The hazards of hazard ratios

L.J. Wei, Harvard University

Thanks to H. Uno, Lu Tian, B. Claggett, Dae Kim, Lihui Zhao, Bo Huang, T. Cai, Zack McCaw, Ray Sun

Time to a clinical event as the endpoint in a comparative study

• How to empirically summarize the “survival” (event-free) time profile for each treatment group?
• Kaplan-Meier (cumulative incidence curve)
• Event rate (at a specific time point)
• Median “survival” time (may not be observable)
• Hazard curve (hard to estimate well nonparametrically)

• Restricted mean survival time (or t-year mean survival time), RMST
Restricted mean survival time (RMST) Difference:

2.2 months; CI: 0.5 to 4.0, \( p=0.014 \)

Metrics for quantifying the group difference

- Event rate difference (or ratio)
- Difference of two median failure times
- Hazard ratio (routinely used in practice)
- Difference (ratio) between two RMSTs.

- (Moving beyond p-value)
- (Ideally using estimate to do testing too, such as logrank test and HR)
How to communicate with patients via various quantifiers for treatment effect?

Amiodarone or an Implantable Cardioverter–Defibrillator for Congestive Heart Failure

**METHODS**
We randomly assigned 2521 patients with New York Heart Association (NYHA) class II or III CHF and a left ventricular ejection fraction (LVEF) of 35 percent or less to conventional therapy for CHF plus placebo (847 patients), conventional therapy plus amiodarone (845 patients), or conventional therapy plus a conservatively programmed, shock-only, single-lead ICD (829 patients). Placebo and amiodarone were administered in a double-blind fashion. The primary end point was death from any cause.
Risks associated with ICD implantation are uncommon but may include:

- Infection at the implant site
- Allergic reaction to the medications used during the procedure
- Swelling, bleeding or bruising where your ICD was implanted
- Damage to the vein where your ICD leads are placed
- Bleeding around your heart, which can be life-threatening
- Blood leaking through the heart valve where the ICD lead is placed
- Collapsed lung (pneumothorax)

Shared decision making between patients and clinicians

GJ is a 79-year-old woman with hypertension, diabetes, osteoporosis, depression, and New York Heart Association class II heart failure with a left ventricular ejection fraction of 30%. She is a potential candidate for an implantable cardioverter-defibrillator (ICD), and you would like to discuss this with her using evidence from a clinical trial. Which of the following statistics would be most helpful in explaining the possible survival benefit of an ICD?

- p-value comparing mortality of ICD and placebo groups was 0.007.
- hazard ratio (HR) for mortality was 0.77.
- absolute risk reduction was 7%, from 36% to 29%, over 5 years.
- number-needed-to-treat (NNT) was 15 over 5 years.
- ICD will prolong life from 49.1 to 51.4 months, an average of 2.3 months, over 5 years.
Study in acute lymphoblastic leukemia comparing inotuzumab with chemotherapy (NEJM, 2016)

Ref. Annals of Internal Medicine, Guideline for Authors
Let us see what Sir David told us..

In an interview, Professor David R. Cox, the creator of the proportional hazards model, stated, “Of course, another issue is the physical or substantive basis for the proportional hazards model. I think that’s one of its weaknesses…”

“If you can’t explain it simply, you just don’t understand it well enough.”

— Albert Einstein
Beyond “translational” what are other advantages of RMST analysis?

*HR does not have a causal treatment effect interpretation.

*When PH assumption is not met, HR is difficult to interpret, which is not a simple average of hazard ratios over time. The parameter HR estimated depending on the censoring distributions.

*RMST based statistics can be more powerful than HR under Non-PH.

*When HR gives significant results, so does RMST.

*For equivalence or non-inferiority studies, RMST does not require a large study like HR (event driven).

*RMST uses more data than HR

**t-year mean survival time**

Uno et al. (2014, JCO)

Pak et al. (2017, JAMA-Oncology)

Uno et al. (2015, Annals of Internal Medicine)
For non-proportional hazards, RMST may be more powerful

Example: ECOG myeloma study

Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial

S Vincent Rajkumar, Susanna Jacobs, Natalie S Collander, Rafael Fonseca, David H Vesole, Michael E Williams, Rafat Abounour, David S Siegal, Michael Katz, Philip R Greipp, for the Eastern Cooperative Oncology Group

Summary

Background High-dose dexamethasone is a mainstay of therapy for multiple myeloma. We studied whether low-dose dexamethasone in combination with lenalidomide is non-inferior to and has lower toxicity than high-dose dexamethasone plus lenalidomide.

Rajkumar et al. (2010, Lancet Oncology)
ECOG Myeloma study (OS, low Dex vs. High Dex)

**Survival function**

![Survival function graph](image)

**Hazard ratio**

![Hazard ratio graph](image)

HR = 0.87 (0.95CI: 0.60 to 1.27), p=0.46

Restricted mean survival time (RMST) Difference:

2.2 months; CI: 0.5 to 4.0, p=0.014
HR gives significant result, so does RMST

Analysis of 7 clinical studies for heart failure

- C. Perego; M. Sbolli; C. Specchia; C. Oriecuia; G. Peveri; M.Metra; M. Fiuzat; L.J. Wei; C.M. O’Connor; M.A. Psotka

Inova Heart Vascular Inst. Falls Church, US; Univ of Bressia, Italy; Univ of Milan, Italy; Duke Univ; SPEDALI CIVILI Hosp, Italy, Harvard Univ.
<table>
<thead>
<tr>
<th>Trial</th>
<th>Treatment(s)</th>
<th>Outcome</th>
<th>Hazard Ratio (HR)</th>
<th>p-value</th>
<th>RMST Difference (months)</th>
<th>p-value</th>
<th>TIME (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSENSUS</td>
<td>Enalapril vs Placebo</td>
<td>All-cause death</td>
<td>0.73 (0.60-0.90)</td>
<td>0.003</td>
<td>2.2 (1.1 - 3.4)</td>
<td>&lt; 0.001</td>
<td>12</td>
</tr>
<tr>
<td>RALES</td>
<td>Spironolactone vs Placebo</td>
<td>All-cause death</td>
<td>0.70 (0.60-0.82)</td>
<td>0.001</td>
<td>2.2 (1.1 - 3.4)</td>
<td>&lt; 0.001</td>
<td>36</td>
</tr>
<tr>
<td>COPERNICUS</td>
<td>Carvedilol vs Placebo</td>
<td>Death or cardiovascular hospitalization</td>
<td>0.73 (0.60-0.90)</td>
<td>&lt; 0.001</td>
<td>1.7 (1.1 - 2.4)</td>
<td>&lt; 0.001</td>
<td>21</td>
</tr>
<tr>
<td>MERIT-HF</td>
<td>Metoprolol CR/XL vs Placebo</td>
<td>All-cause death</td>
<td>0.66 (0.53-0.84)</td>
<td>&lt; 0.001</td>
<td>0.4 (0.2 - 0.7)</td>
<td>&lt; 0.001</td>
<td>18</td>
</tr>
<tr>
<td>SHIFT</td>
<td>Isosorbide dinitrate vs Placebo</td>
<td>Cardiovascular death or HF hospitalization</td>
<td>0.82 (0.74-0.89)</td>
<td>&lt; 0.001</td>
<td>1.0 (0.8 - 1.2)</td>
<td>&lt; 0.001</td>
<td>30</td>
</tr>
<tr>
<td>PARADIGM-HF</td>
<td>Sacubitril/valsartan vs Enalapril</td>
<td>Cardiovascular death or HF hospitalization</td>
<td>0.85 (0.73-0.99)</td>
<td>&lt; 0.001</td>
<td>1.5 (0.8 - 2.0)</td>
<td>&lt; 0.001</td>
<td>41</td>
</tr>
<tr>
<td>DAPA-HF</td>
<td>Dapagliflozin vs Placebo</td>
<td>Cardiovascular death or worsening HF</td>
<td>0.74 (0.64-0.85)</td>
<td>&lt; 0.001</td>
<td>0.9 (0.6 - 1.2)</td>
<td>&lt; 0.001</td>
<td>24</td>
</tr>
</tbody>
</table>

CheckMate 057 Study

**Nivolumab versus Docetaxel in Advanced Nonsquamous Non–Small-Cell Lung Cancer**


Borghaei et al. (2015, NEJM)
A Overall Survival

No. at Risk
Nivolumab 292 232 194 169 146 123 92 32 9 0
Docetaxel 290 244 194 150 111 88 34 10 5 0

Borghaei et al. (2015, NEJM)

C Progression-free Survival

No. at Risk
Nivolumab 292 128 82 58 46 35 17 7 2 0
Docetaxel 290 156 87 38 18 6 2 1 1 0

Borghaei et al. (2015, NEJM)
Issues and concerns

- The violation of the PH assumption is obvious
- When the PH assumption does not hold, the HR is not a valid summary
- Median event time difference as the summary does not reflect the entire temporal profile between two treatment groups
- RMST can use all the data until the last censored or observed time point, but HR can only use the last observed data point.
Our letter to the editor

Nivolumab in Nonsquamous Non–Small-Cell Lung Cancer

TO THE EDITOR: In the article on the CheckMate 057 trial, Bergsagel et al. (Oct. 22 issue) provide data on overall and progression-free survival among patients with advanced nonsquamous non–small-cell lung cancer who were receiving either nivolumab or docetaxel. In this trial, docetaxel initially appeared to have better outcomes than nivolumab, but the trends were reversed after 9 months (Fig. 1 of the article, available at NEJM.org). In such instances in which hazard functions for two treatment groups cross during the study follow-up, it is not clear how to interpret the observed hazard ratios of 0.73 for death and 0.92 for disease progression or death for nivolumab as compared with docetaxel. An alternative is to use the restricted mean survival time to quantify the treatment benefit.13 For overall survival, an estimated restricted mean survival time up to 24 months for nivolumab is the area under the Kaplan-Meier curve up to 24 months, which is 13 months. In other words, future patients receiving nivolumab for 2 years would survive for an average of 13 months. The difference in the restricted mean survival time between the two groups would be 1.7 months (95% confidence interval [CI], 0.4 to 3.1) in favor of nivolumab.14 For progression-free survival, the difference in the restricted mean survival time is 1.3 months (95% CI, 0.2 to 2.3), again in favor of nivolumab. This quantification of treatment benefit has a much clearer clinical interpretation than its hazard-ratio counterpart, especially in cases in which hazard functions for two groups cross.

Takahiro Hazegawa, D.P.H.
Shinagai
Osaka, Japan

Hajime Uno, Ph.D.
Jna–Farber Cancer Institute
Boston, MA

Lee-Jen Wei, Ph.D.
Harvard University
Boston, MA

wei@hsp.harvard.edu

The New England Journal of Medicine

Highly statistical significance may not be clinically significant
ExteNET Study

Neratinib after trastuzumab-based adjuvant therapy in patients with HER2-positive breast cancer (ExteNET): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial


Summary
Background Neratinib, an irreversible tyrosine-kinase inhibitor of HER1, HER2, and HER4, has clinical activity in patients with HER2-positive metastatic breast cancer. We aimed to investigate the efficacy and safety of 12 months of neratinib after trastuzumab-based adjuvant therapy in patients with early-stage HER2-positive breast cancer.

Chan et al. (2016, Lancet Onc)

Disease-free survival including ductal carcinoma in situ (DCIS)

Chan et al. (2016, Lancet Onc)
Issues and concerns

• The PH may be ok, but the hazard ratio is difficult to explain with a short-term follow-up

• What is the gain from the extra treatment clinically?

• No median survival time estimate

Our analysis results for a clear clinical interpretation

Disease-free survival including DCIS

<table>
<thead>
<tr>
<th>Up to 24 months</th>
<th>Estimate</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMST Neratinib</td>
<td>23.43</td>
<td>(23.28, 23.58)</td>
<td></td>
</tr>
<tr>
<td>Placebo</td>
<td>22.84</td>
<td>(22.62, 23.06)</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>0.59</td>
<td>(0.33, 0.86)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.03</td>
<td>(1.01, 1.04)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Non-inferiority studies
EPOETIN safety study

A Randomized, Open-Label, Multicenter, Phase III Study of Epoetin Alfa Versus Best Standard of Care in Anemic Patients With Metastatic Breast Cancer Receiving Standard Chemotherapy


ABSTRACT

Purpose
An open-label, noninferiority study to evaluate the impact of epoetin alfa (EPO) on tumor outcomes when used to treat anemia in patients receiving chemotherapy for metastatic breast cancer.

Methods
Women with hemoglobin = 11.0 g/dL, receiving first- or second-line chemotherapy for metastatic breast cancer, were randomly assigned to EPO 40,000 IU subcutaneously once a week or best standard of care. The primary end point was progression-free survival (PFS). Secondary end points included overall survival, time to tumor progression, overall response rate, RBC transfusions, and thrombotic vascular events.

PFS by Investigator

<table>
<thead>
<tr>
<th>Time Since Random Assignment (months)</th>
<th>% of Patients Progression Free and Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

No. at risk

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Event (%)</th>
<th>Median</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSC</td>
<td>1,048</td>
<td>816 (78)</td>
<td>7.4</td>
<td>7.1 to 7.6</td>
</tr>
<tr>
<td>Epoetin alfa</td>
<td>1,050</td>
<td>841 (80)</td>
<td>7.4</td>
<td>6.9 to 7.6</td>
</tr>
</tbody>
</table>

HR, 1.06: 95% CI, 0.98 to 1.20

Leyland-Janes et al. (2016, JCO)
Issues and concerns

- HR is difficult to interpret due to the lack of a clear benchmark reference
- Such a lengthy study using HR as the summary may not help us to assess the value of EPO

Our analysis results

<table>
<thead>
<tr>
<th></th>
<th>RMST up to</th>
<th>Estimate</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFS</td>
<td>48 months</td>
<td>BSC</td>
<td>11.40</td>
<td>(10.56, 12.23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epoetin alfa</td>
<td>9.87</td>
<td>(9.23, 10.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difference</td>
<td>-1.53</td>
<td>(-2.58, -0.47)</td>
</tr>
</tbody>
</table>
Quantifying long term survival
Example of immunotherapy trial (CheckMate 214)

The NEW ENGLAND
JOURNAL of MEDICINE

Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma


CheckMate 214: Progression-Free Survival

Motzer et al. (2018, NEJM)
The traditional logrank/HR approach

Test: log-rank: p=0.035
Estimation: HR=0.82 (95%CI: 0.68 to 0.99)

Long-term RMST-based analysis

(Horighuchi, Tian, Uno, Cheng et al. 2018, JAMA Onc)

Consider area under the curve only on $[t_1, \tau]$

Difference in RMST $[5 - 27\text{m}]$
1.7m (95%CI: 0.3m to 3.2m)
Estimating the duration of response

Traditional ways to analyze DOR

Empirically, DOR among responders?

- Valid statistical inference cannot be made comparing two treatment groups
- Response is an outcome after randomization
- Under-estimating the treatment effect if there were more responders in the treated group
Traditional ways to analyze DOR

**KM curve for the DOR with responders (reporting median DOR)**

By adjusting KM curve by considering non-responder’s DOR being zero, there is an issue
Method

The duration of response is “time to P/D” – “time to P/D/R”

- Programming code available at https://web.stanford.edu/~lutian/Software.HTML
Conclusions

- RMST can be used for designing the study (JAMA-Oncology, Pak et al. 2017)
- Regression analysis for RMST
- R package: survRM2, and SAS: PROC RMSTREG
Cox model with baseline covariate adjustment (or stratified Cox)

- When two sample PH assumption is ok, the Cox ANCOVA is not valid (incoherent)
- Augmentation procedures (Tsiatis et al.; Tian et al.)

Identifying a high value subgroup of patients who benefit from treatment

- How to use patient baseline information to identify a high value subgroup?
How to use the real-world observational study data?

- How to integrate clinical trial data with observational data to evaluate treatment effect/toxicity?

Totality of evidence on the treatment effect/toxicity?

- For each patient, we have response, progression, death, toxicity information, how can be integrate them to create a clinically interpretable study endpoint?
There is an R package for designing studies with RMST (SSRMST).
• https://cran.r-project.org/web/packages/SSRMST/index.html

Maybe we need to move out of box for design and analysis of clinical studies?