On testing based on restricted mean survival time for time-to-event outcomes

Hajime Uno, PhD
Dana-Farber Cancer Institute
Harvard Medical School

Joint work with
Miki Horiguchi
(Kitasato University, Tokyo)

Annual Conference of the ISCB at Vigo, Spain (Jul 10, 2017)
Consider

Randomized clinical trial to compare two groups
A time-to-event outcome is the primary endpoint
What we need to do

- Testing equality of the two event time distributions (Test)

  *ex. “The survival benefit of this drug is highly significant!!”*

- Estimating magnitude of the treatment effect (0.95CI for a primary summary measure)

Important information for decision-making
Test-estimation coherency

Testing

Test result is SIG

Test result is NS

Estimation

CI of the treatment effect excludes the null value (e.g., HR=1)

Includes the null value

Example of test-estimation coherency

Test: Logrank test

Estimation of the treatment effect:

Cox's estimator (Hazard ratio)
Outline

Â Issues of the conventional practices
Â Alternatives and challenge
Â Restricted mean survival time (RMST) based versatile test
Â Application
Â Numerical studies
Â Conclusions
A conventional practice

Description → Test → Estimation of treatment effect

Kaplan-Meier → Logrank test → HR by Cox PH
A conventional practice

Logrank test is

- the most powerful nonparametric test in PH
- equivalent to the score test for testing HR equals 1 via the Cox's model

(test-estimation coherency)

However, the proportional hazards (PH) assumption does not always hold in practice, which raises some problems
What if non-PH?

**Test:** Logrank is not optimal anymore (may perform rather poorly, e.g., cross-hazard cases)

**Estimation of treatment effect:**

- Interpretation of HR is not obvious as a quantitative summary of the treatment effect
  - HR is not a simple average of the hazard ratio over time
  - HR depends on underlying study-specific censoring distributions (or follow-up time)

- Thus, it may not be useful as the primary summary of the treatment effect for decision making
A conventional practice when non-PH

Test
- Another member of the weighted logrank tests (e.g., Wilcoxon test) is often chosen

Estimation of treatment effect
- Difference in median survival time (without CI) is often reported as the quantitative information of treatment effect

This conventional practice can also be problematic sometimes!
Metastatic breast cancer example

- N=371 (182 on exemestane, 189 on tamoxifen)
- The primary endpoint was PFS
"The 4.1-month difference in median PFS between treatment arms was statistically significant using the Wilcoxon test \((P = .028)\)" Paridaens et al. (2008, JCO)

This is not a correct statement statistically
How does the conventional design and analysis fail?
# Metastatic breast cancer example

Paridaens et al. (2008, JCO)

| Design stage | 
|---|---|
| median PFS | 10.0 mon 7.14 mon |
| Pattern of difference | PH |
| HR | 0.78 |
| Number of events | 278 |
| sample size | 342 |
| The primary test | logrank |
Metastatic breast cancer example
Paridaens et al. (2008, JCO)

<table>
<thead>
<tr>
<th></th>
<th>Design stage</th>
<th>It turned out</th>
</tr>
</thead>
<tbody>
<tr>
<td>median PFS</td>
<td>10.0 mon</td>
<td>9.9 mon</td>
</tr>
<tr>
<td></td>
<td>7.14 mon</td>
<td>5.8 mon</td>
</tr>
<tr>
<td>Pattern of difference</td>
<td>PH</td>
<td>Non-PH</td>
</tr>
<tr>
<td>HR</td>
<td>0.78</td>
<td>0.84 ?</td>
</tr>
<tr>
<td>Number of events</td>
<td>278</td>
<td>319</td>
</tr>
<tr>
<td>sample size</td>
<td>342</td>
<td>371</td>
</tr>
<tr>
<td>The primary test</td>
<td>logrank</td>
<td>p=0.121</td>
</tr>
</tbody>
</table>
Why did this happen?

Failed to guess the pattern of the difference

- Power depends on the underlying true difference between the two survival functions
- No or little information is usually available at design stage

This is a common challenge in many clinical trials
Solution and problem

Versatile tests can capture various patterns of difference between two survival curves

Examples:
- Max/liner comb of weighted logrank tests (e.g., Tarone (1981))
- Adaptively weighted logrank test (Yang and Prentice, 2010)
- Adaptively weighted KM-based test (Uno et al. 2015)
- Min of logrank p-val and RMST permutation test p-val (Royston and Parmer 2016)
Solution and problem

However, a problem with most of the versatile tests is that test-estimation coherency will be a challenge.
Proposal

RMST-based versatile test
Restricted Mean Survival Time (RMST)

\[ = E_{x \min} E, F(z) = \dot{z} \]

Exemestane = 24 mon

Tamoxifen = 10.0 mon

11.5 mon
robust, clinically interpretable, and model-free summary measure

**Standard RMST-based test**

- Using a pre-specified fixed
- Based on the test statistic,

\[ (\cdot) = \hat{\delta}(\cdot) / \hat{\sigma}(\cdot) \]

\( \hat{\sigma}(\cdot) \): standard error estimate of \( \hat{\delta}(\cdot) \)
Proposed RMST-based versatile test

- Uses data-dependent
  - can detect various patterns of the difference

- Has companion quantification procedures to provide a corresponding, robust, clinically interpretable summary of the treatment effect
  - test-estimation coherency
Details of the proposed test

i. Instead of choosing a fixed \( \tau \), we consider a set of \( \tau \), \( \tau = \{ \tau_0, \tau_1, \tau_2, \ldots \} \)

ii. For each \( \tau \), we can calculate the test statistic, \( \tau' E \tau F = \tau'(\tau) / \tau(\tau) \)

iii. The test statistic is then obtained as

\[
\tau = \max \tau' (\tau) \quad \text{(one-sided)}
\]
\[
\tau = \max |\tau' E \tau F | \quad \text{(two-sided)}
\]

iv. The null distribution of \( \tau \) can be derived via a perturbation-resampling procedure
Application
Metastatic breast cancer example

RMST-based versatile test

<table>
<thead>
<tr>
<th>Selected</th>
<th>2, 4, 6, é, 24 } mon</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value</td>
<td>P=0.010</td>
</tr>
<tr>
<td>Difference (10 mon)</td>
<td>1.05 (0.95CI, 0.14 to 1.96) mon</td>
</tr>
</tbody>
</table>
Numerical studies

1. Metastatic breast cancer example
2. Another cancer example (RAINBOW study)
Metastatic breast cancer example

Event time distribution

No difference observed in EXEMESTANE
Metastatic breast cancer example

Other parameters

- Test: two-sided
- Total study time: 24 months
- Time points for: \{ , , , ..., \} (months)
- Sample size: N = 300 (per arm)
- The number of the perturbation-resampling: 5000
- The number of iterations: 2000
- Comparisons:
  - Log-rank test
  - Peto-Prentice-Wilcoxon test
  - Standard RMST-based test ( = months)
## Results

### Size (nominal 0.05)

<table>
<thead>
<tr>
<th>Test</th>
<th>No censoring</th>
<th>Light censoring</th>
<th>Moderate censoring</th>
<th>EXEM. censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logrank</td>
<td>0.048</td>
<td>0.052</td>
<td>0.055</td>
<td>0.054</td>
</tr>
<tr>
<td>Peto-Prentice Wilcoxon</td>
<td>0.054</td>
<td>0.050</td>
<td>0.052</td>
<td>0.048</td>
</tr>
<tr>
<td>Standard RMST</td>
<td>0.053</td>
<td>0.053</td>
<td>0.051</td>
<td>0.055</td>
</tr>
<tr>
<td>Versatile RMST</td>
<td>0.056</td>
<td>0.048</td>
<td>0.047</td>
<td>0.051</td>
</tr>
</tbody>
</table>

### Power

<table>
<thead>
<tr>
<th>Test</th>
<th>No censoring</th>
<th>Light censoring</th>
<th>Moderate censoring</th>
<th>EXEM. censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logrank</td>
<td>0.568</td>
<td>0.572</td>
<td>0.587</td>
<td>0.560</td>
</tr>
<tr>
<td>Peto-Prentice Wilcoxon</td>
<td>0.809</td>
<td>0.815</td>
<td>0.812</td>
<td>0.820</td>
</tr>
<tr>
<td>Standard RMST</td>
<td>0.649</td>
<td>0.644</td>
<td>0.624</td>
<td>0.652</td>
</tr>
<tr>
<td>Versatile RMST</td>
<td>0.932</td>
<td>0.930</td>
<td>0.926</td>
<td>0.934</td>
</tr>
</tbody>
</table>
Another cancer example (RAINBOW study)

- A phase III randomized trial to compare ramucirumab + paclitaxel and placebo + paclitaxel for advanced gastric cancer.
- N=665 (330 on ramucirumab, 335 on placebo)

Wilke et al. (2014, Lancet)
Overall Survival

The KM curves suggest that the PH assumption does not hold (the cumulative residual test, \(p=0.002\))
RAINBOW study

Other parameters

- Test: two-sided
- Total study time: 21 months
- Time points for: = \{ , , , , , , \} (months)
- Sample size: N = 300 (per arm)
- The number of the perturbation-resampling: 5000
- The number of iterations: 2000
- Comparisons:
  - Log-rank test
  - Peto-Prentice-Wilcoxon test
  - Standard RMST-based test ( = months)
# RAINBOW study Results

## Power

<table>
<thead>
<tr>
<th>Test</th>
<th>No censoring</th>
<th>Light censoring</th>
<th>Moderate censoring</th>
<th>RAINBOW censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logrank</td>
<td>0.435</td>
<td>0.467</td>
<td>0.497</td>
<td>0.586</td>
</tr>
<tr>
<td>Peto-Prentice Wilcoxon</td>
<td>0.870</td>
<td>0.875</td>
<td>0.872</td>
<td>0.893</td>
</tr>
<tr>
<td>Standard RMST</td>
<td>0.746</td>
<td>0.746</td>
<td>0.722</td>
<td>0.736</td>
</tr>
<tr>
<td>Versatile RMST</td>
<td>0.940</td>
<td>0.945</td>
<td>0.940</td>
<td>0.945</td>
</tr>
</tbody>
</table>
Summary of numerical studies

- The empirical significance levels of all tests are close to their nominal value of 0.05.

- Although the proposed test is inferior to the logrank test under PH alternatives by theory, the proposed test is dramatically powerful for the pattern of the difference seen in those cancer clinical trials.
Conclusions
Conclusions

Several practical issues arise from the conventional design and analysis using the logrank-HR test estimation practice.

The RMST-based versatile test is

- a model-free, pre-specified test, and
- dramatically powerful for patterns of difference seen in some recent cancer clinical trials.

It also provides corresponding robust and interpretable quantitative information of the treatment effect (test-estimation coherency).
References


Tian et al. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics 2014; 15: 222-233.


Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med 2011; 30: 2409-2421.


References


