Generalized redistribute-to-the-right algorithm: application to the analysis of censored cost data

Hongwei Zhao, ScD
Professor
Department of Epidemiology and Biostatistics
School of Rural Public Health
Texas A&M Health Science Center

and

Shuai Chen,
Department of Statistics,
Texas A&M University

Outline of Topic

- Introduction and Background
- The Redistribute-to-the-Right (RR) Algorithm and the Kaplan-Meier Estimator
- Application of RR to the Estimation of the Survival Function of Costs With Censored Data
- Simulation Studies
- Example
- Conclusions and Discussions

Introduction and Background

- Rising health care costs are are a big concern to the public and policy makers.
- Sky-rocketing costs are associated with new medical interventions, chronic diseases, among others.
- Limited resources are available.
- Economic evaluations of new treatment options become more and more common.
- The goal is to maximize the potential use of resource and optimize health benefits.

Mean and Median Costs Estimation

- The mean costs is usually preferred, since it is associated with the total costs
- The mean is heavily influenced by extreme outliers, which are common in cost data
- The median value represents a central cost value, less influenced by extreme values
- Medians and quantiles have been used for describing money related outcomes, such as housing price, salary
- In order to estimate the median or quantiles of costs, one needs to estimate the survival function of costs

Censoring Problem

- Many economic studies are conducted alongside clinical trials
- Data collection are terminated at the end of the study – survival time and costs are censored
- Numerous well established statistical methodologies and algorithms exist for analyzing censored survival data (e.g., Kaplan-Meier (KM) estimator, Log-rank test, Cox proportional hazards regression).
- However, these methods cannot be used on the medical cost data due to the induced "informative censoring" problem.

Graphical illustration of informative censoring (Lin, 2003)

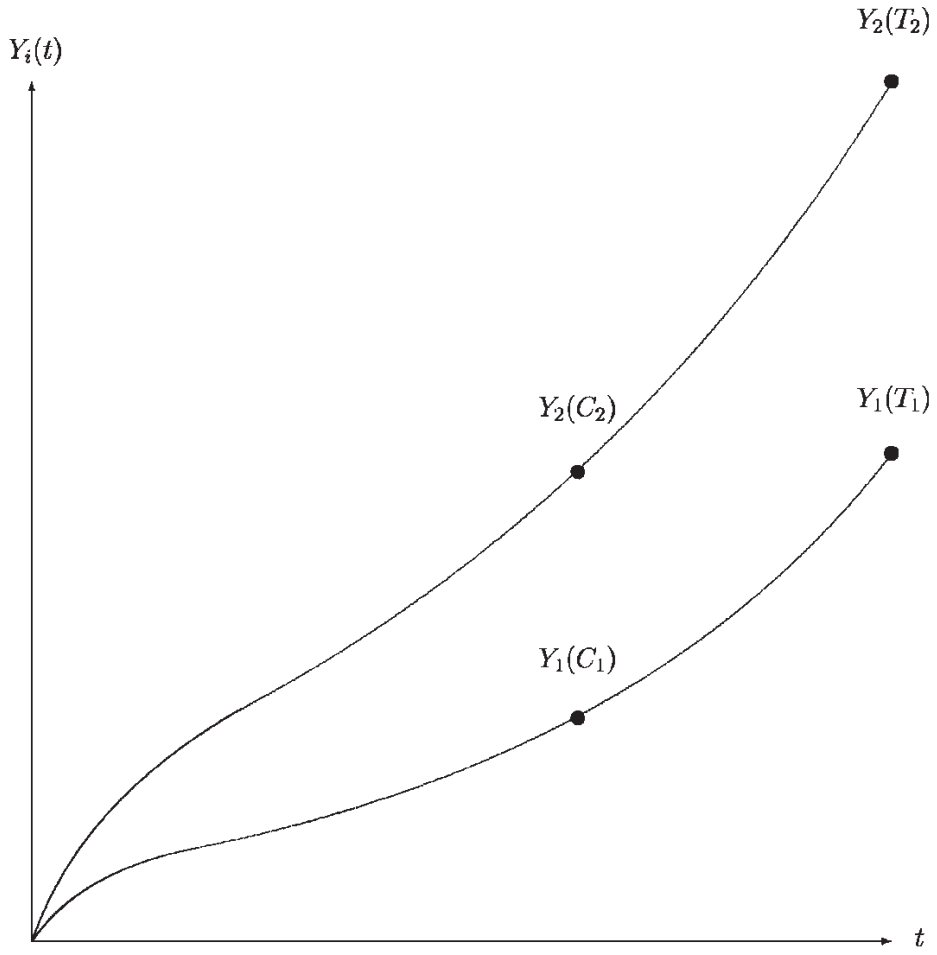


Figure 2. Cumulative costs at survival times and censoring times for two subjects in a heterogeneous population: subject 2 accumulates costs at higher rates than subject 1.

Notations and Assumptions

- T_i: survival time to some events (eg., death)
  C_i: censoring time
  $X_i = \min(T_i, C_i)$: follow-up time
  $\Delta_i = I(T_i \leq C_i)$: indicator of death
  $M_i(t)$: cost accumulated up to time t
  $M_i = M_i(X_i)$: observed total costs
- We can observe the follow-up time X_i, the death indicator Δ_i, and the total cost M_i.
- Cost history $M_i(t)$ $(t \leq X_i)$ may also be available.
- We only focus on the accumulated cost by a time limit L (L is chosen such that a reasonable amount of information is still available at that time)
- Our goal is to estimate the survival function of costs accumulated over time L: $S(x) = Pr\{M(L) > x\}$

The RR Algorithm and the Kaplan-Meier Estimator

- Efron (1967) proposed an RR algorithm which can be used to explain the Kaplan-Meier estimator for survival time.

The RR Algorithm and the Kaplan-Meier Estimator

- Efron (1967) proposed an RR algorithm which can be used to explain the Kaplan-Meier estimator for survival time.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

The RR Algorithm and the Kaplan-Meier Estimator

- Efron (1967) proposed an RR algorithm which can be used to explain the Kaplan-Meier estimator for survival time.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
Step 1:	$\frac{1}{5}$		$\frac{4}{15}(=\frac{1}{5}+\frac{1}{3}\cdot\frac{1}{5})$	$\frac{4}{15}$	$\frac{4}{15}$

The RR Algorithm and the Kaplan-Meier Estimator

- Efron (1967) proposed an RR algorithm which can be used to explain the Kaplan-Meier estimator for survival time.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
Step 1:	$\frac{1}{5}$		$\frac{4}{15}(=\frac{1}{5}+\frac{1}{3}\cdot\frac{1}{5})$	$\frac{4}{15}$	$\frac{4}{15}$
Step 2:	$\frac{1}{5}$		$\frac{4}{15}$		$\frac{8}{15}(=\frac{4}{15}+\frac{4}{15})$

The RR Algorithm and the Kaplan-Meier Estimator

- Efron (1967) proposed an RR algorithm which can be used to explain the Kaplan-Meier estimator for survival time.

$X_i =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
Step 1:	$\frac{1}{5}$		$\frac{4}{15}(=\frac{1}{5}+\frac{1}{3}\cdot\frac{1}{5})$	$\frac{4}{15}$	$\frac{4}{15}$
Step 2:	$\frac{1}{5}$		$\frac{4}{15}$		$\frac{8}{15}(=\frac{4}{15}+\frac{4}{15})$
$\hat{S}^T_{RR}(t) =$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{8}{15}$	$\frac{8}{15}$	0

Then $\hat{S}^T_{RR}(t)$ is the same as the K-M estimator

$$\hat{S}^T_{KM}(t) = \prod_{i:\tau_i \le t} (1 - \frac{d_i}{R_i}).$$

- What does the RR algorithm tell us about the Kaplan-Meier estimator for the survival time?
  - Due to the assumption of independent censoring, a censored time is equally likely to be any observed time larger than itself.
  - The censored time can be replaced by the larger observed times with equal probability, which can be considered as a special form of imputation.
  - The survival function cannot be estimated after the largest event time.

Application of RR to the Estimation of the Survival Function of Costs With Censored Data

- To estimate the survival function of costs $S(x) = Pr(M > x)$, a simple weighted (SW) estimator (Zhao and Tsiatis, 1997) can be formed:

$$\hat{S}_{SW}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(T_i)} I(M_i > x).$$

- The large sample property of this estimator, such as its consistency and asymptotic normality has been established for the outcome of quality adjusted lifetime, but can be applied to the problem of cost estimation.

- One effective way to improve efficiency is to redefine the endpoint for each person, adapting the approach used in the estimation of quality-adjusted life time (Zhao and Tsiatis, 1997).

- For a fixed x, if M_i exceeds x, then this would be known at any time s such that $s \geq s_i(x)$, where $s_i(x) = \inf[s : M_i(s) \geq x]$.

- Redefine $T_i^*(x) = \min\{T_i, s_i(x)\}$, $X_i^*(x) = \min\{T_i^*(x), C_i\}$ and $\Delta_i^*(x) = I(T_i^*(x) \leq C_i)$, we have

$$\widehat{S}_{ZT}(x) = n^{-1} \sum_{i=1}^{n} \frac{\Delta_i^*(x)}{\hat{K}^*\{T_i^*(x)\}} I(M_i > x),$$

  where $\hat{K}^*\{T_i^*(x)\}$ is the Kaplan-Meier estimator for the survival function of the censoring time variable C, evaluated at $T_i^*(x)$, based on data $\{X_i^*(x), \Delta_i^*(x),\ i = 1, \cdots, n\}$.

- This estimator is usually more efficient than the SW estimator, but it cannot be guaranteed to be monotone (Huang, 1998).

A Consistent RR Survival Estimator

- In order to use the RR algorithm for censored costs estimation, we need to find, for a censored observation i, the contributions to its cost from the observations to its right.
- Denote the contribution from the j observation ($X_j > X_i$) as $W_j^{(i)}$.
- Example to get $W_j^{(2)}$ using the RR algorithm:

$X_j =$	1	2	3	4	5
	x	o	x	o	x

A Consistent RR Survival Estimator

- In order to use the RR algorithm for censored costs estimation, we need to find, for a censored observation i, the contributions to its cost from the observations to its right.
- Denote the contribution from the j observation ($X_j > X_i$) as $W_j^{(i)}$.
- Example to get $W_j^{(2)}$ using the RR algorithm:

$X_j =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	0	1	0	0	0

A Consistent RR Survival Estimator

- In order to use the RR algorithm for censored costs estimation, we need to find, for a censored observation i, the contributions to its cost from the observations to its right.
- Denote the contribution from the j observation ($X_j > X_i$) as $W_j^{(i)}$.
- Example to get $W_j^{(2)}$ using the RR algorithm:

$X_j =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	0	1	0	0	0
Step 1:	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

A Consistent RR Survival Estimator

- In order to use the RR algorithm for censored costs estimation, we need to find, for a censored observation i, the contributions to its cost from the observations to its right.
- Denote the contribution from the j observation ($X_j > X_i$) as $W_j^{(i)}$.
- Example to get $W_j^{(2)}$ using the RR algorithm:

$X_j =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	0	1	0	0	0
Step 1:	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
Step 2:	0	0	$\frac{1}{3}$	0	$\frac{2}{3}(=\frac{1}{3}+\frac{1}{3})$

A Consistent RR Survival Estimator

- In order to use the RR algorithm for censored costs estimation, we need to find, for a censored observation i, the contributions to its cost from the observations to its right.
- Denote the contribution from the j observation ($X_j > X_i$) as $W_j^{(i)}$.
- Example to get $W_j^{(2)}$ using the RR algorithm:

$X_j =$	1	2	3	4	5
	x	o	x	o	x
Step 0:	0	1	0	0	0
Step 1:	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
Step 2:	0	0	$\frac{1}{3}$	0	$\frac{2}{3}(=\frac{1}{3}+\frac{1}{3})$
$W_j^{(2)}$:			$\frac{1}{3}$		$\frac{2}{3}$

- For the censored observation i, we use the weighted sum

$$I(M_i > x)^{RR} = \sum_{j=1}^{n} \Delta_j I(X_j > X_i) W_j^{(i)} I(M_j > x)$$

as its replacement indicator.

- The RR survival estimator for the costs is

$$\hat{S}_{RR}(x) = \frac{1}{n} \sum_{i=1}^{n} \{\Delta_i I(M_i > x) + (1 - \Delta_i) I(M_i > x)^{RR}\}.$$

- We can show that this RR^S estimator is mathematically equivalent to the simple weighted estimator $\hat{S}_{SW}(x)$.

- Remarks
  - The weight $W_j^{(i)}$ is actually the conditional probability of an event occurring at X_j given that the subject is alive at X_i (discrete case), which can be easily obtained from Kaplan-Meier estimators:

$$W_j^{(i)} = \hat{P}(T = T_j | T \geq C_i) = \frac{\hat{S}^T(T_j^-) - \hat{S}^T(T_j)}{\hat{S}^T(C_i)} = \frac{1}{n\hat{S}^T(C_i)\hat{K}(T_j)}.$$

  - This estimator does not use the costs from the censored observations, therefore it is not very efficient. We will propose an estimator that will use observed censored costs.

An Improved RR Survival Estimator

- For the censored observation i, we consider the weighted sum

$$I(M_i > x)^{RRimp} = \sum_{j=1}^{n} \Delta_j I(T_j > X_i) W_j^{(i)} I(M_j^{(i)} > x),$$

  where $M_j^{(i)} = M_i(C_i) + M_j - M_j(C_i)$ is the sum of the observed costs for the censored observation i and the additional cost from time X_i to X_j for the observation j.

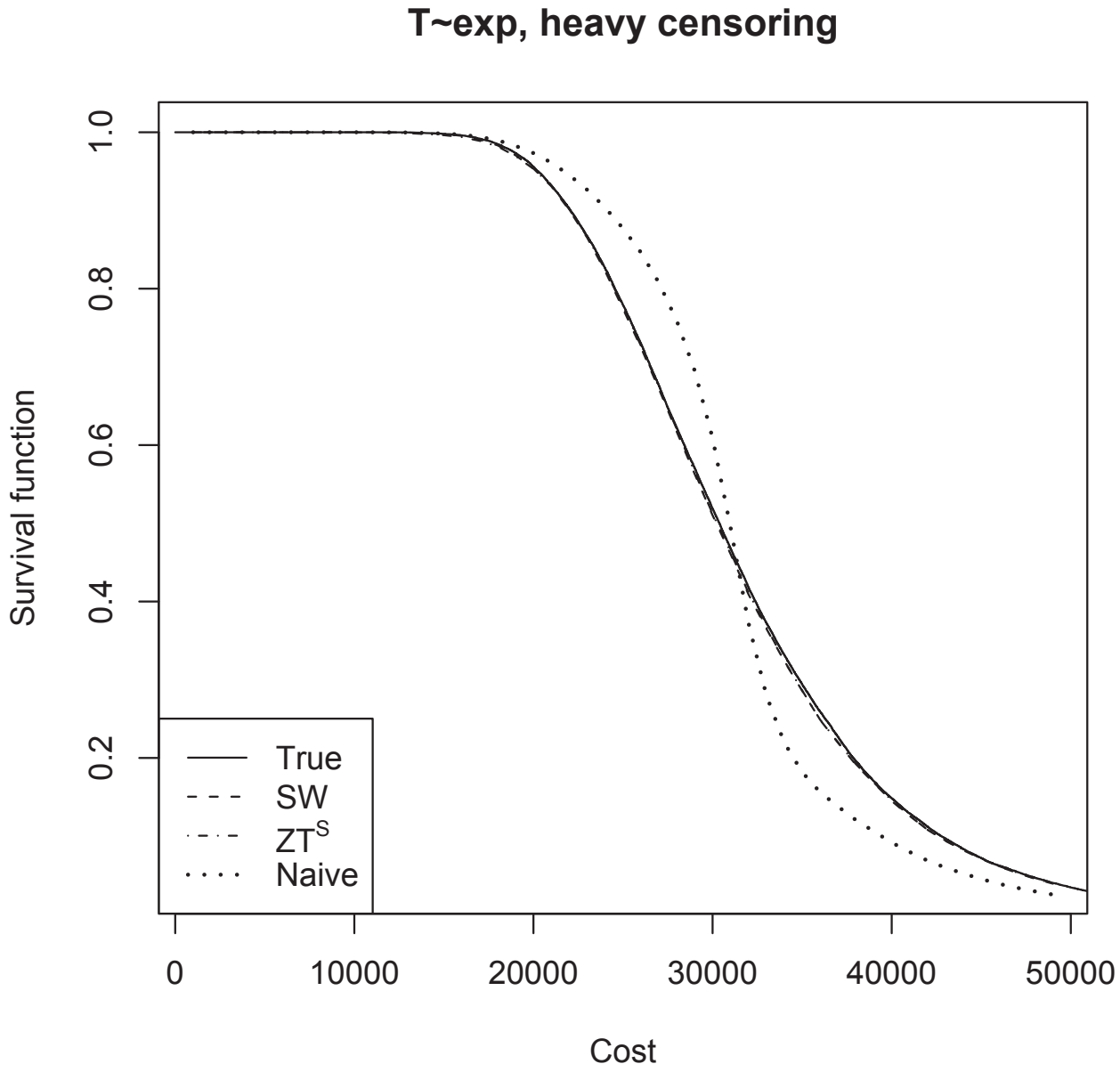
- We propose an improved RR survival estimator:

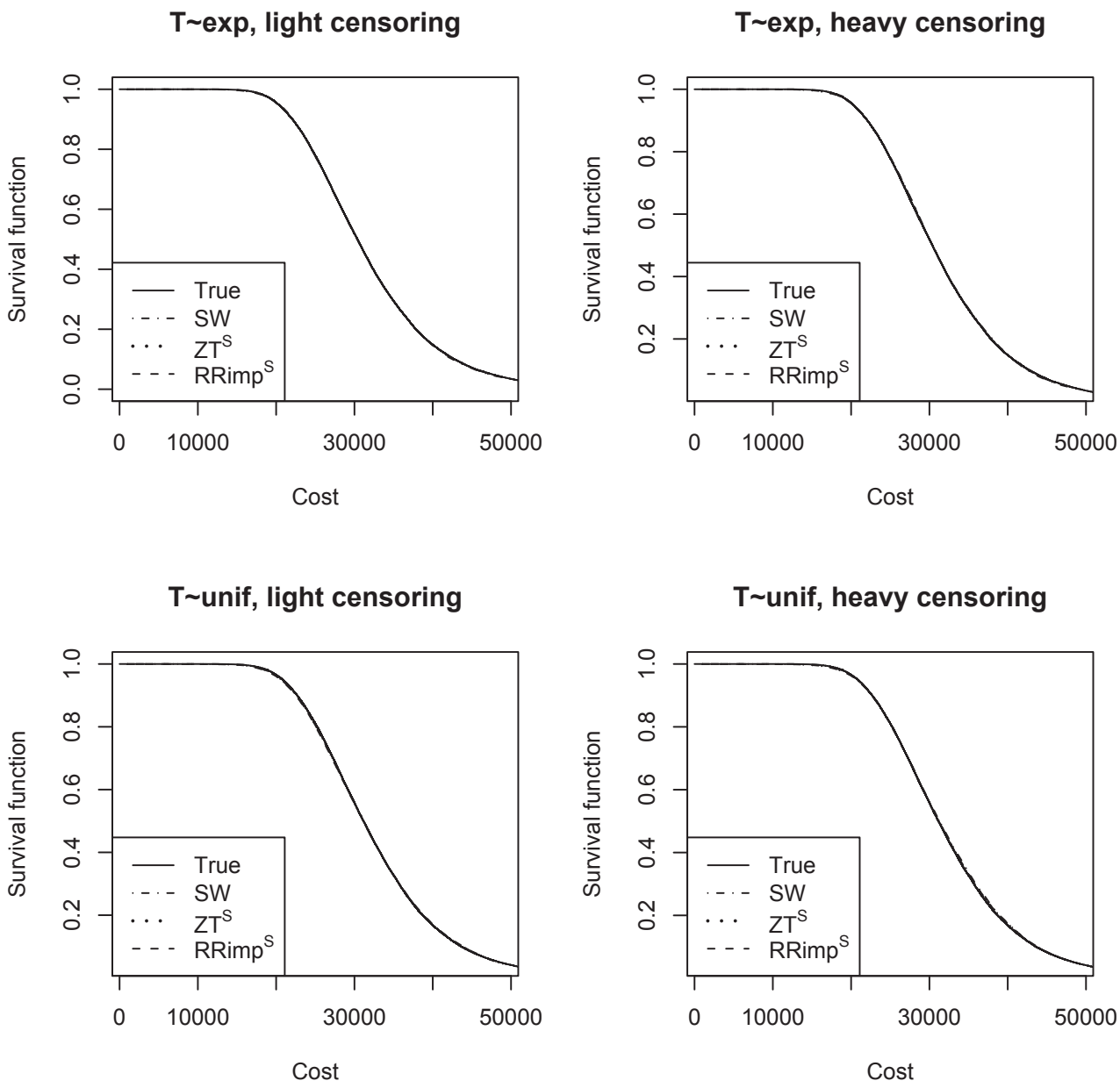
$$\hat{S}_{RRimp}(x) = \frac{1}{n} \sum_{i=1}^{n} [\Delta_i I(M_i > x) + (1 - \Delta_i) I(M_i > x)^{RRimp}]$$

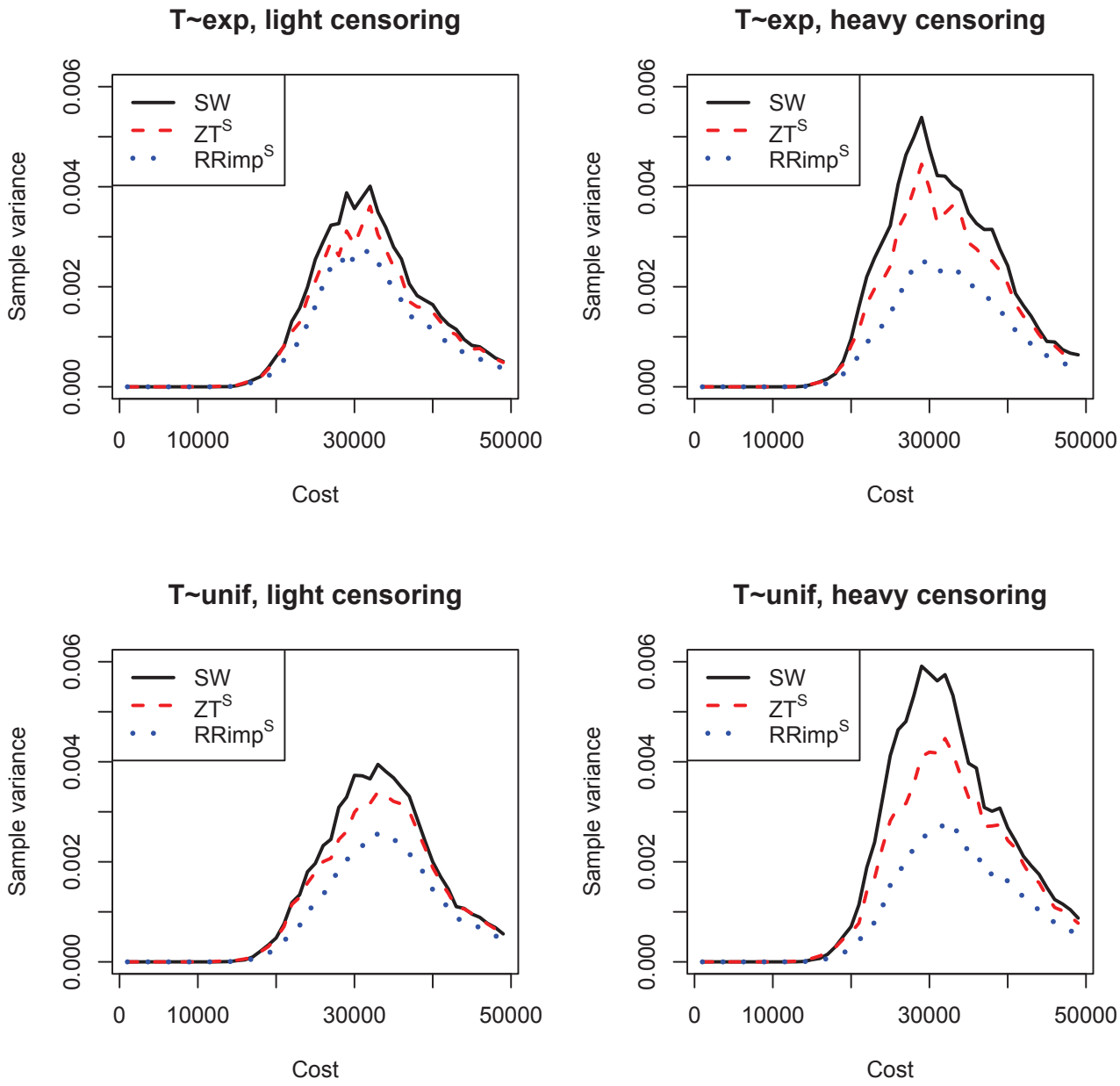
- Unfortunately, $\hat{S}_{RRimp}(x)$ is not always consistent. However, it is monotone, more efficient, and the bias is quite small as seen from the simulation studies.

Simulation Studies

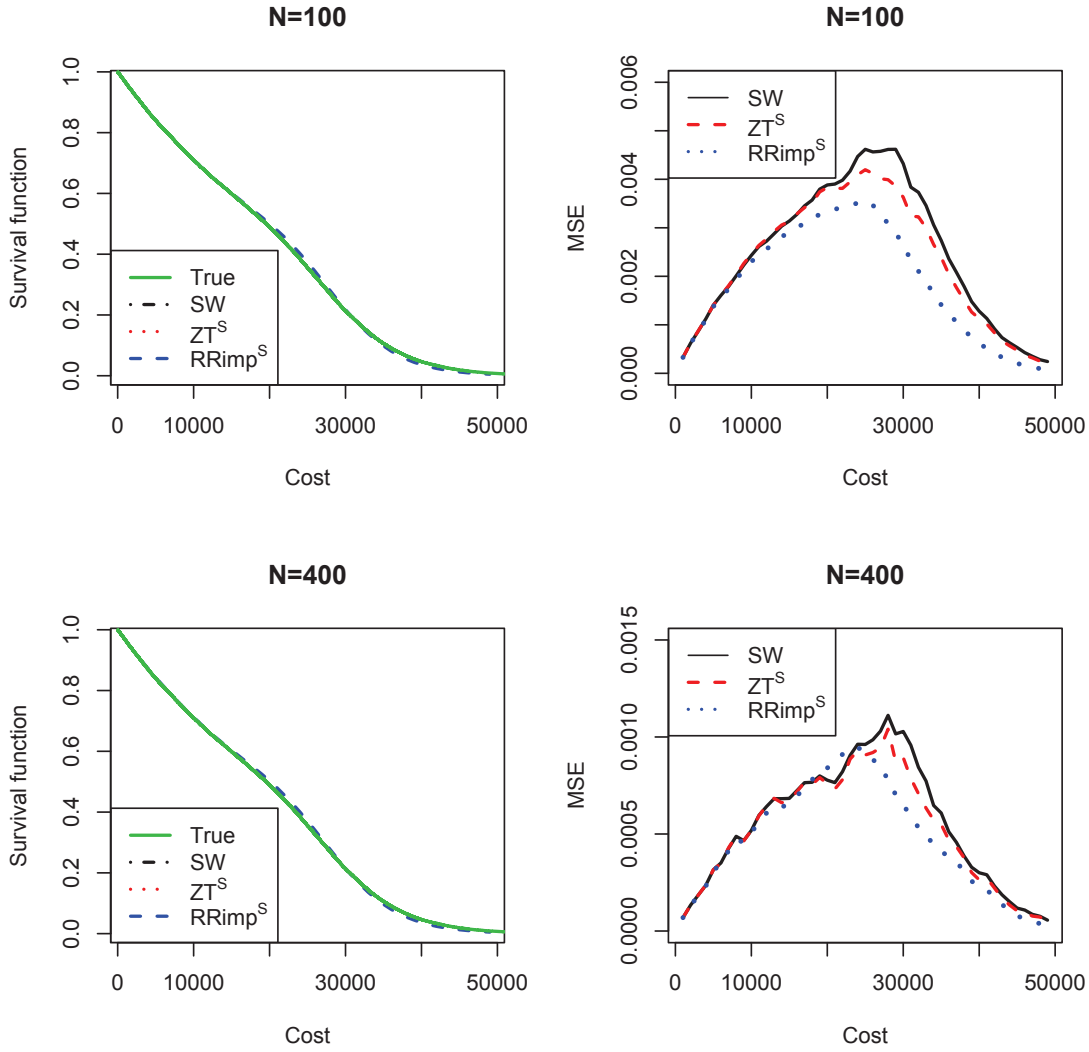
- Survival time $T \sim exp(10)$ or $T \sim Unif(0, 15)$, and truncated at L=10.
- Censoring time $C \sim Unif(0, 22)$ for light censoring, or $C \sim Unif(0, 15)$ for heavy censoring.
- Diagnostic costs$\sim Lnorm(10, 0.245^2)$
  Fixed annual costs (fixed every year)$\sim Lnorm(6, 0.245^2)$
  Random annual costs$\sim Lnorm(4, 0.245^2)$
  Terminal costs$\sim Lnorm(9, 0.632^2)$
- Sample size=100, number of simulations=1000.

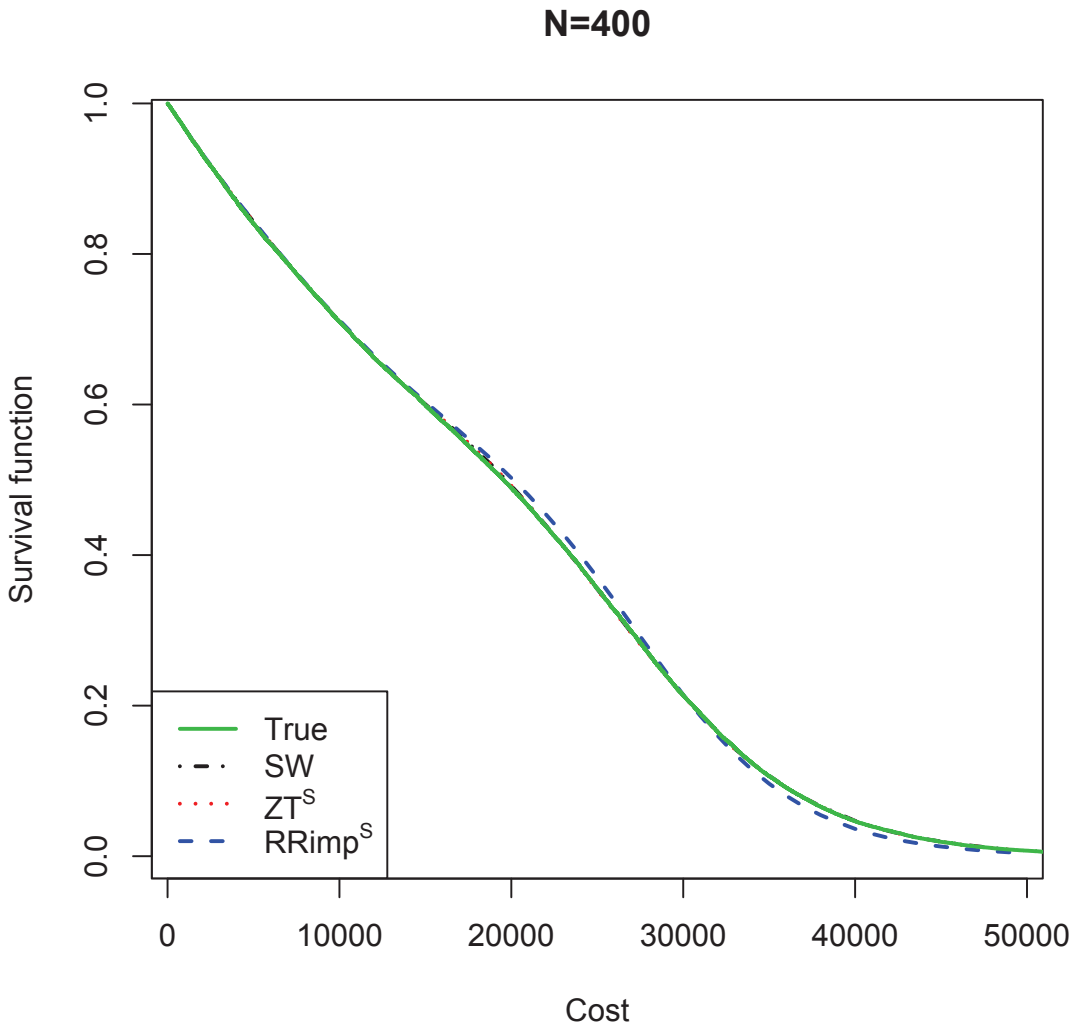

T~exp, heavy censoring
Survival function
1.0
0.8
0.6
0.4
0.2
True
SW
ZT^S
Naive
0
10000
20000
30000
40000
50000
Cost


T~exp, light censoring
T~exp, heavy censoring
T~unif, light censoring
T~unif, heavy censoring
Survival function
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
30000
50000
Cost
True
SW
ZT^S
RRimp^S


T~exp, light censoring
T~exp, heavy censoring
T~unif, light censoring
T~unif, heavy censoring
SW
ZT^S
RRimp^S
Sample variance
0.000
0.002
0.004
0.006
0
10000
30000
50000
Cost

Extreme Case: Fixed$\sim$lnorm(8, 0.245^2); other costs=0; T$\sim$exp, heavy censoring.



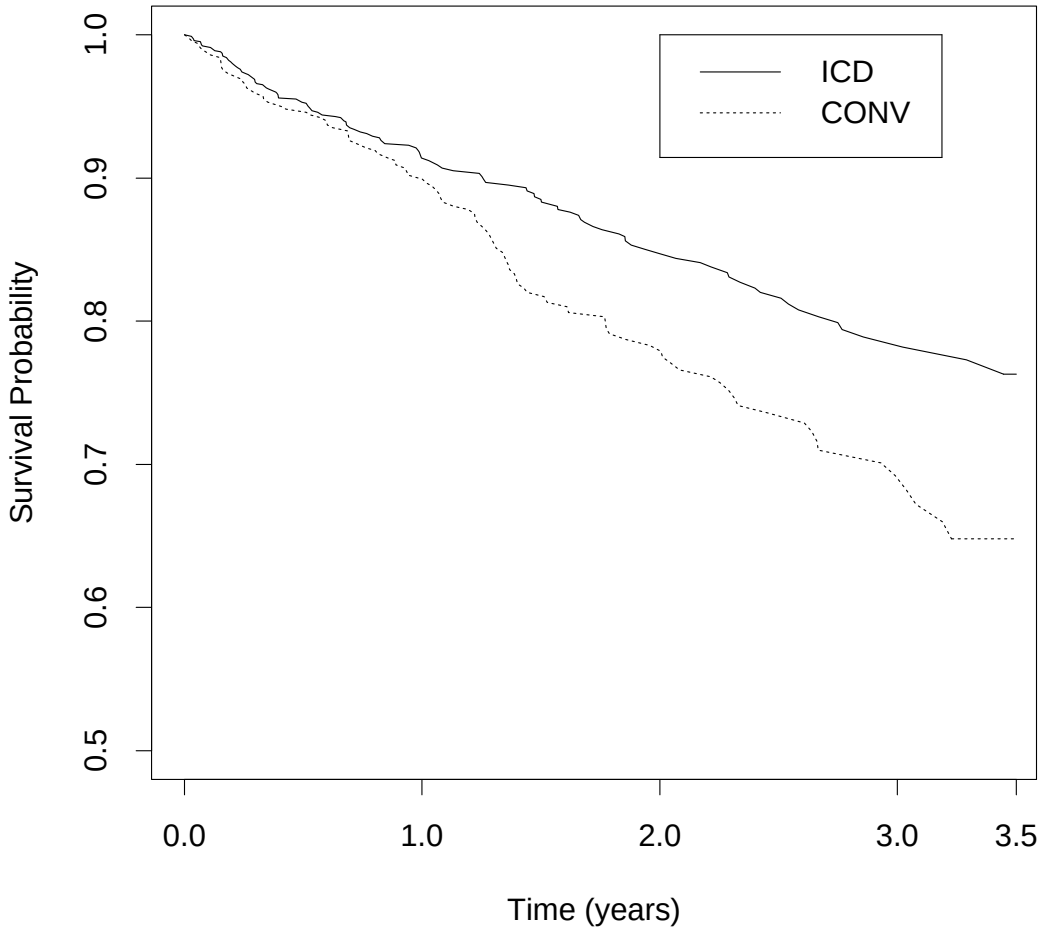

N=400
Survival function
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
50000
Cost
True
SW
ZT^S
RRimp^S

MADIT II Data Example

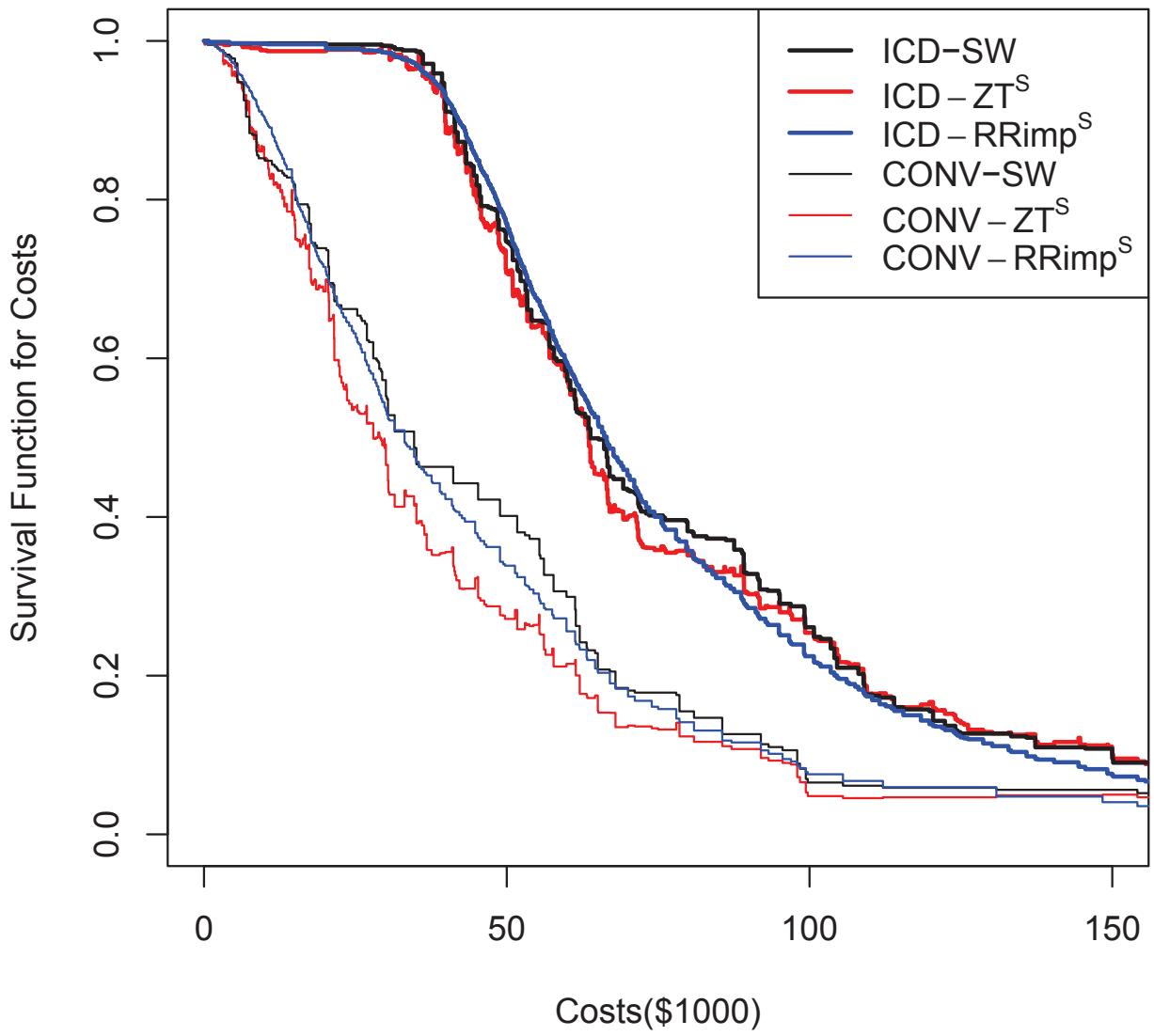
The Multicenter Automatic Defibrillator Implantation Trial -II.

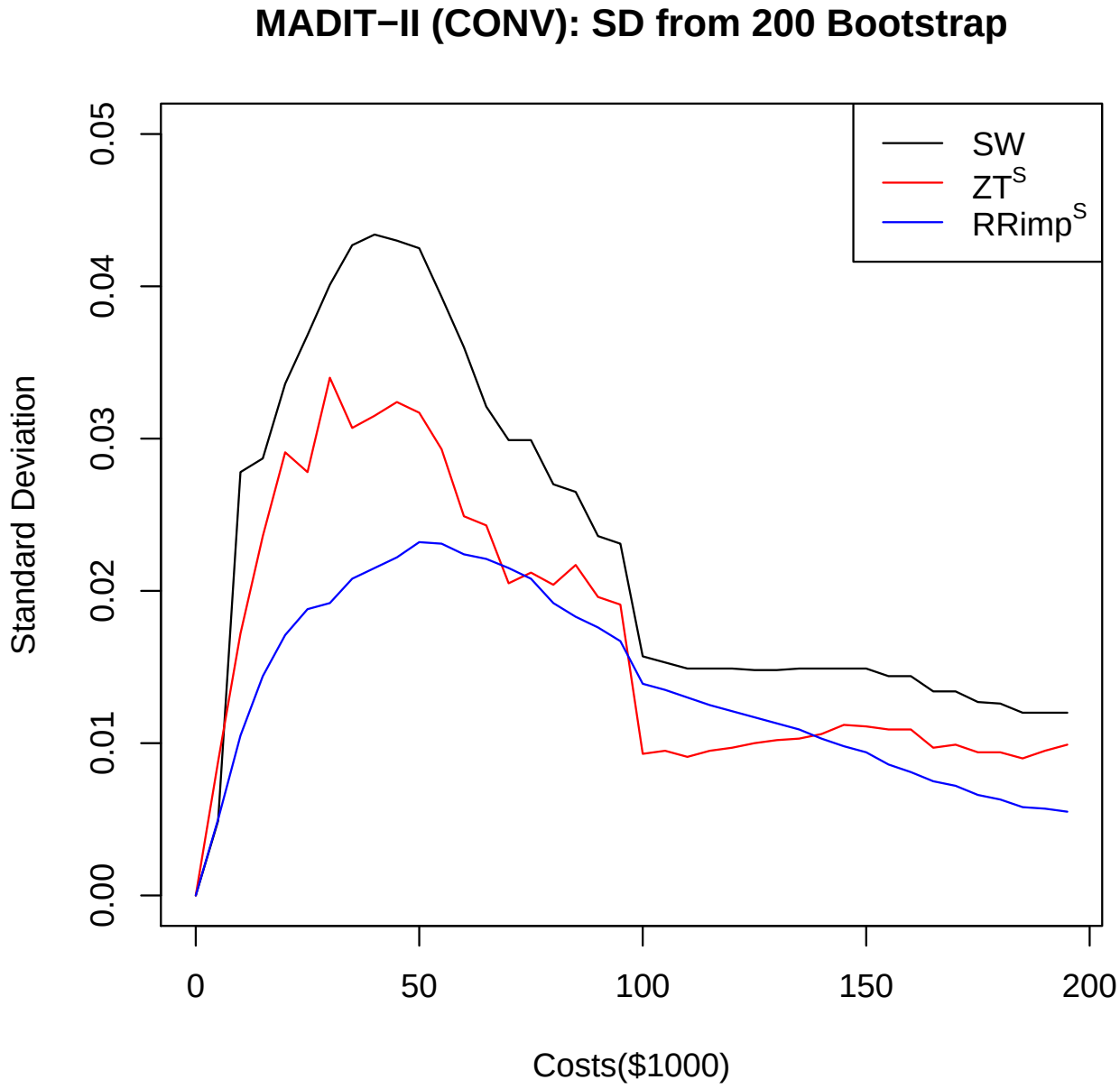
- A fully sequential, randomized clinical trial.
- Primary objective: To evaluate the potential survival benefit of a prophylactically implanted defibrillator (ICD) vs. conventional therapy (CONV).
- Population: Patients with a prior myocardial infarction and advanced left ventricular dysfunction.
- There were 664 (431) patients in the ICD (CONV) arm.
- Follow-up: 11 days to 55 months, averaging 22 months.
- After the trial was completed, it was found that the ICD arm had a survival advantage with an estimated hazard ratio of 0.69 (95% confidence interval, 0.51 to 0.93; p=0.016).

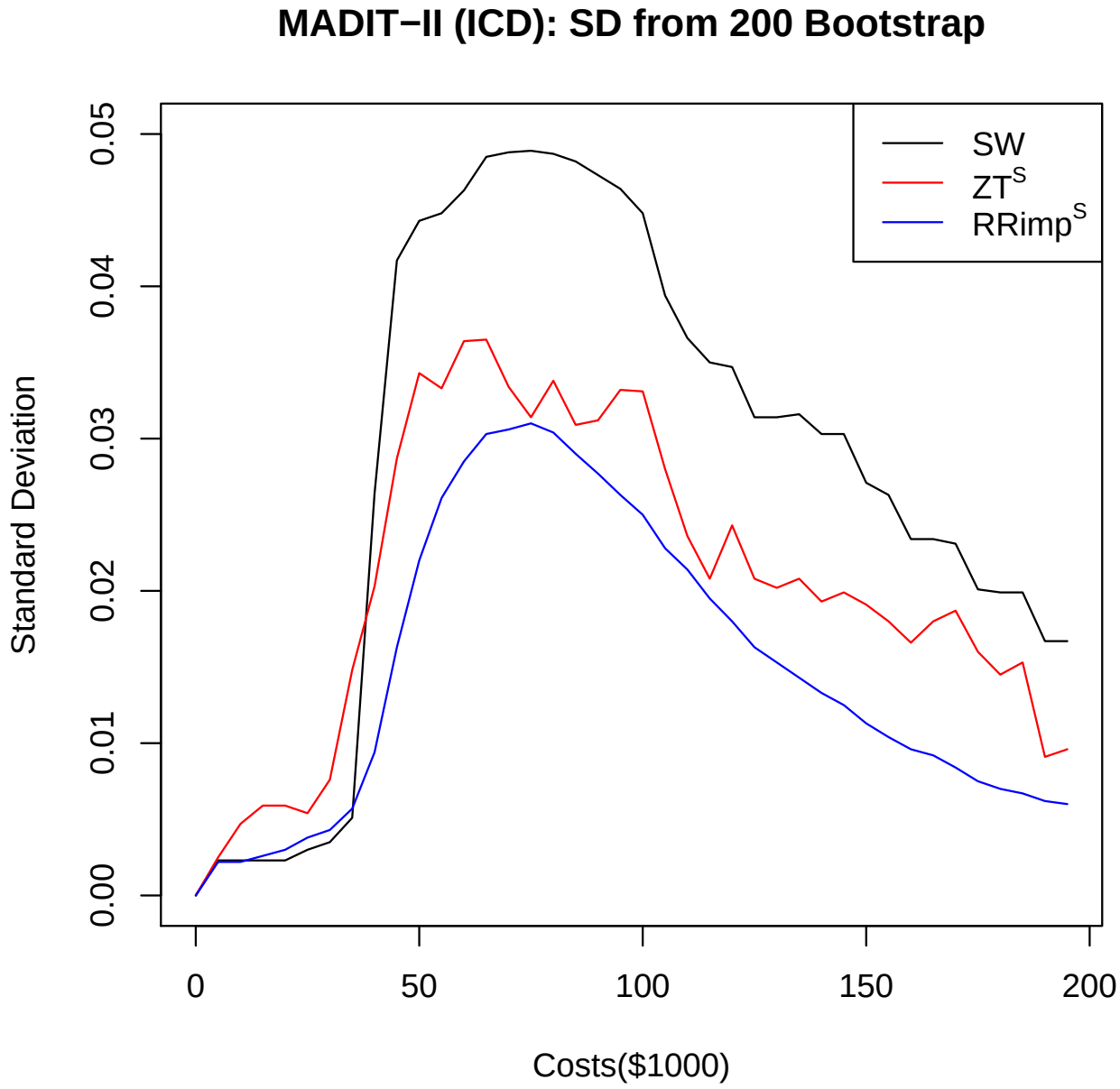
Kaplan-Meier survival plot for the two treatment arms in MADIT II.



- The survival benefit of the implanted cardiac defibrillator was established.
- However, the associated costs of the defibrillators are also high.
- It could have an enormous economical impact, since 55,000-400,000 people in the US annually fit the criteria, and may therefore, benefit from the intervention.
- Secondary Analysis: To evaluate the costs and cost-effectiveness of the ICD as compared to conventional therapy.
- Patients were asked every month on the utilization of hospitalizations, emergency room visits, physician visits, outpatient diagnostic tests and procedures, outpatient surgeries, nursing home visits, and prescription medications.
- Missing data were imputed with multiple imputation methods.
- Costs were discounted using a 3% annual discount rate.


Survival Function for Costs
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Costs($1000)
ICD−SW
ICD – ZT^S
ICD – RRimp^S
CONV−SW
CONV – ZT^S
CONV – RRimp^S


MADIT−II (CONV): SD from 200 Bootstrap
Standard Deviation
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
200
Costs($1000)
SW
ZT^S
RRimp^S


MADIT−II (ICD): SD from 200 Bootstrap
Standard Deviation
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
200
Costs($1000)
SW
ZT^S
RRimp^S

Conclusions and Discussions

- We provide a link between a theoretically derived simple-weighted estimator for the survival function of costs, and an intuitive RR survival estimator.
- We propose an improved RR survival estimator for costs, which is monotone and more efficient than the SW and the ZT^S estimators, but unfortunately, it is not always consistent.
- Further research needs to be conducted in order to find a monotone, consistent and efficient estimator.
- The generalized RR algorithm may also be considered for regression modeling of costs, and for studying other outcomes such as quality-adjusted survival time, repeated events, etc.

Acknowledgement

I would like to thank the following people for making this work possible:

- Dr. Arthur Moss, PI for MADIT-II trial, for providing data to us;
- Dr. Phillip Pfeifer for his original idea on the RR algorithm;
- Dr. Heejung Bang, for many helpful discussions;
- Shuai Chen, for mathematical derivations, and analyzing the MADIT-II example.

This research was supported by R01 HL096575 from the National Heart, Lung, and Blood Institute.

Reference

- Bang H, Tsiatis AA. (2000). Estimating medical costs with censored data. Biometrika 87: 329-343.
- Chen, S. and Zhao, H. "Generalized Redistribute-to-the-right Algorithm: Application to the Analysis of Censored Cost Data". 2012, *Journal of Statistical Theory and Practice*, in press.
- Lin, D. Y., Feuer, E. J., Etzioni, R. and Wax, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53**, 419-434.
- Zhao, H. and Tian, L. (2001). On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics* **57**, 1002-1008.

- Zhao, H, Bang, H., Wang, H. and Pfeifer, P.E. (2007). On the equivalence of some medical cost estimators with censored data". *Statistics in Medicine*, **26**, 4520-4530.
- Zhao, H., Cheng, Y. and Bang, H. (2011). Some insight on censored cost estimators. *Statistics In Medicine*, **30**:2381-2389
- Zhao, H. and Tsiatis, A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*, **84**:339-348.