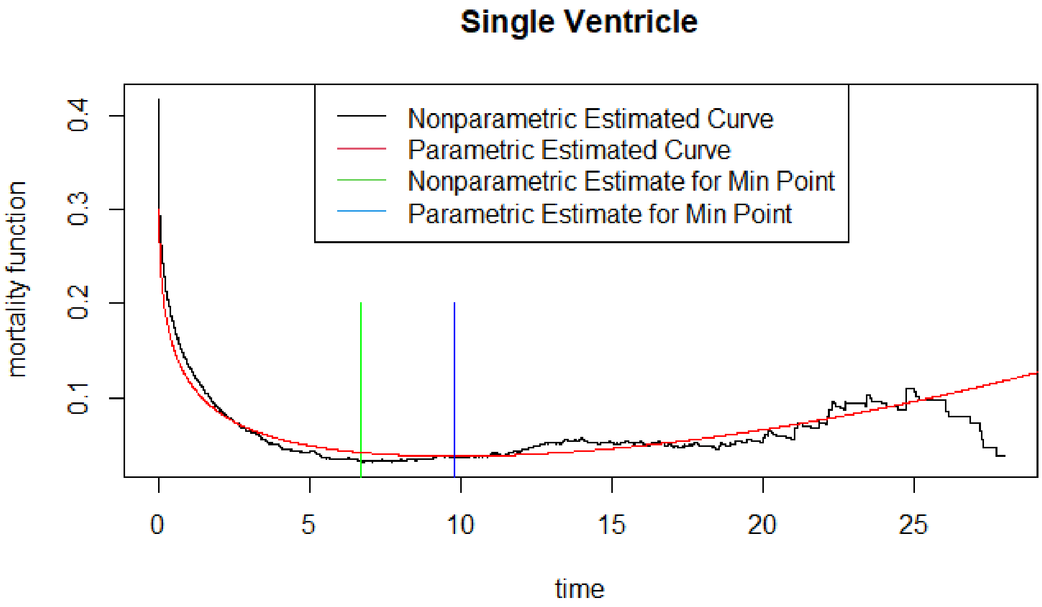

Figure 3: Parametric estimated mortality function and Nonparametric estimated mortality function for Single Ventricle Dataset