SCT Invited Session 18, May 24, 2023, 10.30 am
Society for Clinical Trials Annual Meeting 2023

Thanks to Lu Mao for organizing this session

and

to our chair, Po-Kuei Chen

Extending the Proportional Odds Model to Multiple Event Time Data

David Oakes

Department of Biostatistics and Computational Biology
University of Rochester

david_oakes@urmc.rochester.edu

NO RELEVANT DISCLOSURES

1. INTRODUCTION

- Examine the "WLW" data presented in Wei, Lin and Weissfeld (1989) from a new perspective.
- Compare semi-parametric and parametric analyses of the WLW data odds model.
- Comment on the general specification and simulation of models for ordered event times.

2. THE BLADDER CANCER DATA

- The WLW data were abstracted from Andrews and Herzberg (1985). The accompanying commentary by D.P. Byar stated that data in other publications he lists "are not necessarily identical to those published here".
- Cook and Lawless (2006) indicate that the data are actually inspection times and perhaps should not be analyzed as incidence times. Moreover, multiple tumors were detected at each inspection.
- As stated by WLW, only data up to the fourth recurrence are abstracted.
- I ignore these issues and use the data as presented in WLW.

Table 2. Tumor Recurrence Data for Patients With Bladder Cancer

Treatment group	Follow-up time	Initial number	Initial size	Recurrence time 1	2	3	4	Treatment group	Follow-up time	Initial number	Initial size	Recurrence time 1	2	3	4
1	0	1	1					1	53	3	1	3	15	46	51
1	1	1	3					1	59	1	1				
1	4	2	1					1	61	3	2	2	15	24	30
1	7	1	1					1	64	1	3	5	14	19	27
1	10	5	1					1	64	2	3	2	8	12	13
1	10	4	1	6				2	1	1	3				
1	14	1	1					2	1	1	1				
1	18	1	1					2	5	8	1	5			
1	18	1	3	5				2	9	1	2				
1	18	1	1	12	16			2	10	1	1				
1	23	3	3					2	13	1	1				
1	23	1	3	10	15			2	14	2	6	3			
1	23	1	1	3	16	23		2	17	5	3	1	3	5	7
1	23	3	1	3	9	21		2	18	5	1				
1	24	2	3	7	10	16	24	2	18	1	3	17			
1	25	1	1	3	15	25		2	19	5	1	2			
1	26	1	2					2	21	1	1	17	19		
1	26	8	1	1				2	22	1	1				
1	26	1	4	2	26			2	25	1	3				
1	28	1	2	25				2	25	1	5				
1	29	1	4					2	25	1	1				
1	29	1	2					2	26	1	1	6	12	13	
1	29	4	1					2	27	1	1	6			
1	30	1	6	28	30			2	29	2	1	2			
1	30	1	5	2	17	22		2	36	8	3	26	35		
1	30	2	1	3	6	8	12	2	38	1	1				
1	31	1	3	12	15	24		2	39	1	1	22	23	27	32
1	32	1	2					2	39	6	1	4	16	23	27
1	34	2	1					2	40	3	1	24	26	29	40
1	36	2	1					2	41	3	2				
1	36	3	1	29				2	41	1	1				
1	37	1	2					2	43	1	1	1	27		
1	40	4	1	9	17	22	24	2	44	1	1				
1	40	5	1	16	19	23	29	2	44	6	1	2	20	23	27
1	41	1	2					2	45	1	2				
1	43	1	1	3				2	46	1	4	2			
1	43	2	6	6				2	46	1	4				
1	44	2	1	3	6	9		2	49	3	3				
1	45	1	1	9	11	20	26	2	50	1	1				
1	48	1	1	18				2	50	4	1	4	24	47	
1	49	1	3					2	54	3	4				
1	51	3	1	35				2	54	2	1	38			
1	53	1	7	17				2	59	1	3				

NOTE: Treatment group: 1, placebo; 2, thiotepa. Follow-up time and recurrence time are measured in months. Initial size is measured in centimeters. Initial number of 8 denotes eight or more initial tumors.

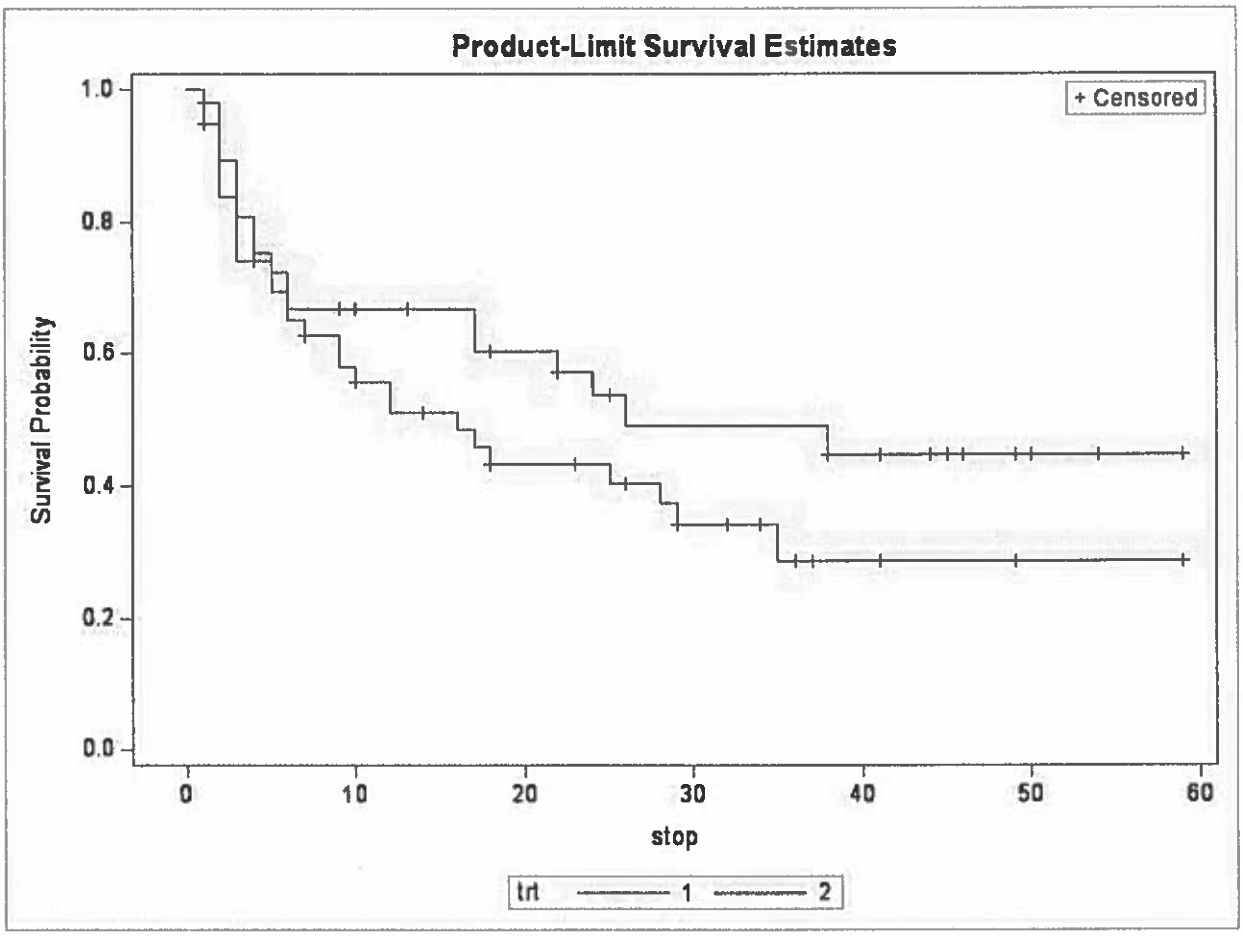
Source: Andrews and Herzberg (1985, pp. 254–259).

$c = (c_1, \ldots, c_K)' = (e'\hat{\Psi}^{-1}e)^{-1}\hat{\Psi}^{-1}e$, where $e = (1, \ldots, 1)'$, has the smallest asymptotic variance among all of the linear estimators (see Wei and Johnson 1985). Notice that, even if the η_k's are unequal, in practice one may still combine the $\hat{\eta}_k$'s to draw a conclusion about the "average effect" of the covariates provided that there are no qualitative differences among the η_k's.

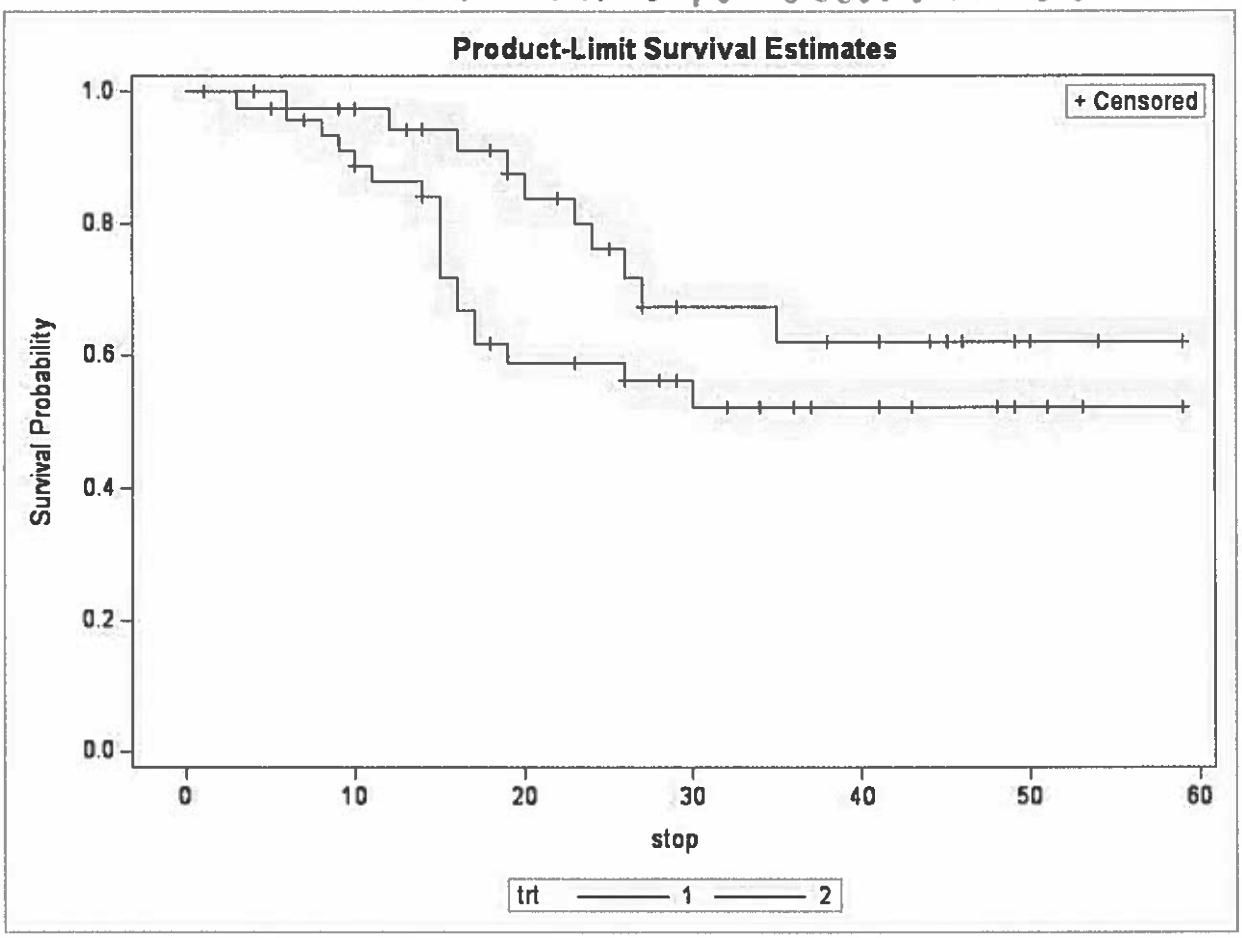
The longitudinal failure time data, such as those described in Section 1, provide us with the opportunity to study the changes of the effects η_k's over time. If we do not prespecify any relationship among the η_k's, we are faced with a multiple inference problem. For simplicity, let us assume that $\eta_k \leq 0$ for all k. If all of the null hypotheses H_k's are true, the standardized estimator $(\tilde{\eta}_1, \ldots, \tilde{\eta}_K)'$, where $\tilde{\eta}_k = \hat{\eta}_k/\hat{\psi}_{kk}^{1/2}$, is approximately normal with mean 0 and covariance matrix $\tilde{\Psi} = \{\tilde{\psi}_{kl}\}$, where $\tilde{\psi}_{kl} = \hat{\psi}_{kl}/(\hat{\psi}_{kk}\hat{\psi}_{ll})^{1/2}$ and $\hat{\psi}_{kl}$ is the (k, l)th element of $\hat{\Psi}$. $\tilde{\eta}_k < d$, where d is the largest value such that $\Pr(\tilde{\eta}_k \geq d, k = 1, \ldots, K \mid H_1, \ldots, H_K) \geq 1 - \alpha$ and α is a prespecified level of significance. The sequential multiple test procedures studied by Marcus, Peritz, and Gabriel (1976), Holm (1979), and Wei and Stram (1988), however, can be applied to the present case to obtain more powerful tests. Now, let $\tilde{\eta}_k^*$ be the kth smallest observed value of the $\tilde{\eta}_k$'s, and let $\tilde{\Psi}^*$ be the corresponding covariance matrix obtained by rearranging the rows and columns of $\tilde{\Psi}$. In addition, let $H_k^* : \eta_k^* = 0$ be the ordered hypotheses from the H_k's according to the order of $\tilde{\eta}_1^*, \ldots, \tilde{\eta}_K^*$,. Furthermore, let $(V_1, \ldots, V_K)'$ be a multivariate normal vector with mean 0 and covariance matrix $\tilde{\Psi}^*$. Starting with the hypothesis H_1^*, reject H_k^* $(k = 1, \ldots, K)$ if $\Pr(\min_{k \leq j \leq K} V_j \leq \tilde{\eta}_k^*) \leq \alpha$, provided that $H_1^*, \ldots, H_{k-1}^*$ have been tested and rejected. It can be shown that the Type I error probability of this procedure is α asymptotically for any

- Total follow-up time, including follow-up beyond the fourth recurrence, is shown for all individuals.
- Unusually, this is listed before the baseline variables - randomized treatment (1=placebo, 2=thiotepa), initial number and initial size. Within each treatment group, individuals are listed in order of follow-up time.
- There is no visually obvious difference between treatment groups in the numbers of recurrences.

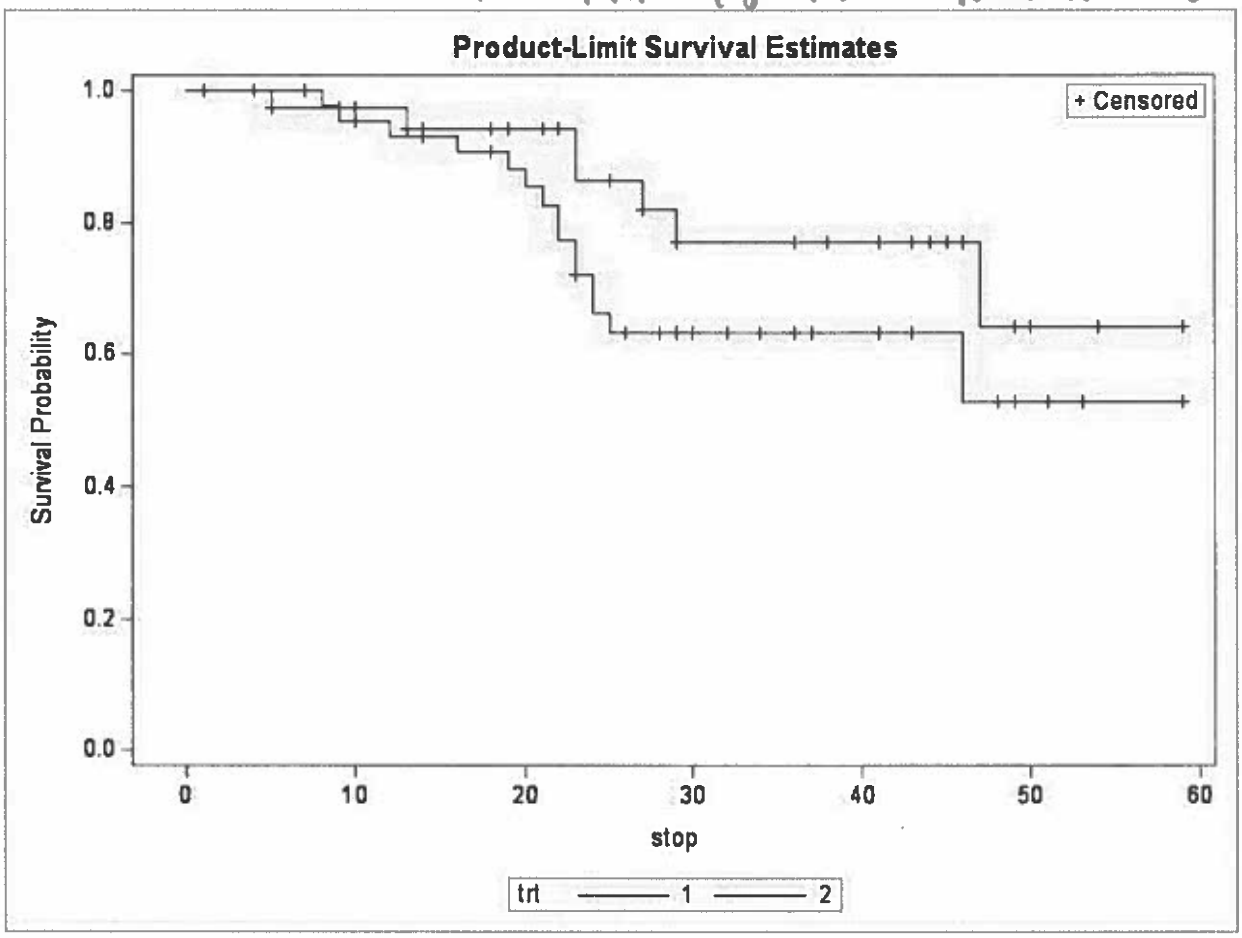
WLW Time to First Recurrence



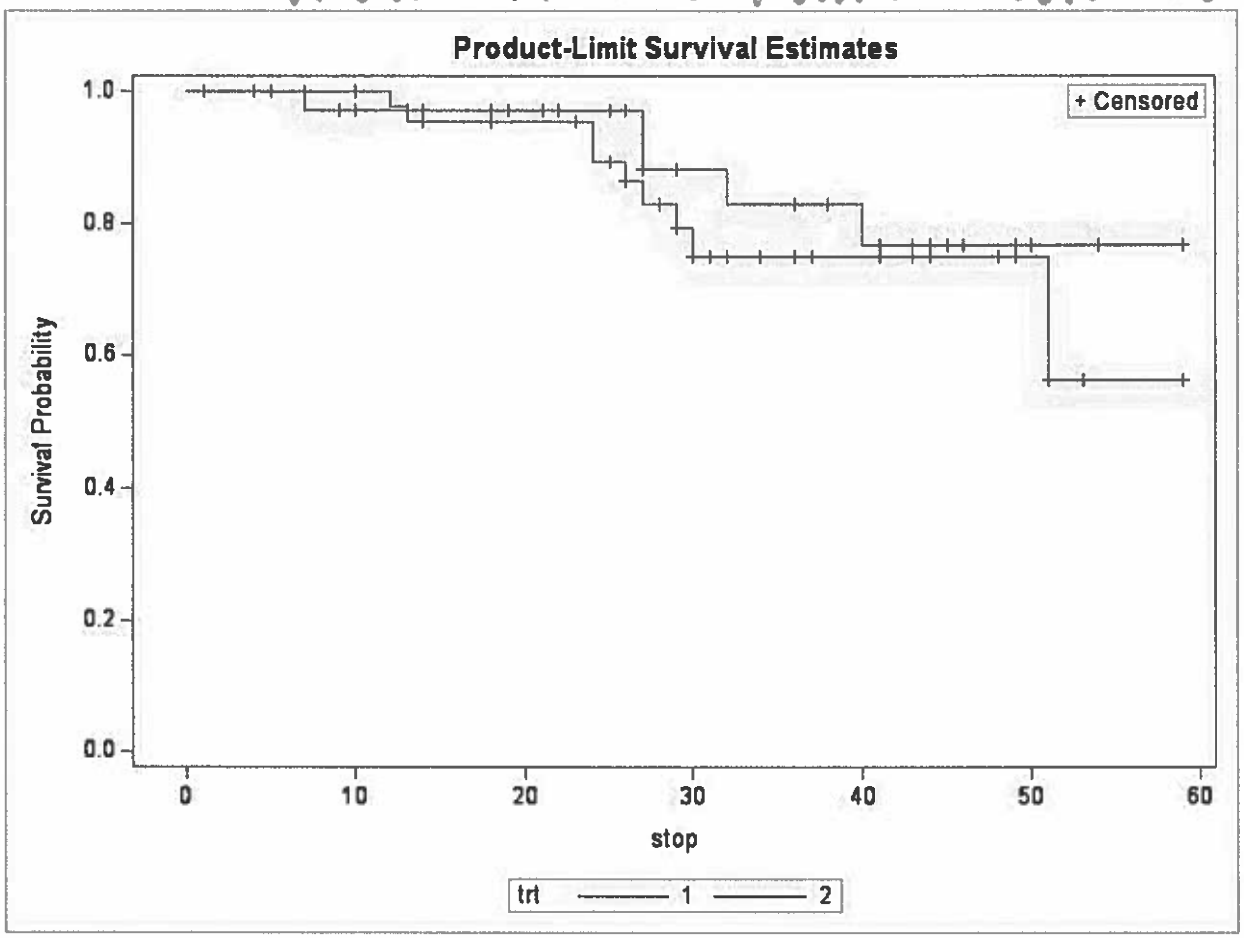
W.L.W. Time to Second Recurrence



W.L.W. Time to Third Recurrence



W.L.W. Time to Fourth Recurrence



- Classify individuals by treatment and the number of recurrences reported:-

Recurrences	0	1	2	3	4	All	Mean
Placebo	18	10	4	6	9	47	1.53
Thiotepa	20	8	3	2	5	38	1.05

- We can create several 2×2 tables by dichotomizing:
  The odds ratios are shown, e.g. (18 x 18/(29 x20) = 0.5586

18 29	28 19	32 15	38 9
20 18	28 10	31 7	33 5
0.5586	0.5263	0.4187	0.6340

- Fit a (discrete) proportional odds model to assess a difference between groups:

  Odds ratio = 0.551 (95% CL 0.249, 1.220), P = 0.14

- Or, compare the distributions by a two-sample t-test:

  Mean difference -0.48 (95% CL -1.129, 0.170), P = 0.14

- Now include the baseline variables number and size, use the individual data for $n = 85$ subjects:

Variable	Odds Ratio	LCL	UCL	P
Treatment	0.469	0.206	1.064	0.07
Number	1.346	1.064	1.703	0.01
Size	1.036	0.783	1.371	0.81

- These results closely parallel those of WLW from the survival model.

- A simple adjustment for follow-up time (c) is to include it in the prediction equation - as another "baseline" variable.
- Assume c is not influenced by outcome data.
- Surprisingly, a time-dependent Cox analysis detects no effect of number of recurrences on follow-up time c even though end of follow-up includes some deaths.
- $\log c$ is a natural transformation - a coefficient of unity implies odds-ratio is proportional to follow-up time.

- Recap previous table...

Variable	Odds Ratio	LCL	UCL	P
Treatment	0.469	0.206	1.064	0.07
Number	1.346	1.064	1.703	0.01
Size	1.036	0.783	1.371	0.81

- Include $\log(t)$ as an additional predictor, last row is the coefficient of $\log(t)$.

Variable	Odds Ratio	LCL	UCL	P
Treatment	0.471	0.203	1.091	0.08
Number	1.404	1.064	1.703	0.006
Size	0.999	0.750	1.333	0.99
$\log c$	0.999	0.267	1.732	0.0075

- Interestingly, the coefficient of $\log c$ is (very) close to unity and other results are broadly similar to the previous table.

3. THE PROPORTIONAL ODDS MODEL IN SURVIVAL ANALYSIS

- We now consider times to a single event, dichotomizing time
- Considering a single binary variable (e.g. treatment) and T_j the time to an event with distributions $F_j(t) = \text{pr}(T_j < t)$, the odds ratio for failure before t is

$$\theta(t) = \frac{F_1(t)}{1 - F_0(t)} \Big/ \frac{F_0(t)}{1 - F_0(t)} = \theta(t).$$

- Bennett's (1983) proportional odds model asserts $\theta(t) = \theta$.
- In contrast to the proportional hazards (Cox) model, here

$$\frac{h_1(t)}{h_0(t)} \to 1 \ (t \to \infty), \text{ for any } \theta.$$

4. PROPORTIONAL ODDS AS A FRAILTY MODEL

- Recall that a (traditional) frailty model involves the inclusion of an unobserved random effect (W say) into the hazard function.
- So, in a two-sample model, an individual with frailty w would have hazard $wb(t)$ if in sample 1 or $\theta wb(t)$ if in sample 2, where θ is the conditional hazard ratio
- The two *individual* survival functions are respectively $\bar{B}(t)^w$ and $\bar{B}(t)^{\theta w}$ where $\bar{B}(t) = \exp\{-\int_0^t b(u)du\}$.
- The observed *population* survival functions are $1-F_0(t) = \bar{F}_0(t) = E\{B(t)^W\}$, $1-F_1(t) = \bar{F}_1(t) = E\{B(t)^{\theta W}\}$.
- Suppose $W \sim \mathcal{E}(1)$. Then

$$\bar{F}_0(t) = \frac{1}{1-\log \bar{B}(t)},\ \bar{F}_1(t) = \frac{1}{1-\theta \log \bar{B}(t)},\ \frac{F_1(t)}{1-F_1(t)} = \theta \frac{F_0(t)}{1-F_0(t)}.$$

- This identity allows the proportional odds model to be fitted by the EM algorithm, treating the W's as missing data.
- Therneau and Grambsch (2000) show that the same estimates can be obtained much more rapidly using using penalized likelihood.
- This approach is incorporated into the "coxph" program under the "frailty" option.
- The program reports two different estimates of standard erororr. My analyses use the larger one "se(coeff)" not "se2".

5. THE LOG-LOGISTIC DISTRIBUTION

- Recall the (quasi)-theorem AL + PH = Weibull*, where AL denotes the accelerated life (scale change) model.
- There is an analog AL + PO = Log-logistic* with

$$F_0(t) = 1 - \frac{1}{1 + (\rho t)^\kappa},$$

for then

$$F_1(t) = 1 - \frac{1}{1 + \theta(\rho t)^\kappa}.$$

- The AL parameter is $\rho\theta^{(1/\kappa)}$.

6. COMPARATIVE ANALYSES

- Present four analyses of the bladder data, two semiparametric using coxph, the other parametric, using survreg.
- Compare proportional hazards estimates to proportional odds estimates.
- Coefficients and standard errors are reported for models including (i) treatment alone (column 1), and (ii) treatment, number and size (columns 2-4).
- Effects are shown on the log-scale for both proportional hazards and proportional odds model.
- The estimates and standard errors from "survreg" using the accelerated life parameterization have been converted to the PH and PO parameterization
- Analyses are presented for each recurrence and then for all recurrences.

Coxph fits for each event and for all events

Model	treatment	treatment	number	size
I PH	-0.37 (0.30)	-0.53 (0.32)	0.24 (0.076)	0.070 (0.10)
I PO	-0.39 (0.41)	-0.59 (0.43)	0.34 (0.12)	0.10 (0.14)
II PH	-0.57 (0.39)	-0.63 (0.39)	0.14(0.092)	-0.078 (0.13)
II PO	-0.82 (0.48)	-1.01 (0.50)	0.17(0.13)	-0.14 (0.17)
III PH	-0.62 (0.46)	-0.70 (0.46)	0.17(0.10)	-0.21 (0.18)
III PO	-0.78 (0.54)	-1.00 (0.56)	0.23(0.15)	-0.23 (0.21)
IV PH	-0.43 (0.56)	-0.64 (0.58)	0.33 (0.13)	-0.21 (0.23)
IV PO	-0.50 (0.63)	-0.80 (0.67)	0.40 (0.17)	-0.18 (0.26)
ALL PH	-0.48 (0.33)	-0.59 (0.31)	0.21 (0.066)	-0.052 (0.095)
ALL(*) PO	-0.68 (0.33)	-0.98 (0.35)	0.32 (0.10)	-0.017 (0.12)
ALL(**) PO	-0.61 (0.25)	-0.83 (0.36)	0.28 (0.069)	-0.065 (0.086)

Survreg fits for each event and for all events

Model	treatment	treatment	number	size
I Weibull	-0.46 (0.30)	-0.61 (0.31)	0.25 (0.076)	0.06 (0.10)
I Loglogistic	-0.51 (0.42)	-0.69 (0.43)	0.34 (0.12)	0.09 (0.14)
II Weibull	-0.60 (0.39)	-0.63 (0.39)	0.14(0.092)	-0.078 (0.13)
II Loglogistic	-0.85 (0.49)	-1.08 (0.51)	0.21(0.13)	-0.17 (0.17)
III Weibull	-0.72 (0.46)	-0.80 (0.46)	0.22 (0.11)	-0.21 (0.18)
III Loglogistic	-0.87 (0.55)	-1.10 (0.57)	0.26 (0.15)	-0.24 (0.21)
IV Weibull	-0.53 (0.56)	-0.70(0.57)	0.33 (0.13)	-0.21 (0.23)
IV Loglogistic	-0.58 (0.63)	-0.87(0.67)	0.35 (0.17)	-0.19 (0.26)
ALL Weibull	-0.55 (0.35)	-0.66 (0.33)	0.23 (0.070)	-0.057(0.10)
ALL Loglogistic	-0.67 (0.46)	-0.90 (0.45)	0.31 (0.11)	-0.071(0.14)

7. A MODEL FOR ORDERED EVENT TIMES

- WLW did not present a model for generating data satisfying proportional hazards for ordered failure times.
- Yang and Ying (2001) note that if the joint survivor function $\bar{F}(t_1, t_2) = \text{pr}(T_1 > t_1, T_2 > t_2)$ exhibits positive dependence in the sense of having Clayton odds ratio

$$\bar{F}\frac{\partial^2 \bar{F}}{\partial t_1 \partial t_2} \Big/ \left(\frac{\partial \bar{F}}{\partial t_1}\frac{\partial \bar{F}}{\partial t_2}\right) > 1$$

  then $\bar{F}^\theta(t_1, t_2)$ is also a survivor function for all $\theta > 0$.
- $\bar{F}$ and $\bar{F}^\theta$ have the same support and marginal distributions satisfying proportional hazards.
- They state that their approach extends to k-variate distributions but with different θ_j for k the ordered variates.
- They did not study families other than proportional hazards.

- We now present a simple construction for arbitrary bivariate families - extensions to k-dimensional are straightfoward but notationally complex.
- If the joint distribution of T_1 and T_2 is supported on $t_2 > t_1$ then $F_1(t) = \text{pr}(T_1 \leq t) \geq F_2(t) = \text{pr}(T_2 \leq t)$ since the latter event implies the former.
- Assuming continuity of the marginal distributions, converting to integrated hazards yields $H_1(t) \geq H_2(t)$.
- For simplicity we assume that $H_1(t) \geq \psi H_2(t)$ where $\psi > 1$, and that each H_j is strictly increasing.
- We construct an absolutely continuous joint distribution with support on $t_2 > t_1$ and with marginals F_1 and F_2.

- We shall assume (for now) the existence of a continuous bivariate distribution (Y_1, Y_2) supported on $y_2 > y_1$ and with Y_1 and Y_2 each exponential, with means θ_1 and $\theta_2 = \psi\theta_1$ (this will be shown below).

- The inverse cumulative hazard functions $H_1^{(-1)}(y)$ and $H_2^{(-1)}(y)$ are both monotone increasing and have $H_1^{(-1)}(y) < H_2^{(-1)}(y)$. Set $T_j = H_j^{(-1)}(Y_j/\theta_j)$ for $j = 1, 2$.

- Since the Y_j/θ_j are unit exponential,

$$\begin{aligned}
\text{pr}(T_j > t) &= \text{pr}\{H_j^{(-1)}(Y_j/\theta_j > t)\} = \text{pr}\{Y_j/\theta_j > H_j(t)\} \\
&= \exp\{-H_j(t)\} = 1 - F_j(t); \\
Y_1 < Y_2 &\Rightarrow \theta_1 H_1(T_1) < \theta_2 H_2(T_2) \\
&\Rightarrow \theta_2 H_2(T_1) < \theta_2 H_2(T_2) \Rightarrow T_1 < T_2,
\end{aligned}$$

  showing both required properties hold.

- The construction is not reversible.
- *Example:* Let $Y_2 = Y_1 + Y$, where Y_1 and Y are independent $\mathcal{E}(1)$.
- Then

$$\bar{F}(y_1, y_2) = \exp(-y_1)\{1 + (y_2 - y_1)\}\exp(-y_2), (y_2 > y_1),$$

$$\bar{F}_1(y_1) = \exp(-y_1), \quad \bar{F}(y_2) = (1 + y_2)\exp(-y_2),$$

$$H_1(y_1) = y_1, \quad H_2(y_2) = y_2 - \log(1 + y_2)$$

and $Y_1 < Y_2$ does not guarantee that $H_1(Y_1) < H_2(Y_2)$ or even that $H_1(Y_1) < \psi H_2(Y_2)$ for any fixed ψ.

8. THE BIVARIATE EXPONENTIAL DISTRIBUTION

- We start with a bivariate exponential distribution with $T_1 \leq T_2$ but positive mass along $T_1 = T_2$.
- Generate $T_1 \sim \mathcal{E}(\theta_1)$, $T_2 = T_1 + Z$, where $Z = I \cdot T$, where $I \sim \text{Be}(1 - \theta_1/\theta_2)$ and $T \sim \mathcal{E}(\theta_2)$.
- A simple calculation, or appeal to the lack of memory property of the exponential, shows that $T_2 \sim \mathcal{E}(\theta_2)$.
- This model is sometimes used in cancer studies for the joint distribution of survival time (T_2) and progression-free survival time (T_1).

- The degenerate distribution with $T_1 \sim \mathcal{E}(\theta_1), T_2 = \psi T_1$ also has the required marginal properties.
- Here $\text{corr}(Z, T_1 = 1)$.
- A "compromise" is to take

$$T_2 = \alpha T_1 + I \cdot T,$$

  where $1 < \alpha < \phi$, $T \sim T_2$ and $I \sim \text{Be}(1 - \alpha/\phi)$.
- The same representation applies, with θ_1 replaced by $\alpha\theta_1$
- This distribution assigns positive probability to the line $t_2 = \alpha t_1$ and a continuous density in the region $t_2 \geq \alpha t_1$.
- Easily, $\text{cov}(T_1, T_2 - T_1) = (\alpha - 1)\theta_1^2$.

- Fianlly. we replace α by a random variable A, say, with density $g(a)$ over $(1, \phi)$ and independent of T_1. (A separate realization of A is drawn for each pair). This smears out the discontinuity along the line $t_2 = at_1$ resulting in an absolutely continuous distribution over $t_1 < t_2$.

- The resulting joint density, $f(t_1, t_2)$ over $t_1 < t_2$, is discontinuous, leading to possible issues in maximum likelihood fitting.

- To simulate this distribution:

  (i) Draw $T_1 \sim \mathcal{E}(\theta_1)$;

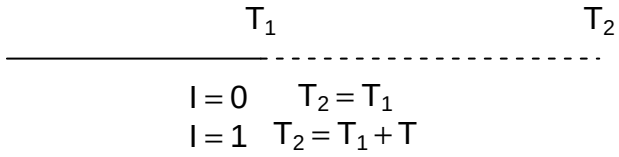
  (ii) Draw $A \sim g(a)$, independent of T_1;

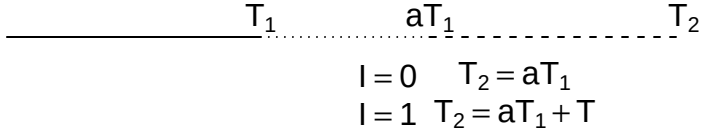
  (iii) Draw $I \sim \text{Be}(1 - A/\psi)$;

  (iv) Draw $T \sim \mathcal{E}(\theta_2)$;

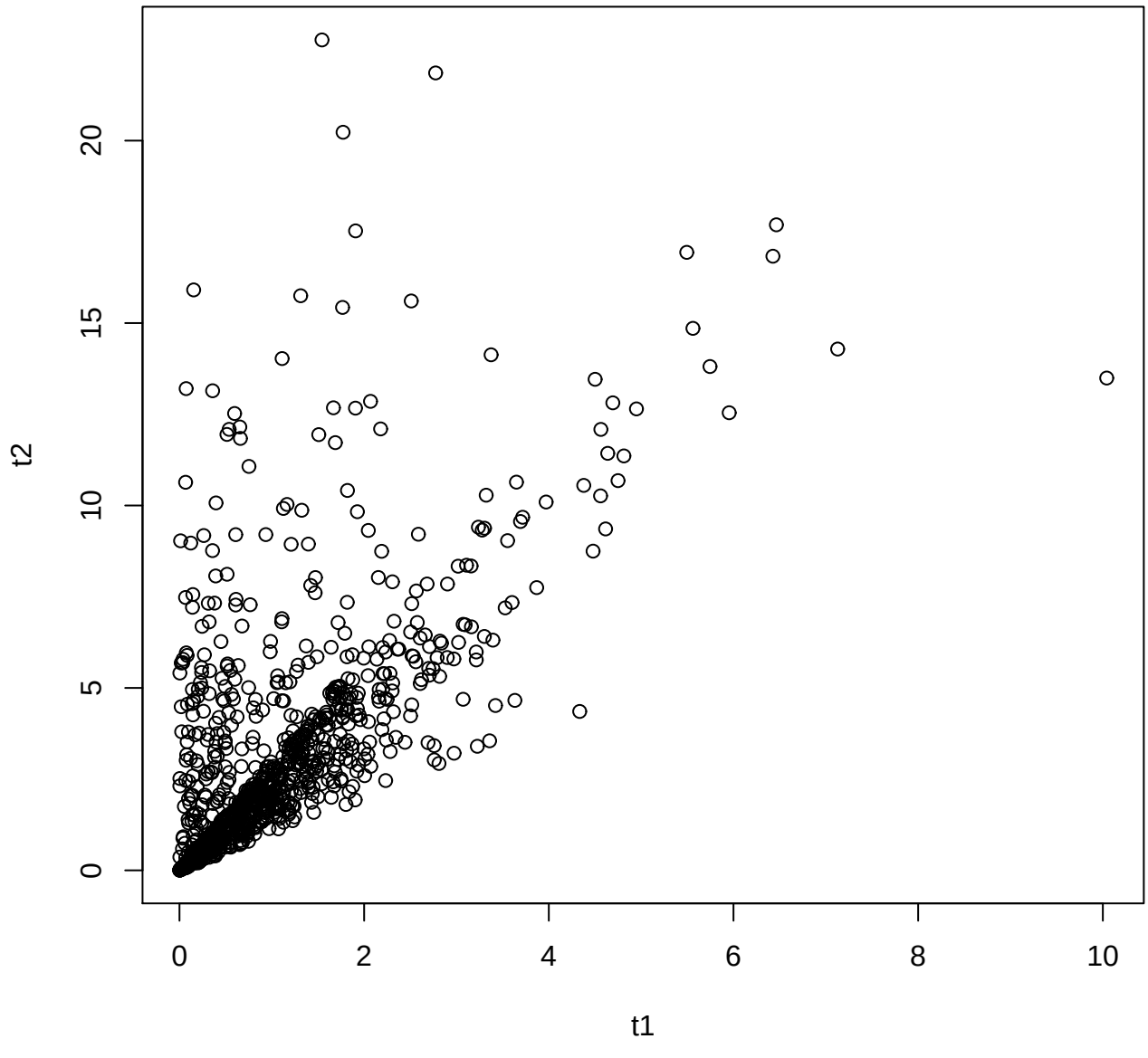
  (v) Calculate $T_2 = A \cdot T_1 + I \cdot T$.

Construction of a Bivariate Exponential Distribution





```
>
> setEPS()
> postscript("bivarplots.eps")
> u <- runif(1000)
> v <- runif(1000)
> t1 <- -log(u)
> #t1 is exp with mean 1.
> a <- runif(1000,1,3)
> #a is unform over (1,3) determining the multiplier
> t <- -3*log(v)
> #t, like t2, will be exponential with mean 3
> ind <- rbinom(1000, 1, 1-a/3)
> t2 <-  a*t1 + ind*t
> mean(t2)
[1] 3.089
> sd(t2)
[1] 3.219799
> plot(t1, t2)
> dev.off()
null device
          1
> q()
```


t2
0
5
10
15
20
0
2
4
6
8
10
t1

- The conditional density of T_2 given T_1 can be derived but is complicated and has discontinuities.
- The model may be extended to k variates, iterating the previous construction to generate the conditional density of T_k give T_{k-1}.
- We could use the same value of a for each k, resample from the same density, or allow different densities $g_k(a)$ for each k.

9. CONCLUSION

- We can incorporate an individual-level frailty into the construction (since multiplying the baseline hazards $H_j(\cdot)$ by the same random quantity W preserves their ordering.
- The resulting *marginal* distributions will be the same as if we included independent frailties for each it event (which might not preserve the ordering).
- This is one route to a model for multiple events with the ordered event times satisfying proportional odds.
- The model differs from an Andersen-Gill model with frailty.

10. REFERENCES

Wei, L.J., Lin. D.Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J. Am. Statist. Assoc, 84, 1065-1073.

Andrews, D.F. and Herzberg, A.M. (1985) Data: New York, Springer,

Therneau, T.M., and Grambsch, P.M. (2000) Modeling Survival Data: Extending the Cox Model. New York, Springer.

Cook, R. J. and Lawless, J.F. (2006) The Statistical Analysis of Recurrent Events. New York, Springer,

Yang, Y. and Ying, Z. (2001) Marginal proportional hazards models for multiple event-time data. (2001). Biometrika, 88, 581-586.

THANK YOU FOR YOUR ATTENTION

How should we compare survival outcomes with delayed treatment effects?

Rick Chappell
Professor, Departments of Statistics and
of Biostatistics & Medical Informatics
University of Wisconsin
chappell@stat.wisc.edu

Disclosures:

- **Student support from Merck, Inc.**

- **Acknowledgements:**

Useful discussions with Keaven Anderson and Amarjot Kaur of Merck.

Intellectual and programming contributions by Mitchell Paukner, former student at Wisconsin.

Twin themes of this talk:

- **Interpretability**
- **Power**

Outline

I. Some Historical Background

II. Rank Tests in Trials with Late Effects

III. Choice of Scale: Milestone Tests

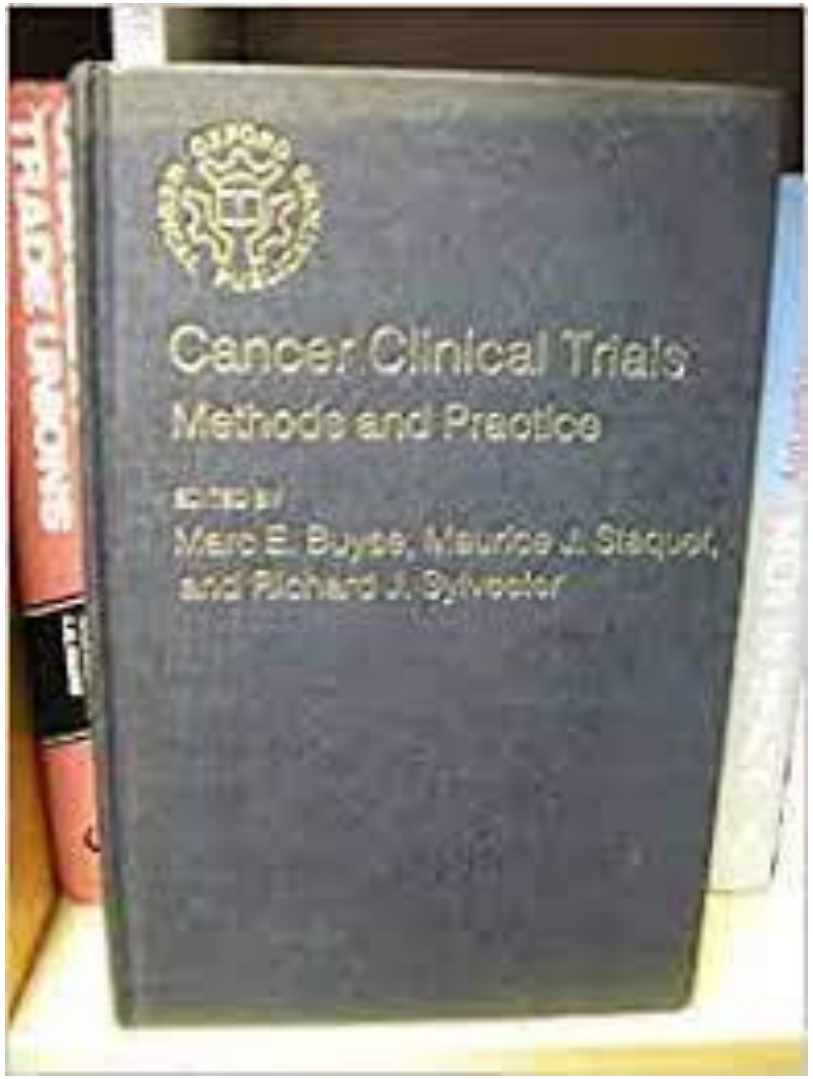
IV. An Intuitive Approach to Concentrating on Late Differences: Window Mean Life

I. Background - What I "knew" in 1984

"The questions are the same but the answers have all changed"

(The punchline from an old joke).

When I first became interested in survival analysis (~1984), what were the accepted means of comparing two survival curves from randomized data?



The most authoritative book at the time (Buyse, Staquet & Sylvester, 1984) had a chapter ”Comparison of Survival Curves” by Breslow.

It was dominated by the section “Two Sample Rank Tests for Censored Data” which described the **log-rank** and **Wilcoxon** tests.

These were the only tests computed by SAS (and no other packages accommodated survival data).

Chappell's theory: many of our methods are (often unknowingly to us) relics of pre- or early- computing exigencies.

Relevance here:

- Rank tests are easy to compute;
- Ditto the proportional hazards model, even though it is iterative; it relieves us of the need to estimate the hazard function and assess distributional fit.

Other quantities might be of greater interest.

Paul Meier: the emphasis should be on the model's **naturalness and ease of interpretation.**

We should **adapt traditional models** to the data if possible.

The Meier Prize

The "**Meier Prize**": One dollar to whomever can find an instance of the proportional hazards model or log-rank (Wald exponential score) test applied to uncensored data.

(Never awarded, continued to be endowed by myself.)

Censoring ought to be the statistician's problem, not the reader's. Why not modify existing familiar methods to accommodate it?

Cox Regression: his admission

- The log-rank test is the score test for a binary covariate in Cox regresson
- Its associated estimate is the hazard ratio.
- What do hazard ratios mean? Why not differences?
- Even Cox had his doubts:

"The proportional hazards model is not associated with a simple underlying generating process (as opposed to an accelerated life model)"

[Cox, 1997, in Lin & Fleming, eds, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*]

Consequences of the PH assumption

- Power for logrank tests and its extensions involving Cox PH is clearly best under proportionality; other weights (e.g., “Wilcoxon” analog decreasing weights) are more powerful under other models (e.g., proportional odds).

Differences between these tests can even be used as a diagnostic for PH (Chappell, *Biometrika*, 1989).

- In some circumstances (Martinez, Sinha, Wang, Lipsitz, Chappell, *Stat Meth. Med. Res.* 2017), PO-based tests’ size is closer to nominal than PH-based ones.

- Cox regression is not in general collapsible; omitting covariates may violate (or induce) PH

Some Ways Effects can be Defined for a Time-to-Event Trial

1. Survival function $S(t)$ at all followup times
2. Hazard function $\lambda(t)$ at all followup times
3. Event rate by a given time, e.g., $S(5)$
4. Median or restricted mean time to event

More Recent Ways Effects can be Defined for a Time-to-Event Trial

5. Accelerated life (e.g., “5-STAR” by Mehrotra & West, *Stat Med*, 2020)

6. Probability of better performance by “X” months (Buyse, *et al. J Clin. Epi.* 2021)

7. Weighted hazard functions.

The most commonly referenced (but far from the only) choice is the Fleming-Harrington weight function:

- *FH(p,q)* weights are estimates of $S(t_i)^p \times [1-S(t_i)]^q$.

8. “Versatile” tests, which maximize the test statistic over a range fns (Kosorok, Ying & Prentice, Karrison, Anderson, Uno ...). A popular modern popular choice (with variations):

- ”Max Combo” = maximum of Fleming-Harrington *(0,0)*, *(0,1)*, *(1,0)*, and *(1,1)* test statistics.
- Other variations exist with different weight functions.

Choices 1. – 6. are all “directional”: they test for superiority.

II. Rank Tests in Trials with Late Effects: "Sensibility and the Problem with Increasing Weights"

- Directional tests examine the hypothesis of one treatment being *superior* on some scale.

  Others merely examine differences.

- Versatile, or omnibus, rank tests such as Max Combo use the maximum of several choices, usually including ones with increasing weights.

- These can be more powerful against a range of alternatives ("versatile") but are less interpretable than directional tests.

- They arose in the design of prevention and immunotherapy trials with presumably late differences.

The problem with increasing weights

- *"FH(p,q)* weights are estimates of $S(t_i)^p \times [1-S(t_i)]^q$."

Thus the *(0,1)* and *(1,1)* weights will *increase*.

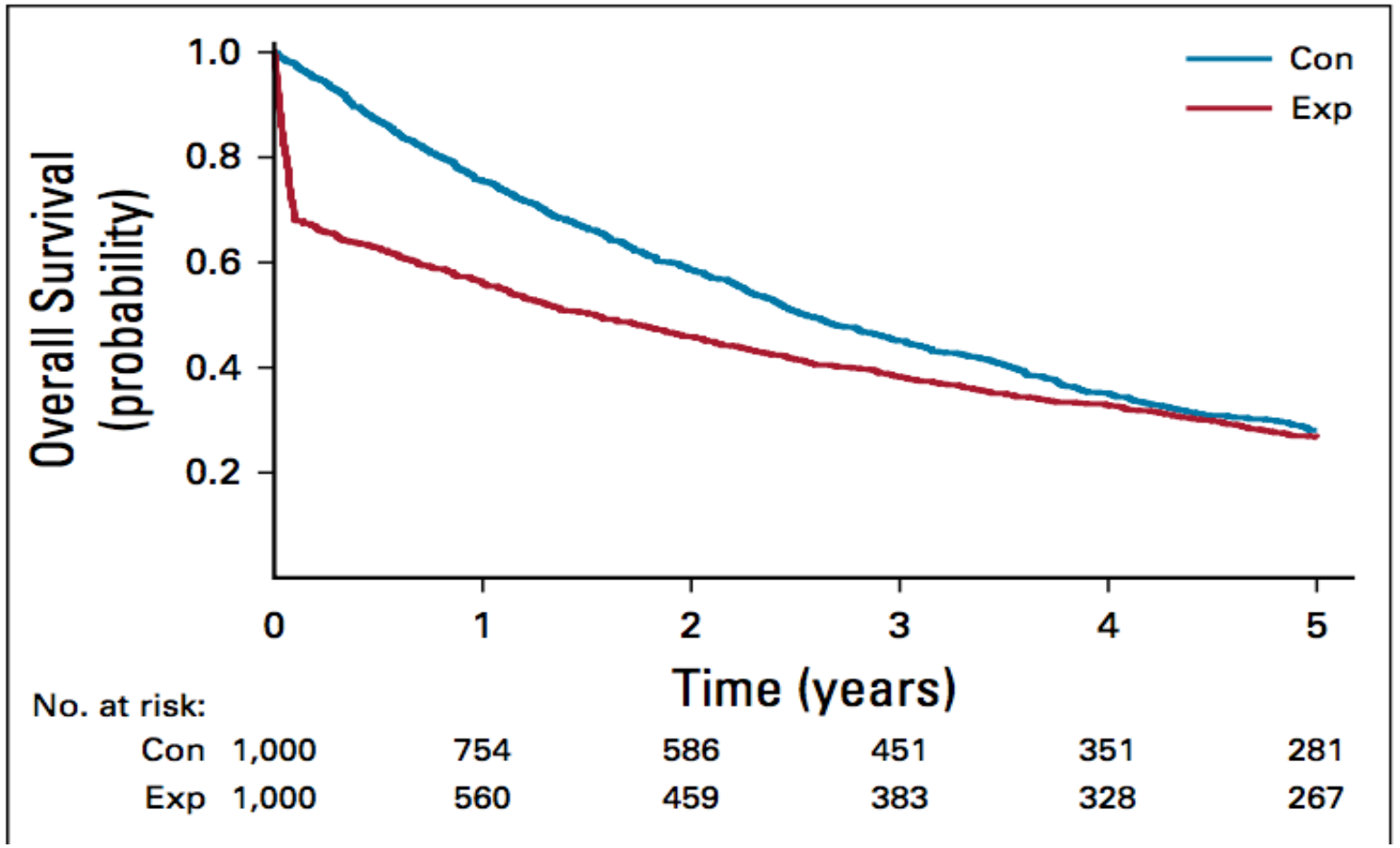
- This is problematic (a euphemism) for the same reason that the post-randomized three-week cardiac event exclusion rule was a bad idea, as pointed out in papers by DeMets and Sackett in the early 1980s.

Digression: discussion of the old "three [etc.]-week rule" in clot-busting trials.

- The Wilcoxon-analog rank test uses decreasing weights [in the modern version, an estimate of 1 – S(t)]. It is suitable for cases where late failures are less important (perhaps diseases in the aging).

- But why would we want to downweight early events? When are early failures preferable to late ones?

- For example, Korn, *et al.* (*JCO*, 2019) recently published a hypothetical but realistic paradoxical example of a one-sided test of superiority in which a survival curve which was rejected *in favor of* another despite being completely dominated by it.



A **FH(0,1) rank test rejects equality *in favor of*** the Exp (red curve) treatment with a one-sided p-value of .0046. This is due to the increasing weight function which minimizes the effect of the Exp treatment's early decline.

- Magirr & Burman (*SiM*, 2018) addressed this by developing “modestly weighted logrank tests”.
- They note that even Fleming and Harrington (their book, p. 267) say that a subclass of their tests aren’t consistent under stochastic ordering.
- They propose weights which control this risk at some cost in power but with better performance in the presence of delayed effects than the unweighted logrank tests.
- (The earliest proposal I can find for downweighting is in the Women’s Health Initiative design paper by Prentice in *CCT*. It was later abandoned in favor of a logrank test.)

Choice of Scale in Trials with Late Effects: "Sensibility"

Isn't it reasonable to demand that, when comparing two treatment groups, an earlier failure in one group shouldn't increase the evidence in its favor?

- E.g., suppose we come up with a test statistic in favor of Group A.
- We are then told that a group A failure of *5 years* should have been *5 months*.
- We then correct the mistake and rerun the analysis.
- Shouldn't it be impossible for the test statistic to increase?

Notions of Sensibility

Alas, this cannot be achieved in practice:

- Oakes ("A note on the Kaplan-Meier estimator", *Am. Stat.,* 1993) showed the the KM estimator can increase when a failure time is moved up;
- The log-rank test can be shown to have a similar property;
- This even holds for the t-test with uncensored data.
- It's a danger for any test with > 1 parameter.

A Notion of Sensibility: "Asymptotic Sensibility"

Consider a control survival curve C (eliding the time argument) and treatment curve T*, some comparison function $\Delta(.)$, with null/alternative hypotheses

$$H_0: \Delta(T^*,C) \geq 0 \text{ and } H_1: \Delta(T^*,C) \leq 0.$$

Then a test is "sensible" with respect to $\Delta(.)$ if, for any curve T such that $T^* \geq T$, asymptotic relative efficiency for the test of T* in the above hypotheses exceeds that for T.

"If one treatment curve is inferior to another there is asymptotically less evidence to show it better than the control."

“Difference” vs. “Directionality”

“We show that versatile tests, while achieving robustness to departures from proportional hazards, may lose interpretation of directionality (superiority or inferiority) and can only be seen to test departures from equality. ”
(Chen, Wang, Chen, Zheng, Chappell, and Dey, *Clin Trials*. 2020).”

III. Choice of Scale: Milestone Tests ("Diff. in x-year Survival")

- Why not test for equality at a pre-specified time?
- Kalbfleisch and Prentice: "Such a procedure would not make efficient use of the data."
- But it would certainly be robust, directional, and interpretable.
- From **Jiren Sun**'s simulations at the University of Wisconsin:

Milestone survival vs logrank test

Simulations (continued)

Ten thousand simulations were conducted for every scenario with n = 200 subjects in each arm, and with three different levels of censoring.

- No Censoring
- Moderate Censoring: Censoring is distributed exponentially with $\lambda = 0.02$ for each trial arm.
- Heavy Censoring: Censoring is distributed exponentially with $\lambda = 0.08$ for each trial arm.

Milestone survival vs logrank test

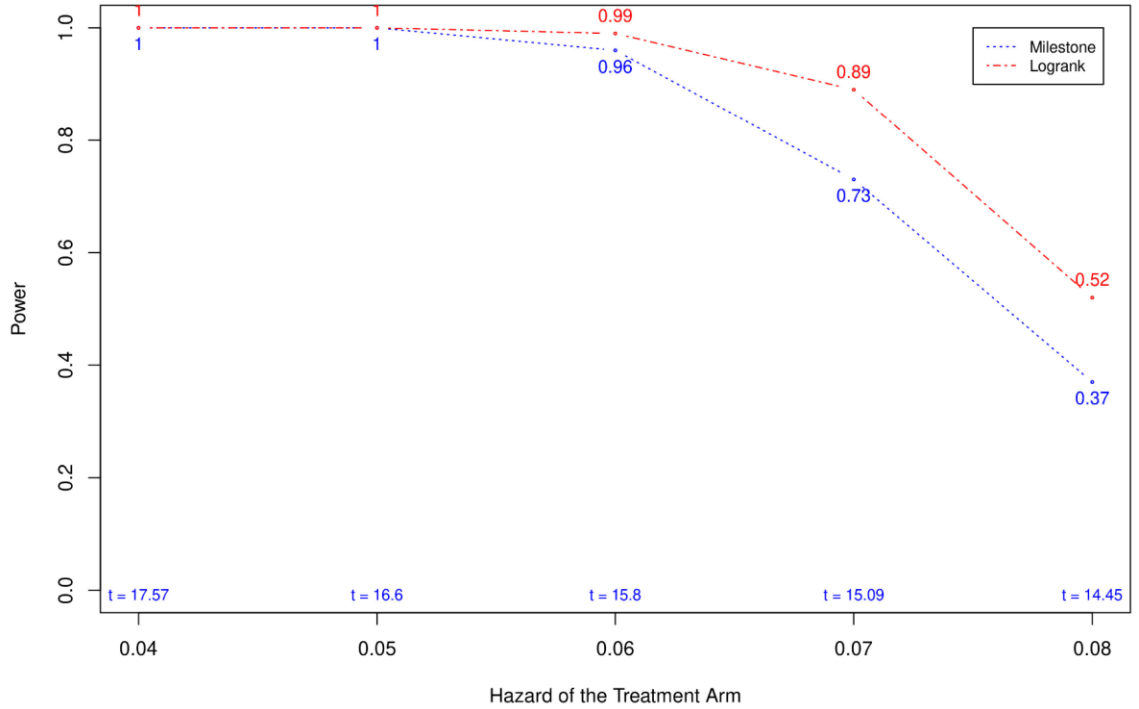
Simulations

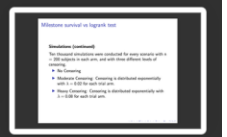
We consider a two-arm RCT with right censoring. In each simulation scenario, the control arm is distributed exponentially with median 7 months ($\lambda_C = 0.1$), while the treatment arms in each of the three different survival scenarios are distributed as

- Proportional: Treatment arm is distributed exponentially with λ_T ranging from 0.04 to 0.08.
- Moderate Effect Delay: Treatment arm is distributed piecewise exponentially with $\lambda_T = 0.1$ from $t = 0$ to $t = 4$, and λ_T ranging from 0.04 to 0.08 after $t = 4$.
- Large Effect Delay: Treatment arm is distributed piecewise exponentially with $\lambda_T = 0.1$ from $t = 0$ to $t = 8$, and λ_T ranging from 0.04 to 0.08 after $t = 8$.

Milestone survival vs logrank test

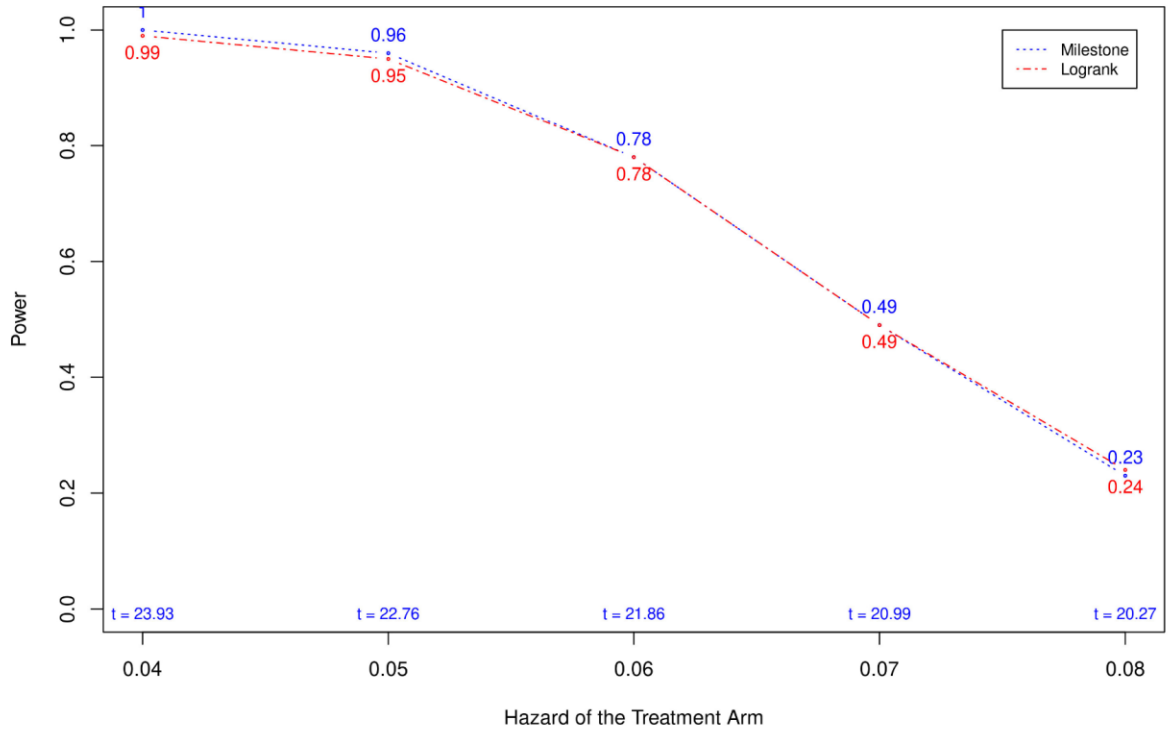
- Moderate Censoring
  - Proportional Hazards

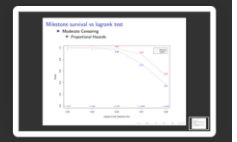




Milestone survival vs logrank test

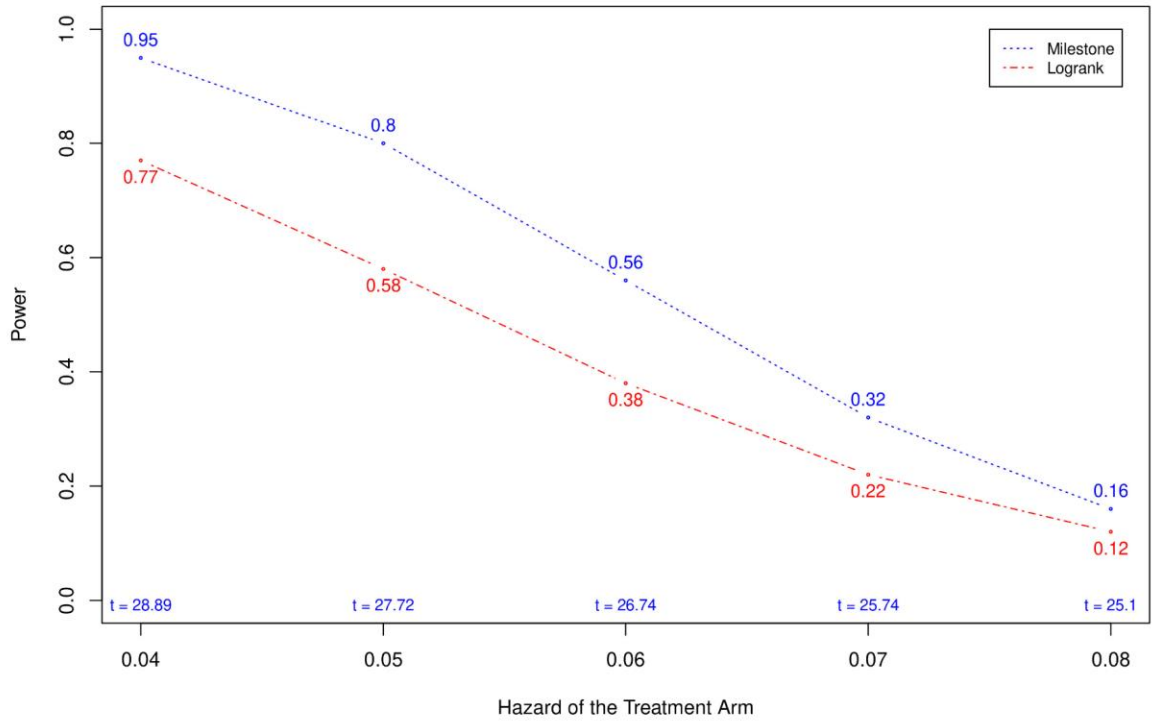
- Moderate Censoring
  - Moderate Effect Delay

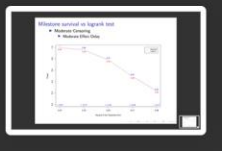




Milestone survival vs logrank test

- Moderate Censoring
  - Large Effect Delay

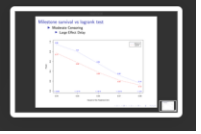




Milestone survival vs logrank test

Which test yields higher power in the following scenarios?

	Proportional	Moderate Effect Delay	Large Effect Delay
No Censoring	Logrank	Logrank	Milestone
Moderate Censoring	Logrank	Logrank/Milestone	Milestone
Heavy Censoring	Logrank	Milestone	Milestone



Note: Milestone tests were conducted at the optimal times
However: Results are quite robust to the choice of those times

“Difference” vs. “Directionality”

- Back to Meier. He liked the mean.
  It is certainly interpretable.

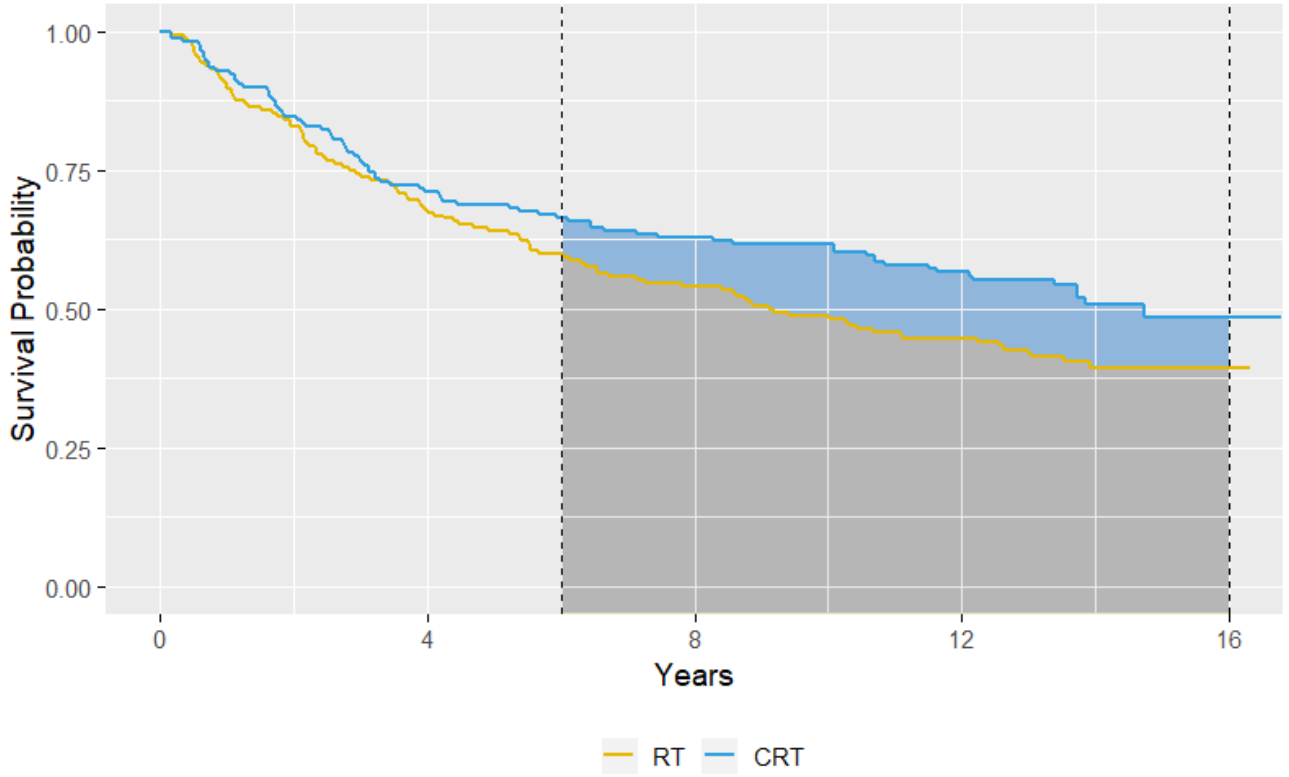
- However, the mean is the area under the survival curve and if the curve doesn’t go to 0 then it isn’t estimable.

- Restricted Mean Life is the area between 0 and a finite time “τ”. It is interpreted as the average gain for the better treatment between 0 and τ.

- It is directional and interpretable, but lacks power for late differences.

Is there a compromise?

IV. Choice of Scale: An Intuitive Approach to Concentrating on Late Differences: Window Mean Life



Window Mean Life (Paukner & Chappell, *Stat. Med.* 2021, 2023)

- Data from Lee, ... Chappell*, ... et al.* ("Randomized trial of radiotherapy plus concurrent-adjuvant chemotherapy vs. radiotherapy alone for regionally advanced nasopharyngeal carcinoma", *JNCI* 2010) were used to construct Kaplan Meier curves in the Figure.

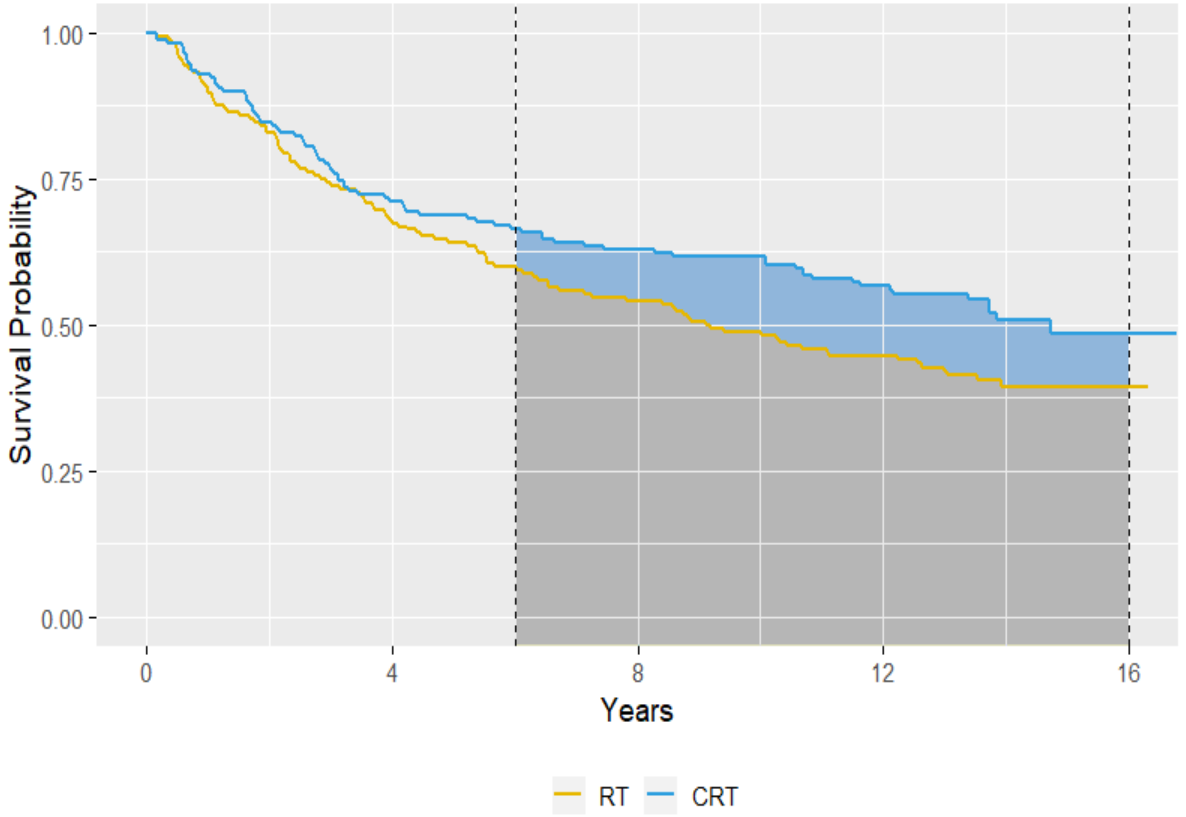
- Subjects with advanced nasopharyngeal carcinoma were treated with either radiotherapy alone (RT) or a combination of radiotherapy and concurrent-adjuvant chemotherapy (CRT).

- WML is applied, combining features of RMST and Milestone tests.

Window Mean Life

- The window was set with $\tau_0 = 6$ and $\tau_1 = 16$

- $\hat{\mu}_{CRT}(6, 16) = 5.8$ years
- $\hat{\mu}_{RT}(6, 16) = 4.7$ years
- $\hat{\Delta}(6, 16) = 1.1$ years



Subsetting Restricted Mean Life increases power to detect late differences while remaining "sensible" – it doesn't ignore or even downweight early events.

Log-rank p-value = .047; WML p-value = .033.

Versatile Window Mean Life (Paukner & Chappell, *Stat. Med.* 2023)

- Window mean life can be made versatile: the maximum test statistic can be taken over several windows.
- The distribution of the maximum can be found.
- The windows can be overlapping or disjoint.
- The associated tests are directional.
- Interpretation can be made as average time gained in the "winning" window.

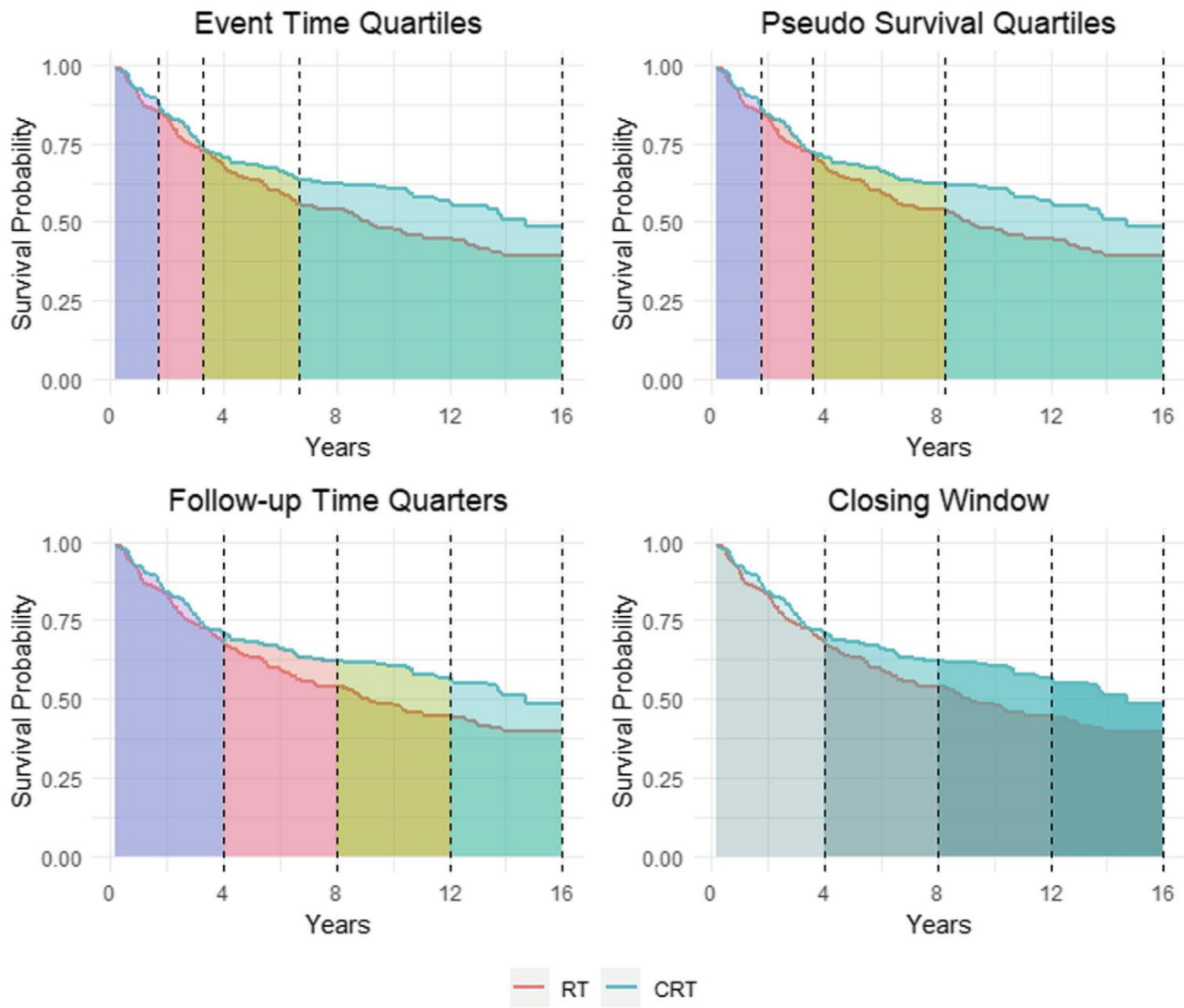


FIGURE 2 Four suggested window selection methods applied to the survival curves from the Lee et al nasopharyngeal carcinoma study

Conclusion

- We may not be able to “Have it all” – there is no one universal solution – but we can flexibly pick the best estimand for a given situation.

- We have a rich new menu from which to choose.

META-ANALYSIS OF VARIABILITY IN SURVIVAL OUTCOMES IN PRECISION ONCOLOGY TRIALS

Ying Lu, PhD

Department of Biomedical Data Science

Stanford University School of Medicine

Department of Biomedical Data Science AT STANFORD MEDICINE

Stanford MEDICINE | Center for Innovative Study Design (CISD)

http://med.stanford.edu/cisd.html

Disclosure

NO conflict of interests for this research project

Research grant to Stanford University

- UCB Biopharmaceutical Inc.

Consultant (DSMBs)

- Roche, Gilead, Nektar, WCG, and DPClinics

Outline

- Introduction
- Methods
- Results
- Discussion and conclusion

INTRODUCTION

- Molecular targeted therapy has revolutionized the landscape of cancer treatment
  - matching patients' tumor molecular profile with the therapeutic targets
  - resulting better treatment responses and less systematic toxicity
  - hundreds targeted cancer drugs have been approved by FDA in the past 23 years (https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies)

Breast Ca	Lung	skin	Lymphoma	Leukemia	Liver	Colorectal
25	32	18	35	33	16	11

- Despite the benefits and promises of increasingly refined therapies, the heterogeneity in survival outcomes for biomarker-enriched alone versus enriched and non-enriched cohorts combined is not fully understood.

INTRODUCTION

- In a randomized clinical trial, let B denote a biomarker status; R be the treatment status; π be the proportion of marker positives; and Y be the outcome measures of interests:
  - B=1 ⇒ marker "+"; B=0 ⇒ marker "-"
  - $\pi = P(B = 1)$
  - R=1 ⇒ by a targeted drug; R=0 ⇒ under standard of care arm
  - a larger value of Y ⇒ more desirable outcome
- Commonly used linear model:

$$Y(B, R) = \mu_{0,0} + \alpha B + \beta R + \gamma BR + \varepsilon$$

- Meta-analysis for oncology trials has been focused on the average treatment difference across trials.

INTRODUCTION

- Null hypotheses to be tested:

$$H_{0,+}: \beta + \gamma = 0 \text{ and } H_{0,1}: \alpha + \gamma = 0$$

- When $\gamma > 0$, $\mu_{1,1} - \mu_{1,0} > \mu_{0,1} - \mu_{0,0}$ and $\mu_{1,1} - \mu_{0,1} > \mu_{1,0} - \mu_{0,0}$

  R significantly benefits more the positives than the negatives

- Furthermore, when $\beta + \gamma > 0$,

$$\mu_{+,1} = E(Y|R = 1) > \mu_{+,0} = E(Y|R = 1)$$

  R benefits all patients

- Meta-analysis methods for treatment mean effect have been well studied.

INTRODUCTION

- However, mean difference doesn't reflect treatment heterogeneity.
- A larger variance means higher heterogeneous responses.
- For the linear model before, let $\sigma_{b,r}^2$ be the variance of Y for $B = b$ and $R = r$, and $V(\varepsilon) = \sigma^2$,

$$\sigma_{+,r}^2 = V(Y|R = r) = \pi\sigma^2 + (1-\pi)\sigma^2 + \left(\mu_{1,r} - \mu_{0,r}\right)^2 \pi(1-\pi)$$

- When $\mu_{1,r} \neq \mu_{0,r}$, $\sigma_{+,r}^2 > \sigma^2 = \sigma_{1,r}^2 = \sigma_{0,r}^2$
- When $\alpha \neq 0, \sigma_{1,0}^2/\sigma_{+,0}^2 < 1$, B is a prognostic marker
- When $\gamma > 0$ and $\alpha \geq 0$, $\left(\sigma_{1,1}^2/\sigma_{+,1}^2\right)/\left(\sigma_{1,0}^2/\sigma_{+,0}^2\right) < 1$, B is a predictive marker.
- Meta-analysis method for heterogeneity is less studied.

INTRODUCTION

- Research Goals:
  - The primary objective of this study is to develop a feasible and rigorous meta-analysis procedure to evaluate treatment heterogeneity for survival outcomes from precision cancer trials.
  - The secondary objective is to assess whether biomarker-based treatments result not only prolonged survival time but also increased consistency in treatment response for the targeted subgroups in cancer trials across different types of cancer.

INTRODUCTION

Challenges # 1: metric for heterogeneity

- For pivotal oncology trials, the primary endpoint is overall survival time (OS).
- Progress-free survival time (PFS) is a surrogate for OS.
- OS is asymmetric and its variance is often positively correlated with the mean survival time

Distribution	Survival Function ($y \geq 0 \geq$)	Mean (m)	Variance
Exponential	$exp(-y/\lambda)$	λ	m^2
Weibull	$exp\left(-(y/\lambda)^k\right)$	$\lambda\Gamma(1+1/k)$	$[\Gamma(1+2/k)/\Gamma^2(1+1/k)-1]m^2$
Log-Normal	$1-\frac{1}{2}\left[1+erf\left(\frac{ln(y)-\mu}{\sigma\sqrt{2}}\right)\right]$	$exp(\mu+\sigma^2/2)$	$[exp(\sigma^2)-1]m^2$

INTRODUCTION

- **Coefficient of variation** (CV), defined as the SD/Mean, reflects the heterogeneity better for survival time
  - Exponential: CV=1
  - Weibull: CV=$\sqrt{\Gamma(1+2/k)/\Gamma^2(1+1/k)-1}$
  - Log-normal: CV=$\sqrt{exp(\sigma^2)-1}$
- Meta-analysis methods for CV have been studied in ecological and evolutionary literature for normally distributed outcomes but not for the survival endpoints.
  - Senior et al., 2020, Nakagawa et al., 2015, Winkelbeiner et al., 2019

INTRODUCTION

Challenges # 2: Parametric models for OS (PFS)

- Can we use parametric distributions to model OS (PFS) from oncology trials?
  - YES!
  - Patient level survival distribution can be parametrized (Plana et al., 2022)
    - a comprehensive analysis of ~150 published phase 3 oncology trials of 220,000 imputed OS or PFS events
    - Two-parameter Weibull or Log-Normal, or Gomperz-Makeham distributions can accurately fit OS or PFS distributions for trials and biomarker defined subsets
    - Weibull and Log-Normal distributions provide equivalent good fits to the individual imputed data
- To evaluate heterogeneity, Log-Normal distribution is a nature choice
  - Qazi et al., 2007
- In this presentation, we use the log-normal distribution to model trial OS (PFS) distributions.

INTRODUCTION

Challenges # 3: Trial selection and individual level patient data

- A meta-analysis protocol has been developed for meta-analysis of breast cancer (and lung cancer) trials according PRISMA criteria
  - Search criteria (inclusion and exclusion criteria for trials to be selected)
  - Information presented (K-M survival curves) for intent-to-treat population and targeted marker positive subpopulation
  - Review, trial characteristics and data collections
- Imputation of individual patient level survival data
  - Individual patient level data from the trial sponsors were not available.
  - a reverse engineering method was used to impute event time and censoring based on published K-M curves using algorithm by Guyot et al., 2012.
  - This algorithm has been used by others (Mukhopadhyay et al., 2022, Plana et al., 2022)

METHODS

Mathematical model specifications:

- Observed data $\left(Y_{i,j}, \delta_{i,j}, R_{i,j}, G_{i,j}\right)$: i is the i^{th} participant in j^{th} trial
  - $Y_{i,j}$ is the last observed time
  - $\delta_{i,j}$ is the event indicator, “1” for death and “0” for censored observation
  - $R_{i,j}$ is the treatment assignment, “1” for treatment arm and “0” for control arm
  - $G_{i,j}$ is the indicator of population, “+” for ITT population and “1” for targeted subpopulation
- Since we don’t have individual marker status, we perform analyses for ITT and targeted populations independently.

METHODS

Choice of Parametric versus Non-Parametric:

- Kaplan-Meier (non-parametric) survival curves were reported.
- Restricted mean survival times (RMST) can be derived for both arms from each trial.
- However, it is difficult to perform uniformed RMST analysis due to various length of trials.
- Parametric models allow extrapolation of survival curves to any length of follow-up.
- We choose the log-normal model to fit the survival curves.

METHODS

Mathematical model specifications:

- Likelihood function for j^{th} trial separately for "+" and "1" population:

$$\mathcal{L}_{j,G} = \prod_{i=1, G_{i,j}=G}^{I_j} \phi^{\delta_{ij}} \left(\frac{ln(y_{ij}) - \mu_{R_{i,j},j,G}}{\sigma_{R_{i,j},j,G}} \right) \Phi \left(\frac{\mu_{R_{i,j},j,G} - ln(y_{ij})}{\sigma_{R_{i,j},j,G}} \right)^{1-\delta_{i,j}}$$

- We can use R-package "flexsurv" to estimate parameters for arm R in j^{th} trial for population :

$$\begin{bmatrix} \hat{\mu}_{R,j,G} \\ \hat{\sigma}_{R,j,G} \end{bmatrix} \approx N\left(\begin{bmatrix} \mu_{R,j,G} \\ \sigma_{R,j,G} \end{bmatrix}, I_{R,J,G}^{-1} \right), I_{R,J,G}^{-1} = \Sigma_{R,J,G}$$

- The analytical form of $\Sigma_{R,J,G}$ can be derived according to Swan (1969)

METHODS

Mathematical model specifications:

- To estimate $\widehat{CV}_{j,G} = \hat{\sigma}_{R,j,G}/\hat{\mu}_{R,j,G}$, we follow the suggestion by Senior et al. (2020) to perform log-transformation to stabilize the ratio
  - point estimation with bias correction,

$$ln\left(\widehat{CV}_{R,j,G}\right) = ln\left(\hat{\sigma}_{R,j,G}\right) - \ln\left(\hat{\mu}_{R,j,G}\right) + \frac{1}{2}\frac{\Sigma_{R,j,G}(2,2)}{\hat{\sigma}^2_{R,j,G}} - \frac{1}{2}\frac{\Sigma_{R,j,G}(1,1)}{\hat{\mu}^2_{R,j,G}}$$

  - variance estimation

$$sd\left(ln\left(\widehat{CV}_{R,j,G}\right)\right) = \frac{\Sigma_{R,j,G}(2,2)}{\hat{\sigma}^2_{R,j,G}} + \frac{\Sigma_{R,j,G}(1,1)}{\hat{\mu}^2_{R,j,G}} - 2\frac{\Sigma_{R,j,G}(1,2)}{\hat{\sigma}_{R,j,G}\hat{\mu}_{R,j,G}}$$

  - these formulas are extensions of Senior et al. (2020) for censoring data
  - trial level statistics is $CVR_{j,G} = CV_{1,j,G}/CV_{0,j,G}$

METHODS

Choice of CVRs:

- The $ln\left(\widehat{CVR}_{j,G}\right) = log\left(\widehat{CV}_{1,j,G}\right) - log\left(\widehat{CV}_{0,j,G}\right)$ captures heterogeneity of treatment versus control in the j^{th} trial for population G.
- The probability of survival beyond human life span is greater than 0.
- To be relevant to human oncology trial, our analyses focus on RMST up to τ

$$E\{Y_{R,J,G}(\tau)\} = \int_0^{\tau} \Phi\left(\frac{\mu_{R,j,G} - ln(y)}{\sigma_{R,j,G}}\right) dy$$

$$V\{Y_{R,J,G}(\tau)\} = 2\int_0^{\tau} y\Phi\left(\frac{\mu_{R,j,G} - ln(y)}{\sigma_{R,j,G}}\right) dy - E^2\{Y_{R,J,G}(\tau)\}$$

- The $CVR_{j,G}(\tau) = \sqrt{V\{Y_{R,J,G}(\tau)\}}/E\{Y_{R,J,G}(\tau)\}$
- τ = 12, 24, 36, 60 months, and the max (observed follow-up time of all trials).

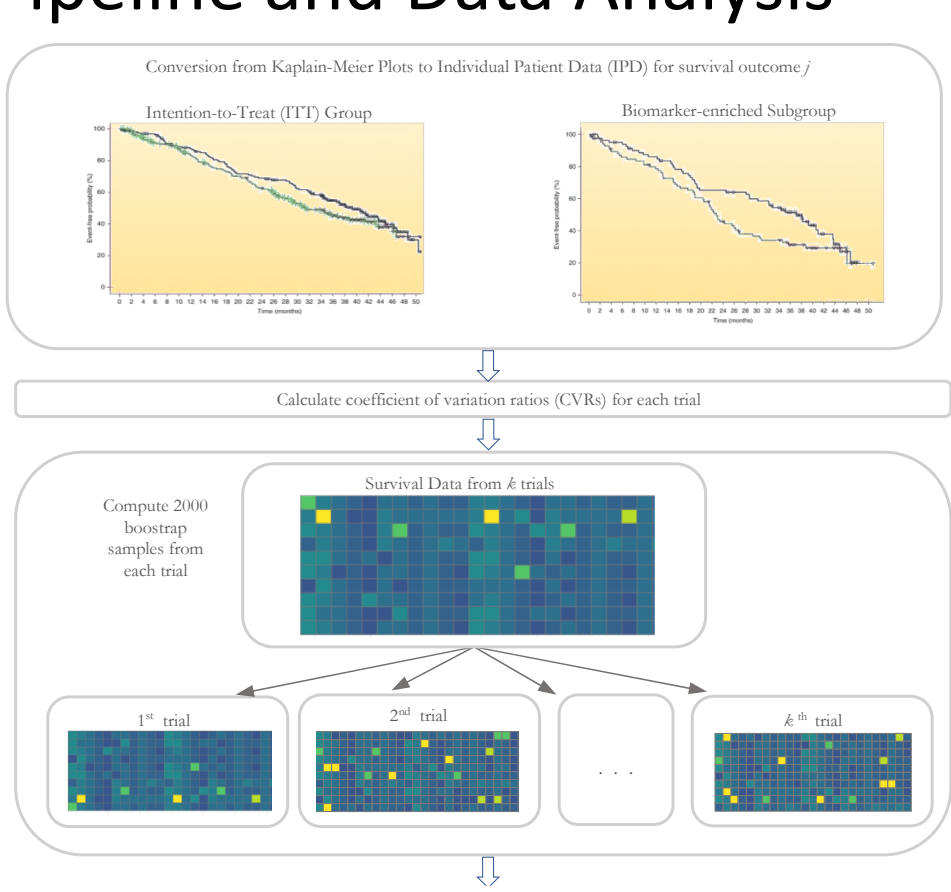
METHODS

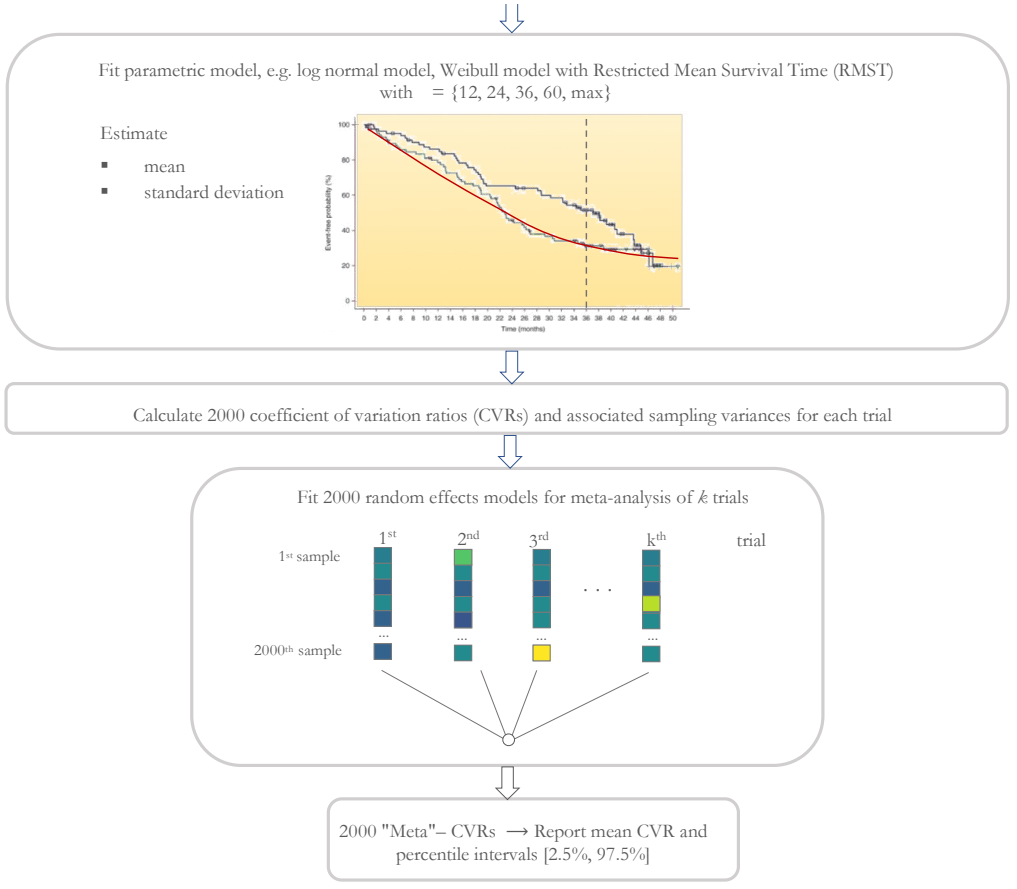
Meta CVR_G

- The meta $CVR_G(\tau) = \sum_{j=1}^{J} w_j CVR_{j,G}(\tau) / \sum_{j=1}^{J} w_j$
- w_j can be $1/V\left(CVR_{j,G}(\tau)\right)$ or other reasonable weights
- Bootstrap method is used to construct the 95% empirical confidence intervals

METHODS

Pipeline and Data Analysis





RESULTS

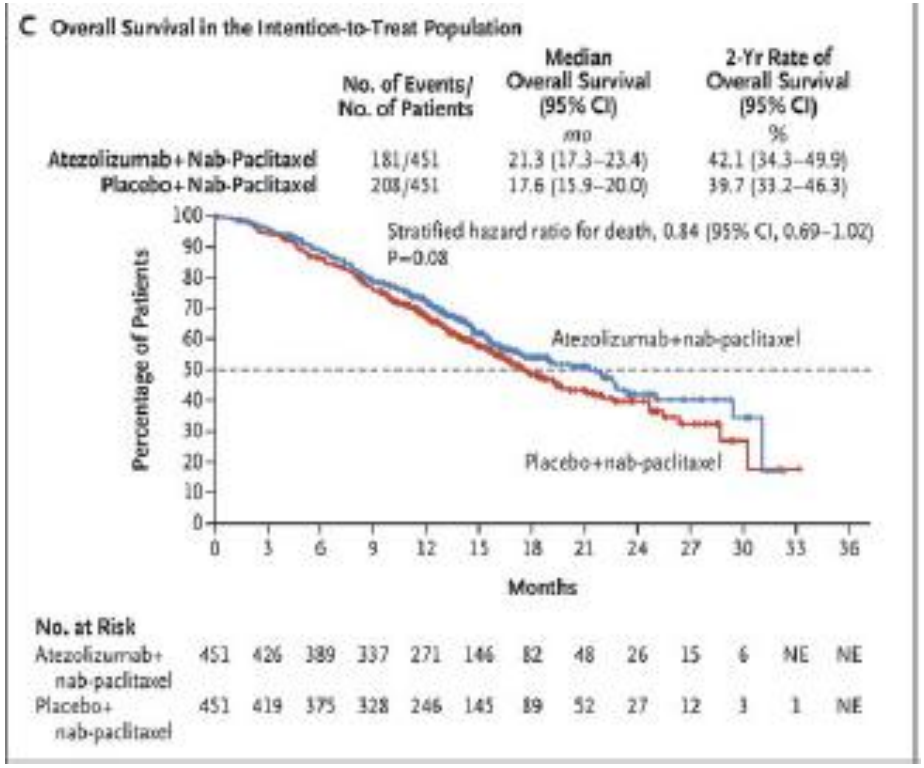
Trial Cohort

Name	Authors	Yr Pub	Trial Design			Trial Population		
			Treatment group	Control Group	Phase	N	Mean Age (yr)	Biomarker for stratification
SOLAR-1	F. Andre et al.	2020	Alpelisib plus fulvestrant	Placebo plus fulvestrant	III	341	-	PIK3CA
EMERALD	F-C. Bidard et al.	2022	Elacestrant	Endocrine therapy	III	477	63	ESR1
IMpassion 130	P. Schmid et al.	2018	Atezolizumab plus nab-Paclitaxel	Placebo plus nab-Paclitaxel	III	902	55	PD-(L)1
IMpassion 131	D. Miles et al.	2021	Paclitaxel plus atezolizumab	Placebo plus paclitaxel	III	651	55	PD-(L)1
Keynote-355	J. Cortes et al.	2022	Pembrolizumab plus chemotherapy	Placebo plus chemotherapy	III	847	52	PD-(L)1
SAFIRO2	T. Bachelot et al.	2021	Durvalumab	Maintenance chemotherapy	II	199	56	PD-(L)1
FAKTION	R. H. Jones et al.	2022	Fulvestrant plus capivasertib	Placebo	II	140	62	PIK3CA/ PTEN status

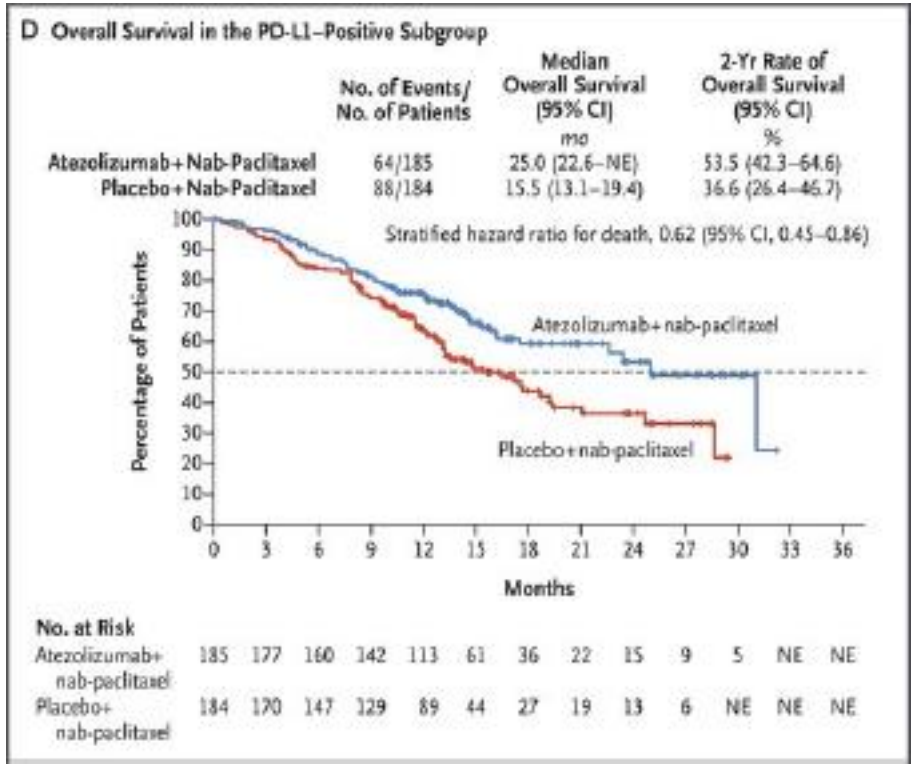
RESULTS

Survival curves from trials (Impassion 130 trial)

ITT Population

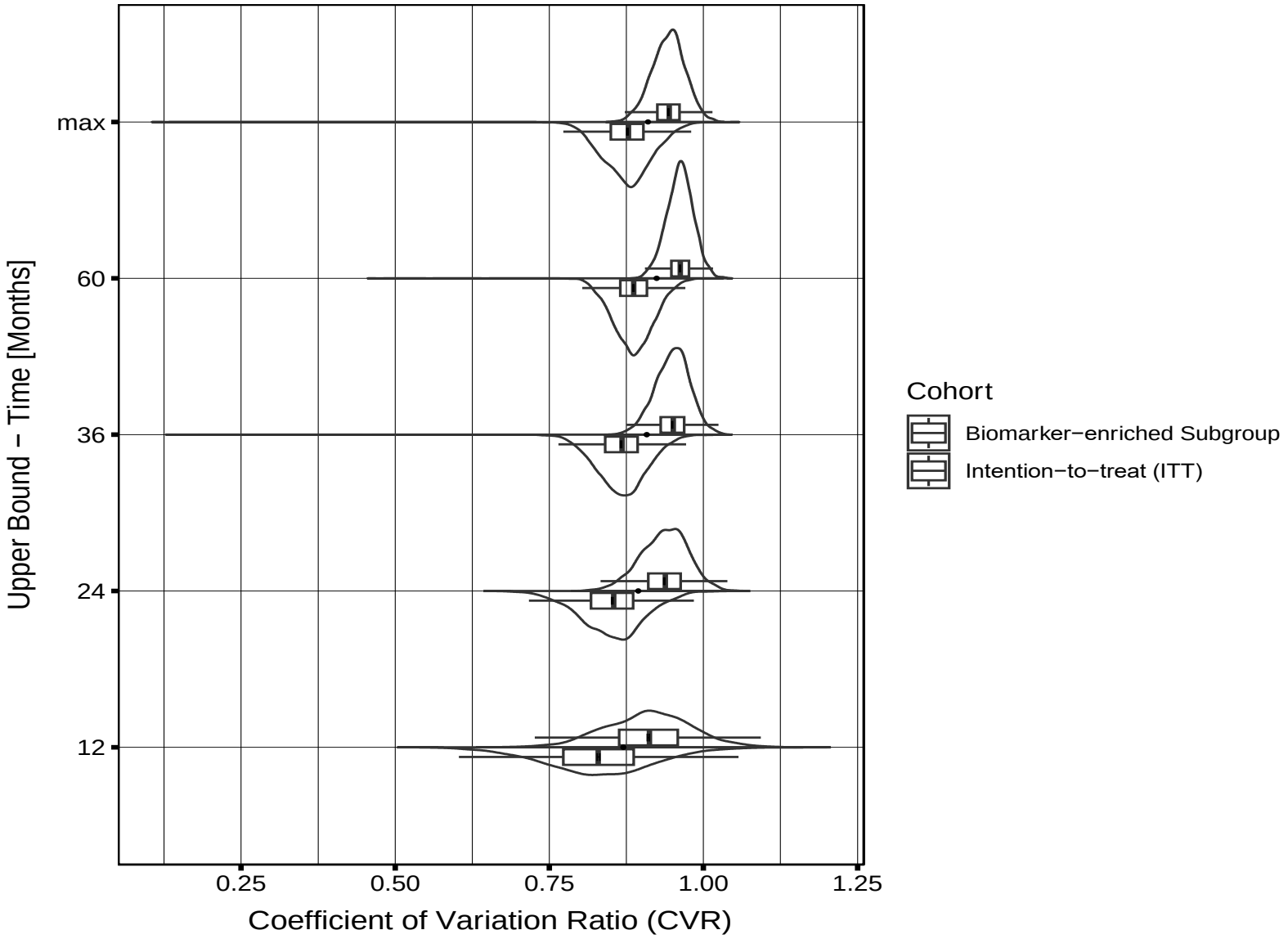


Marker "+" Population



RESULTS

1. Treatment arm always had smaller coefficient of variation than the placebo arm.
2. Biomarker-enriched subgroup always had smaller CV than the IT population.
3. The precision of CVRs improved with longer τ.



RESULTS

CVRs from individual trials for OS with τ=36 months.

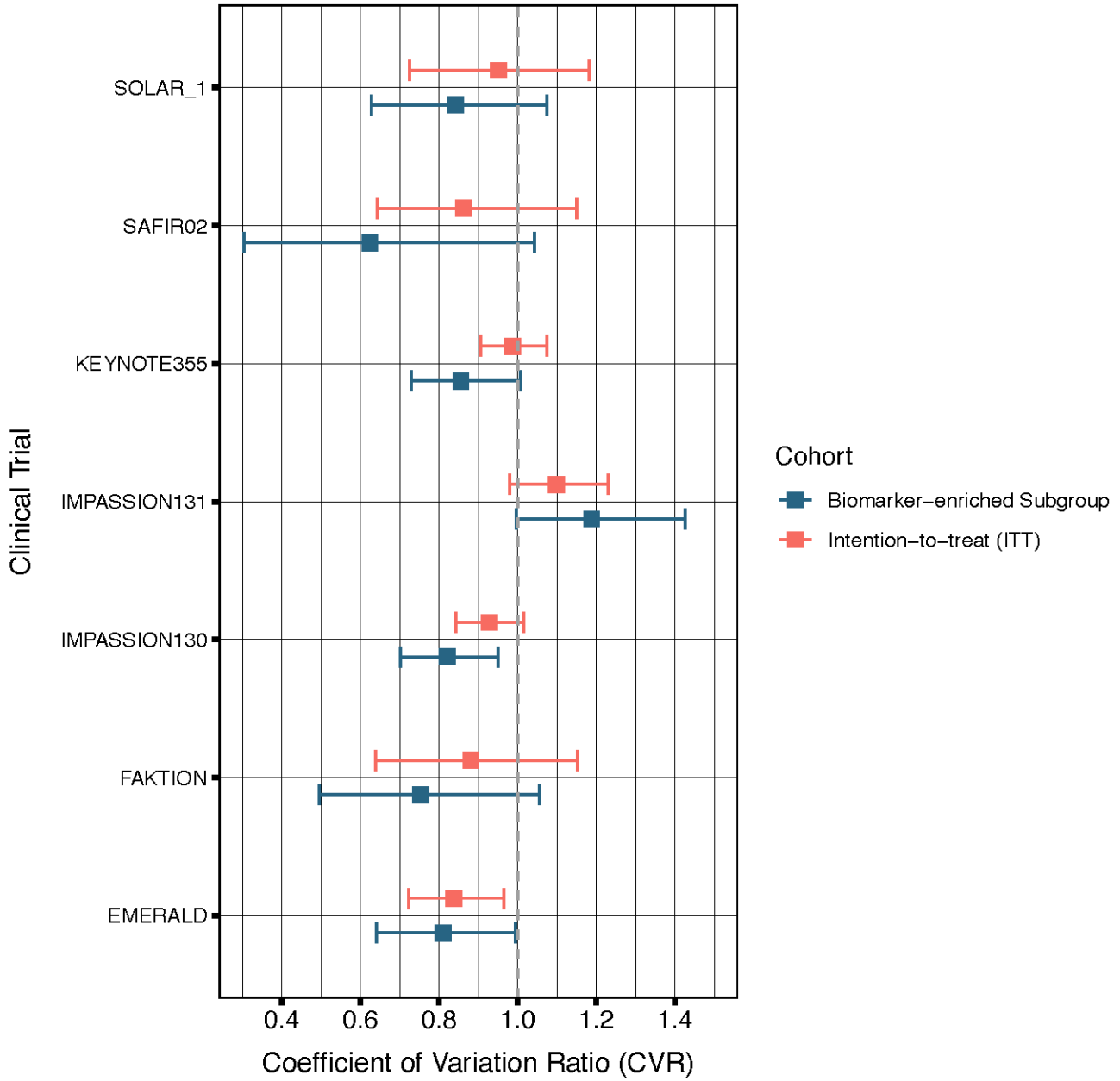
The aggregated CVR

ITT:

0.950 [0.892, 1.002]

Biomarker enriched subgroup:

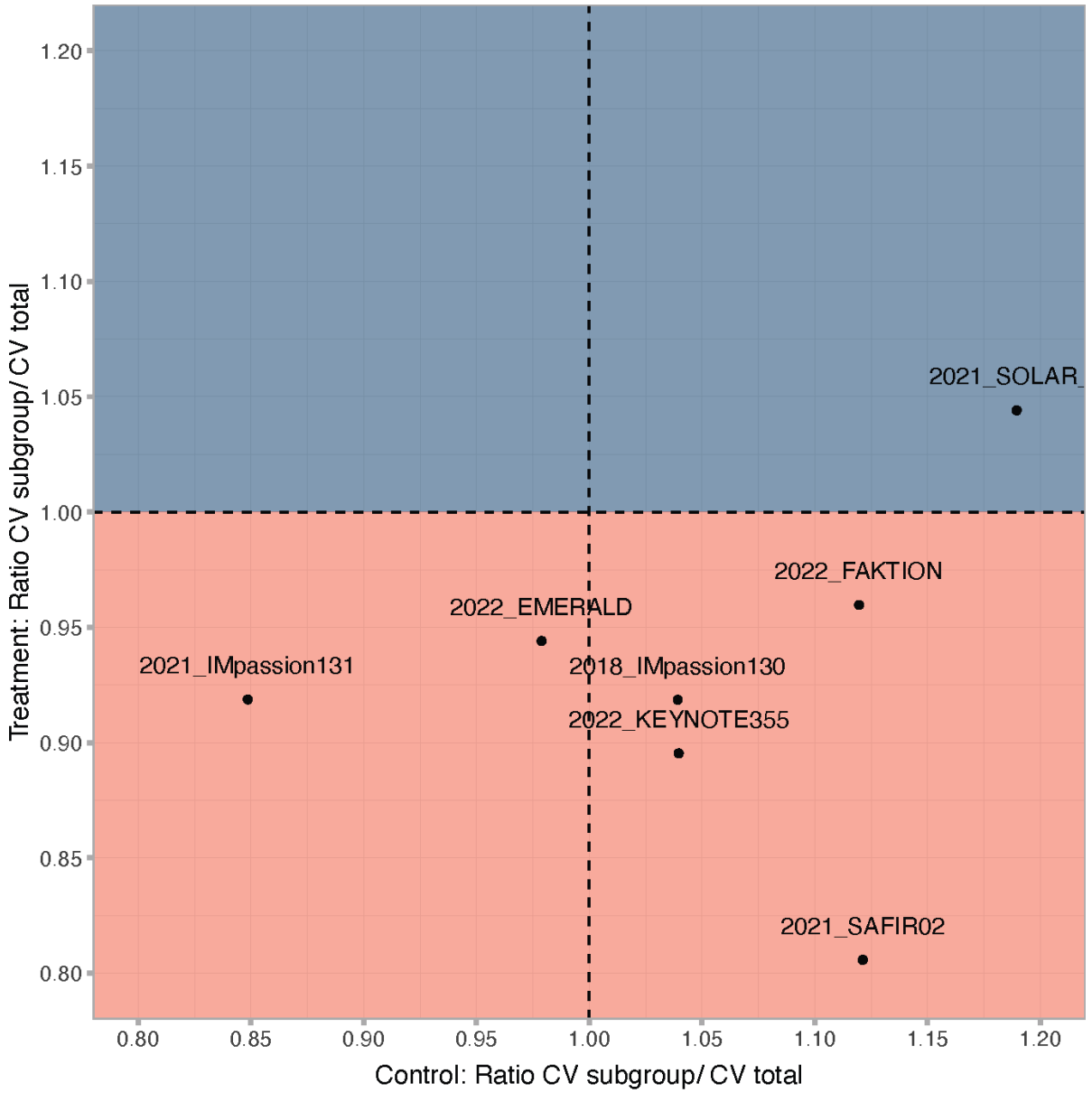
0.865 [0.790, 0.945]



RESULTS

Observed CVR within arms

1. Biomarker-enriched subpopulation most likely to have smaller CVs than ITT population (7 out of 8 below 1).
2. Biomarker-enriched subpopulation have similar or higher CVs than ITT population (more above 1).
3. IMpassion131 is the only trial that is above the identical line.



DISCUSSION AND CONCLUSION

- We developed a feasible approach to construct coefficient of variation ratios (CVRs) for survival outcomes using Kaplan-Meier curves from precision oncology trials.
- We identified seven breast cancer trials with a total of 3,557 participants and successfully reconstructed individual patient data (IPD) for treatment and control groups and two outcomes (OS and PFS).
- Preliminary results show a trend towards lower coefficient of variation ratios for biomarker-enriched subgroups comparing to their respective intent-to-treat cohorts, suggesting more homogeneous survival response in biomarker-enriched subpopulation.

DISCUSSION AND CONCLUSION

- Examination of other tumor types
  - Lung cancer, prostate cancer, lymphoma, skin cancer
- Regression model: association of trial characteristics with CVRs
- Translation of lower coefficient of variation ratios (CVRs) to commonly used measures for clinical benefits
  - e.g., proportion of patients survived beyond τ
- Impact of choice of parametric distributions
  - e.g., test the Weibull distribution or other distributions
- Comparison of alternative measure of heterogeneity
  - IQR/Median
- Alternative modeling based on proportional hazard treatment effect

REFERENCES

1. Nakagawa et al.: 2015, Meta-analysis of variation: ecological and evolutionary applications and beyond, Methods in Ecology and Evolution 6(2), 143-152.
2. Mukhopadhyay et al.: 2022, Log-rank test vs MaxCombo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in immuno-oncology trials, JAMA Oncol. 8(9):1294-1300.
3. Plana et al.: 2022, Cancer patient survival can be parametrized to improve trial precision and reveal time-dependent therapeutic effects, Nature Communications 13:873.
4. Redd R, et al.: 2022, kmdata: A Database of Reconstructed Individual Patient Level Data from Oncology Trials, R package version 1.0.1.
5. Senior et al.: 2020, Revisiting and expanding the meta- analysis of variation: The log coefficient of variation ratio, Research Synthesis Methods 11(4), 553-567.
6. Swan: 1969, Computing maximum-likelihood estimates for parameters of the normal distribution from grouped and censored data, JRSC series C, 18(1), 65-69.
7. Viechtbauer: 2010, Conducting meta-analysis in R with metafor package, Journal of Statistical Software, 36(3), 1-48.
8. Winkelbeiner et al.: 2019, Evaluation of differences in individual treatment response in schizophrenia spectrum disorders: a meta-analysis, JAMA psychiatry 76(10, 1063-1073.

Acknowledgement

- Significant contributions from students and collaborators:
  - 
    Max Schueessler, MD, PhD Student in Biomedical Informatics, Department of Biomedical Data Science, Stanford University
  - 
    Maike Hohberg, PhD, Department of Medical Statistics, University Medical Center Göttingen
  - 
    Pascal Geldsetzer, MD, PhD, MPH, Department of Primary Care and Population Health, Stanford University
  - Elizaveta Skarga, MD Candidate, Heidelberg Institute of Global Health, Heidelberg University
- My effort is supported by the Stanford Cancer Institute (NCI 4P30CA124435)

Thank you very much for your attention

QUESTIONS?

Restricted mean time in favor of treatment

A fine-grained analysis

Lu Mao, PhD

Department of Biostatistics and Medical Informatics

University of Wisconsin-Madison

Joint work with Tuo Wang

SCT

SOCIETY FOR CLINICAL TRIALS

School of Medicine and Public Health

UNIVERSITY OF WISCONSIN–MADISON

Disclosure Statement

No disclosures

Introduction – RMT-IF

- **Restricted mean time in favor (RMT-IF)** (Mao, 2023, *Biometrics*)
  - A new effect-size estimand for *hierarchical* composite endpoints
  - Net average time a treated has a better outcome than an untreated in a fixed time window
    - Remission → relapse → metastasis → death
    - Event-free → hospitalized once → hospitalized twice → ⋯ → death
  - Singular survival endpoint: = difference in restricted mean survival time (RMST) (Royston & Parmar, 2011; Tian et al., 2018; McCaw et al., 2019)
- **Stagewise effects**
  - Additive decomposition of overall effect
    - (Relapse vs remission) + (metastasis vs non-metastatic) + (death vs survival)

extra RMST

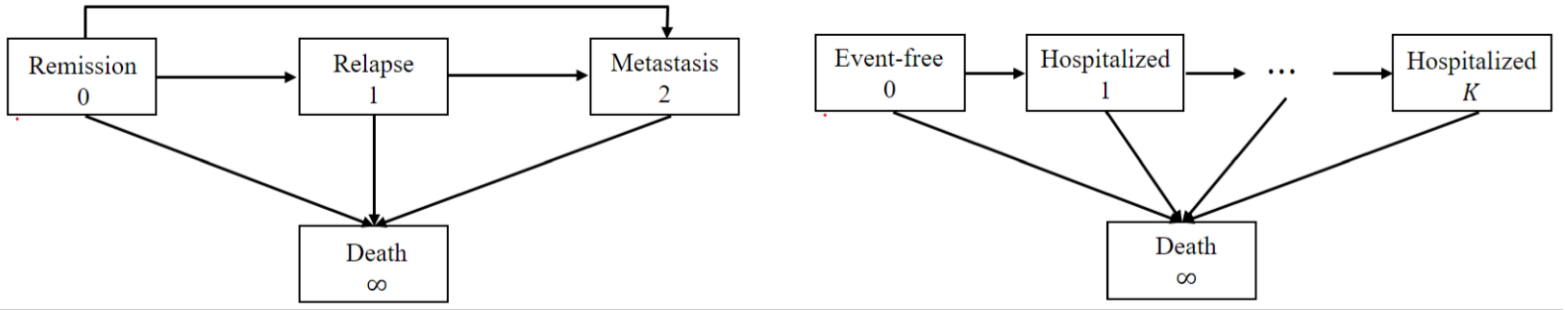
  - Extra time gained before each stage of disease progression
  - Used in secondary analysis of component-wise effects

Introduction – RMT-IF

- **Limitations**
  - Inadequacy of stage-wise composition
    - Extra pre-metastatic time gained -- in remission? post-relapse?
    - Extra lifetime gained -- in remission? post-relapse? post-metastasis?
  - Joint tests unavailable of decomposed units
    - May be more sensitive to component-specific changes “under the hood”
- **A fine-grained analysis** (Mao & Wang, 2023, *J Biopharm Stats*)
  - Decompose stagewise effects (main components) further down to granular level (subcomponents)
  - Construct multi-degree-of-freedom (df) tests based on main and subcomponents

Methods – Review

- **Overall estimand**
  - $Y^{(a)}(t)$: Multistate outcome of generic patient in group a (1: active treatment; 0: control)
    - 0: initial event-free status; $1, \ldots, K$: increasingly adverse states; ∞: death



  - τ-RMT-IF ($\tau > 0$)

$$\mu(\tau) = \underbrace{E\left[\int_0^\tau I\{Y^{(1)}(t) < Y^{(0)}(t)\}\,\mathrm{d}t\right]}_{\text{Avg. time treatment favorable}} - \underbrace{E\left[\int_0^\tau I\{Y^{(0)}(t) < Y^{(1)}(t)\}\,\mathrm{d}t\right]}_{\text{Avg. time control favorable}}$$

Methods – Review

- **Stagewise decomposition**

$$\mu(\tau) = \sum_{k=1}^{K,\infty} \mu_k(\tau)$$

where $\mu_k(\tau) = E\left[\int_0^\tau I\{Y^{(1)}(t) < Y^{(0)}(t) = k\}\,\mathrm{d}t\right] - E\left[\int_0^\tau I\{Y^{(0)}(t) < Y^{(1)}(t) = k\}\,\mathrm{d}t\right]$

- $\mu_1(\tau)$: extra pre-relapse time gained (as opposed to relapse)
- $\mu_2(\tau)$: extra pre-metastasis time gained (as opposed to metastasis)
  - In remission (0) or post-relapse (1)
- $\mu_\infty(\tau)$: extra life time gained (= net RMST)
  - In remission (0), post-relapse (1), post-metastasis (2)

Methods – Subcomponents

- **Further decomposition**

$$\mu_k(\tau) = \sum_{j<k} \mu_{jk}(\tau)$$

where $\mu_{jk}(\tau) = E\left[\int_0^\tau I\{Y^{(1)}(t) = j, Y^{(0)}(t) = k\}\,\mathrm{d}t\right] - E\left[\int_0^\tau I\{Y^{(0)}(t) = j, Y^{(1)}(t) = k\}\,\mathrm{d}t\right]$

  - Net average pre-state k time gained in state j
    - $\mu_{02}(\tau), \mu_{12}(\tau)$: net average pre-metastasis time gained in remission and post-relapse, respectively
    - $\mu_{0,\infty}(\tau), \mu_{1,\infty}(\tau), \mu_{2,\infty}(\tau)$: net average survival time gained in remission, post-relapse, and post-metastasis, respectively

Methods – Subcomponents

Losing \ Winning	1	2	⋯	K	∞
	$\mu_1(\tau)$	$\mu_2(\tau)$	⋯	$\mu_K(\tau)$	$\mu_\infty(\tau)$
0	$\mu_{01}(\tau)$	$\mu_{02}(\tau)$		$\mu_{0K}(\tau)$	$\mu_{0,\infty}(\tau)$
1		$\mu_{12}(\tau)$	⋯	$\mu_{1K}(\tau)$	$\mu_{1,\infty}(\tau)$
⋮				⋮	⋮
$K-1$				$\mu_{K-1,K}(\tau)$	$\mu_{K-1,\infty}(\tau)$
K					$\mu_{K,\infty}(\tau)$

A graphical dissection of $\mu(\tau) = \sum_{k=1}^{K,\infty} \mu_k(\tau) = \sum_{k=1}^{K,\infty} \sum_{j<k} \mu_{jk}(\tau)$

Methods – Re-expression

- **Time-to-event formulation**: suppose $Y^{(a)}(t) \leq Y^{(a)}(s)$ for all $t \leq s$ (progressivity)
  - $T_k^{(a)} = \inf\{t: Y^{(a)}(t) \geq k\}$: Time of transition into state k or higher
  - $Y^{(a)}(\cdot)$ completely determined by $T_1^{(a)}, \ldots, T_K^{(a)}, T_\infty^{(a)} \equiv T_{K+1}^{(a)}$
  - State probabilities ↔ survival functions $S_k^{(a)}(t) = \text{pr}\left(T_k^{(a)} > t\right)$ $(k = 1, \ldots, K+1)$
    - $Y^{(a)}(t) = k \Leftrightarrow T_k^{(a)} \leq t < T_{k+1}^{(a)}$
    - $\text{pr}\left(Y^{(a)}(t) = k\right) = S_{k+1}^{(a)}(t) - S_k^{(a)}(t)$

Methods – Estimation

- **Re-expression** of subcomponents ($\mu_{j,K+1}(\tau) \equiv \mu_{j,\infty}(\tau)$)
  - $\mu_{jk}(\tau) = \int_0^\tau \mathrm{pr}\{Y^{(1)}(t) = j, Y^{(0)}(t) = k\}\,\mathrm{d}t - \int_0^\tau \mathrm{pr}\{Y^{(0)}(t) = j, Y^{(1)}(t) = k\}\,\mathrm{d}t$

$$= \int_0^\tau \left(S_{j+1}^{(1)}(t) - S_j^{(1)}(t)\right)\left(S_{k+1}^{(0)}(t) - S_k^{(0)}(t)\right)\mathrm{d}t - \int_0^\tau \left(S_{j+1}^{(0)}(t) - S_j^{(0)}(t)\right)\left(S_{k+1}^{(1)}(t) - S_k^{(1)}(t)\right)\mathrm{d}t$$

- **Observed data**: $\left(X_k^{(a)}, \delta_k^{(a)}\right)$ $(k = 1, \ldots, K+1)$
  - $X_k^{(a)} = \min\left(T_k^{(a)}, C^{(a)}\right), \delta_k^{(a)} = I\left(T_k^{(a)} \le C^{(a)}\right), C^{(a)}$: random censoring time
- Nonparametric estimator $\hat{\mu}_{jk}(\tau)$: plug in Kaplan-Meier estimator $\hat{S}_k^{(a)}(t)$
  - Robust variance derived by functional delta method

Methods – Joint Tests

- **Three types of tests**
  - Overall

$$H_0: \mu(\tau) = 0$$

    based on $\hat{\mu}(\tau)$: χ_1^2

  - Main components

$$H_0: \mu_1(\tau) = \cdots = \mu_\infty(\tau) = 0$$

    based on $\{\hat{\mu}_1(\tau), \ldots, \hat{\mu}_\infty(\tau)\}^{\mathrm{T}}$: χ_{K+1}^2

  - Subcomponents

$$H_0: \mu_{01}(\tau) = \mu_{02}(\tau) = \mu_{12}(\tau) = \cdots = \mu_{K,\infty}(\tau) = 0$$

    based on $\left\{\hat{\mu}_{01}(\tau), \hat{\mu}_{02}(\tau), \hat{\mu}_{12}(\tau), \cdots, \hat{\mu}_{K,\infty}(\tau)\right\}^{\mathrm{T}}$: $\chi_{(K+1)(K+2)/2}^2$

Methods – Recurrent Events

- **Recurrent events and death**
  - $Y^{(a)}(t) = 0,1, \ldots K$: cumulative number of recurrent events (K potentially large)
  - Re-bundling the subcomponents
    - $\mu_{0,\infty}(\tau)$: extra life time gained event-free
    - $\mu_{1+,\infty}(\tau) = \sum_{j=1}^{K} \mu_{j,\infty}(\tau)$: extra life time gained having experienced at least one nonfatal event
    - $\mu_{0,1+}(\tau) = \sum_{k=1}^{K} \mu_{0,k}(\tau)$: extra time gained event-free in the living
    - $\mu_{1+,R}(\tau) = \sum_{k=2}^{K} \sum_{j=1}^{k-1} \mu_{j,k}(\tau)$: extra time gained with fewer but nonzero events in the living

$$\mu(\tau) = \underbrace{\mu_{0,1+}(\tau) + \mu_{1+,R}(\tau)}_{\mu_R(\tau)} + \underbrace{\mu_{0,\infty}(\tau) + \mu_{1+,\infty}(\tau)}_{\mu_\infty(\tau)}$$

Software – R-Package `rmt`

Package 'rmt'

May 25, 2021

Type Package

Title Restricted Mean Time in Favor of Treatment

Version 1.0

Author Lu Mao

Maintainer Lu Mao <lmao@biostat.wisc.edu>

`rmtfit` *Estimate restricted mean times in favor of treatment*

Usage

```
rmtfit(id, time, status, trt, type = "multistate", ...)
```

Arguments

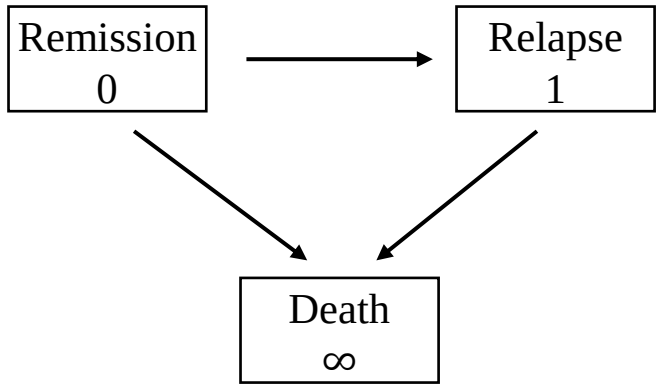
`id`	A vector of id variable.
`time`	A vector of follow-up times.
`status`	For `type="multistate"`, k = entering into state k ($K+1$ represents death) and 0 = censoring; For `type="recurrent"`, 1 = recurrent event, 2 = death, and 0 = censoring;
`trt`	A vector of binary variable for treatment group.
`type`	`"multistate"` = multistate data; `"recurrent"` = recurrent event data.

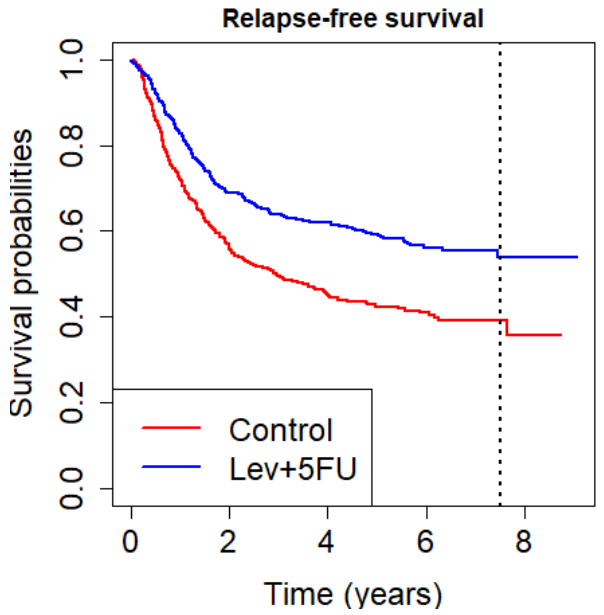
Software – R-Package `rmt`

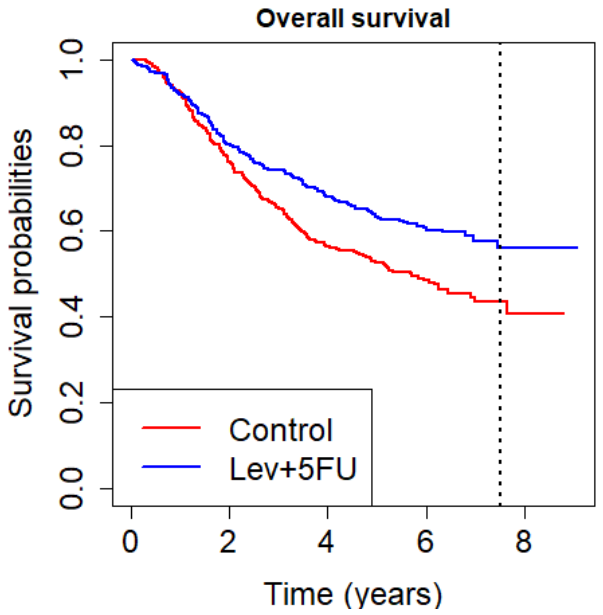
```
> obj <- rmtfit(id, time, status, trt)
> obj_sub <- dissect(obj, tau = 3.0)
> obj_sub
Call:
rmtfit.default(id = id, time = time, status = status, trt = trt)

Restricted mean time in favor of group "1" by time tau = 3:
                 Estimate   Std.Err Z value  Pr(>|z|)
Overall          0.598838  0.189976  3.1522 0.0016206 **
Death            0.174255  0.140569  1.2396 0.2151084
    vs State 0   0.135488  0.064662  2.0953 0.0361410 *
    vs State 1   0.063326  0.050268  1.2598 0.2077573
    vs State 2  -0.024559  0.054056 -0.4543 0.6495950
State 2          0.295115  0.088653  3.3289 0.0008720 ***
    vs State 0   0.215703  0.059633  3.6172 0.0002978 ***
    vs State 1   0.079412  0.040621  1.9549 0.0505891 .
State 1 (vs 0)   0.129468  0.055798  2.3203 0.0203251 *

Overall chi-square test:
X-squared = 9.55291, df = 1, p-value = 0.002;
Joint chi-square test on main components:
X-squared = 13.53824, df = 3, p-value = 0.0036;
Joint chi-square test on subcomponents:
X-squared = 20.78061, df = 6, p-value = 0.002.
```

Real Examples – A Colon Cancer Trial

- A historic colon cancer trial (Moertel et al., 1990)
  - Population: 619 patients with stage C disease
  - Treatment arms:
    - Levamisole + fluorouracil (*n=304*)
    - Observation (Control) (*n=315*)







$\tau = 7.5$ years

RMEST=12.3 months (p<0.001)

RMST=7.4 months (p=0.004)

Real Examples – A Colon Cancer Trial

Analysis of the colon cancer trial by the RMT-IF (months) of combined treatment.

	$\tau = 2.5$ years			$\tau = 5.0$ years			$\tau = 7.5$ years		
	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value
Death	0.6	0.6	0.321	3.6	1.5	0.018	7.4	2.6	0.004
vs Remission	0.8	0.5	0.108	4.1	1.4	0.003	8.1	2.3	<0.001
vs Relapse	−0.2	0.1	0.069	−0.5	0.3	0.150	−0.7	0.5	0.155
Relapse	2.1	0.4	<0.001	3.4	0.7	<0.001	4.2	0.9	<0.001
Overall	2.7	0.8	0.001	7.1	1.9	<0.001	11.6	3.0	<0.001

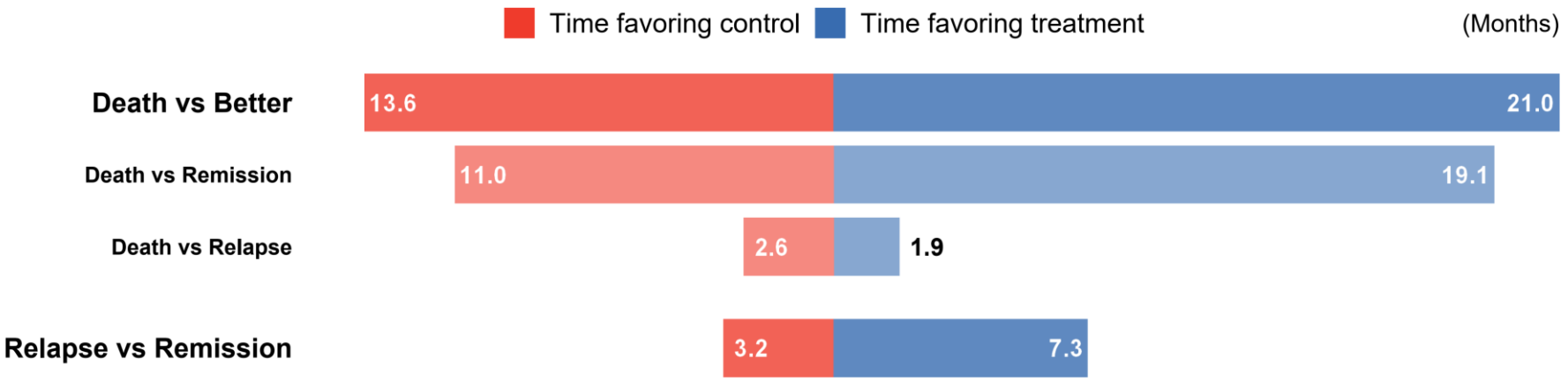
Note: Est, estimate; SE, standard error.

- By 7.5 years, combined treatment gains the patient $\mu(\tau)$ =11.6 months extra time in a more favorable state
  - $\mu_\infty(\tau) = 7.4$ months extra survival time, driven by gains in remission ($\mu_{0,\infty}(\tau) = 8.1$ vs $\mu_{1,\infty}(\tau) = -0.7$)
  - $\mu_1(\tau) = 4.2$ months extra pre-relapse time in the living

Real Examples – A Colon Cancer Trial

- **Favorability plot** ($\tau = 7.5$ years)

```
ggrmtif(obj_sub, unit = "months")
```



  - Joint tests (χ_2^2 and χ_3^2) more significant than overall test (χ_1^2) – heterogeneous component-wise effects

Real Examples – A Cardiovascular Trial

- HF-ACTION (O'Connor et al., 2009)
  - Endpoints: recurrent hospitalizations and death

Summary statistics for the HF-ACTION study subgroup.

		Usual care ($N = 221$)	Exercise training ($N = 205$)	Overall ($N = 426$)
Age	($\leq$60 years)	122 (55.2%)	128 (62.4%)	250 (58.7%)
	(>60 years)	99 (44.8%)	77 (37.6%)	176 (41.3%)
Follow-up	(months)	28.6 (18.4, 39.3)	27.6 (19.0, 40.2)	28 (18.7, 39.5)
Death		57 (25.8%)	36 (17.6%)	93 (21.8%)
Freq. Hosp.	0	51 (23.1%)	60 (29.3%)	111 (26.1%)
	1–3	114 (51.6%)	102 (49.8%)	216 (50.7%)
	4–10	49 (22.2%)	39 (19%)	88 (20.7%)
	>10	7 (3.2%)	4 (2%)	11 (2.6%)

Categorical variables are summarized by $N(\%)$ and continuous variables by median (inter-quartile range).

Real Examples – A Cardiovascular Trial

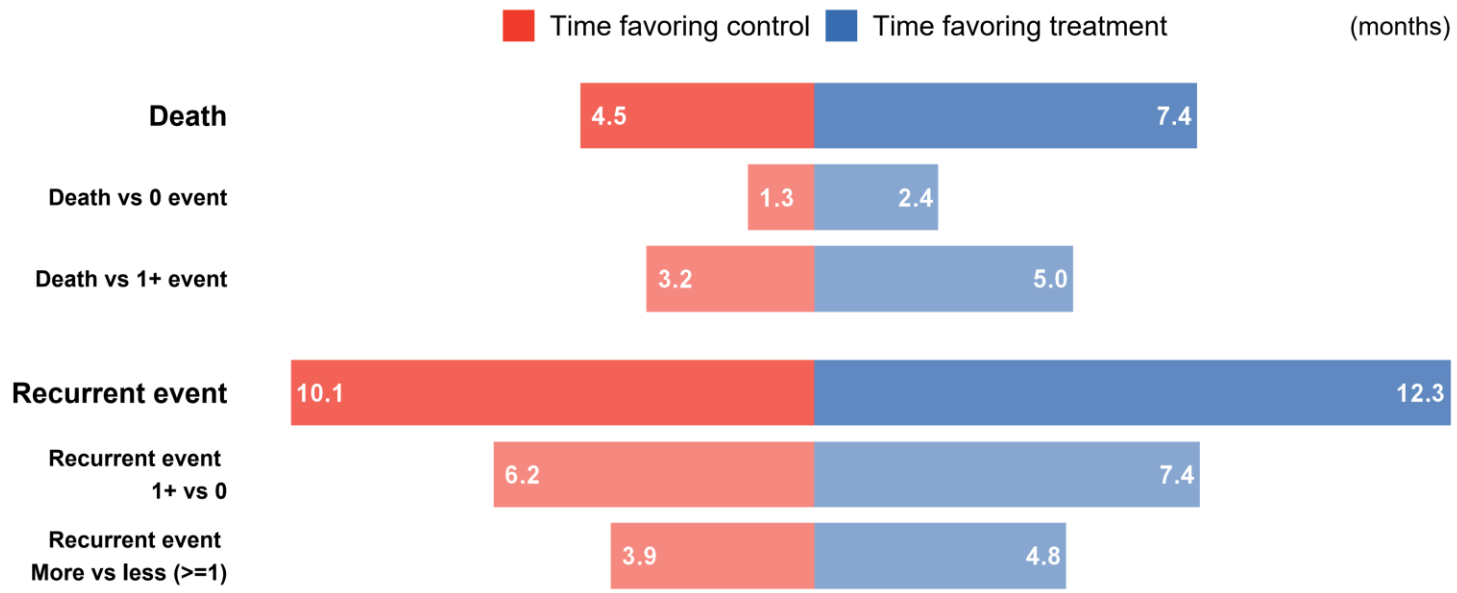
- Analysis of the HF-ACTION trial by the RMT-IF (months) of exercise training ($\tau = 4$ years)

	Estimate	**SE**	***p*-value**
Death	2.9	1.37	0.037
vs Event-free	1.1	0.52	0.032
vs 1+ events	1.8	0.99	0.076
Recurrent events	2.2	1.56	0.161
1+ vs 0	1.3	1.20	0.314
More vs less (≥ 1)	0.9	0.80	0.215
Total	5.1	2.13	0.018

- χ_2^2 test on main components: *p*-value 0.039; χ_4^2 test on subcomponents: *p*-value 0.173
  - Less significant than overall χ_1^2 with *p*-value 0.018 (component-wise effects similar)

Real Examples – A Cardiovascular Trial

- Favorability plot ($\tau = 4$ years)



Summary & Conclusion

- **Further decomposition** of estimand
  - State-to-state comparison: where the gained time is spent
  - Provides finer details to treatment mechanism
  - Useful in secondary analysis
- **Joint tests** of decomposed units
  - Superior to χ_1^2 test with variable component-wise effect sizes
  - Pre-specify which test to use in protocol
- **Software**
  - R-package `rmt`

https://cran.r-project.org/package=rmt

Acknowledgments

- This research was supported by
  - NIH-NHLBI grant **R01HL149875**

    Novel Statistical Methods for Complex Time-to-Event Outcomes in Cardiovascular Clinical Trials
- HF-ACTION study data are provided by BioLINCC of NHLBI

References

- Mao, L. (2023). On restricted mean time in favor of treatment. *Biometrics*, 79, 61–72.
- Mao, L. and Wang, T. (2023). Dissecting the restricted mean time in favor of treatment, *Journal of Biopharmaceutical Statistics*, doi: 0.1080/10543406.2023.2210658.
- McCaw, Z. R., G. Yin, and L.-J. Wei (2019). Using the restricted mean survival time difference as an alternative to the hazard ratio for analyzing clinical cardiovascular studies. *Circulation*, 140, 1366–1368.
- Moertel, C.G., Fleming, T.R., Macdonald, J.S., Haller, D.G., Laurie, J.A., Goodman, P.J. et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, 322, 352–358.
- O'Connor, C. M., Whellan, D. J., Lee, K. L., Keteyian, S. J., Cooper, L. S., Ellis, S. J. et al. (2009) Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *Journal of the American Medical Association*, 301, 1439—1450.
- Royston, P. and Parmar, M.K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30, 2409–2421.
- Tian, L., H. Fu, S. J. Ruberg, H. Uno, and L.-J. Wei (2018). Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics* 74, 694–702.