Statistics•Collaborative

Once in Love with Hazard Ratios

Janet Wittes, PhD
Department of Health Research Methods, Evidence, and Impact
McMaster University
June 27, 2022

The problem

- You have done a randomized controlled trial
  - Two treatments – experimental and control
  - Q: which treatment leads to a lower mortality rate?
- What do you do?

1956

- You know you cannot just consider Pr{alive} at end of study
- You refer to demographers and actuaries[1]
- You construct a life table
  - You bin your data into years (or months)
  - You calculate the probabilities of 1 year, 2 year,…, survival[2]
- And, if you are savvy, you know how to calculate SEs[3]

1. John Graunt and Edmond Halley around 1650
2. Greenwood M, Yule GU (1920). An inquiry …. *JRSS*; **83**:255–279.

3 Greenwood M (1926). "The natural duration of cancer". *Reports on Public Health and Medical Subjects*. London: Her Majesty's Stationery Office. 33: 1–26.

What you would see in 1956

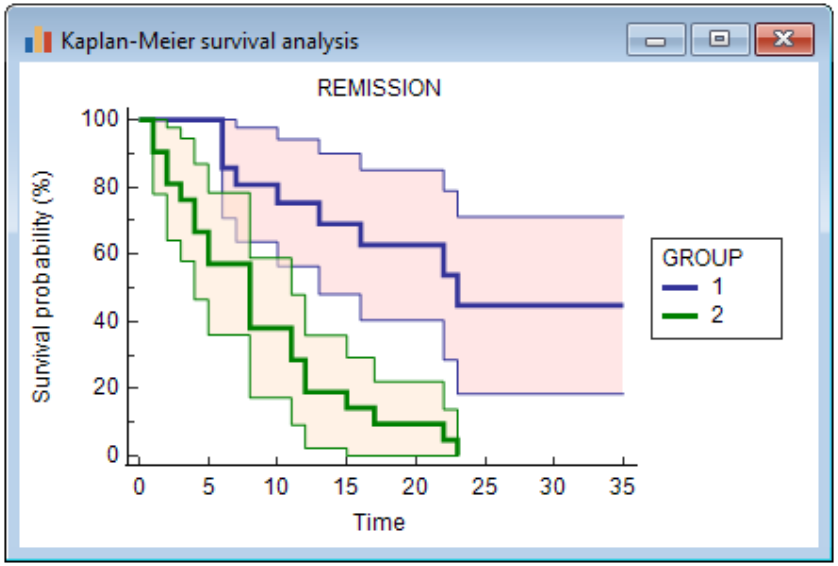
Table 1. Life table for the total population: United States, 2008

Spreadsheet version available from: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/NVSR/61_03/Table01.xls.

Age (years)	Probability of dying between ages x and $x+1$	Number surviving to age x	Number dying between ages x and $x+1$	Person-years lived between ages x and $x+1$	Total number of person-years lived above age x	Expectation of life at age x
	q_x	l_x	d_x	L_x	T_x	e_x
0–1	0.006593	100,000	659	99,425	7,812,389	78.1
1–2	0.000461	99,341	46	99,318	7,712,964	77.6
2–3	0.000281	99,295	28	99,281	7,613,646	76.7
3–4	0.000219	99,267	22	99,256	7,514,365	75.7
4–5	0.000172	99,245	17	99,237	7,415,109	74.7
5–6	0.000155	99,228	15	99,221	7,315,872	73.7
6–7	0.000139	99,213	14	99,206	7,216,651	72.7
7–8	0.000126	99,199	12	99,193	7,117,445	71.7
8–9	0.000110	99,187	11	99,181	7,018,252	70.8
9–10	0.000093	99,176	9	99,171	6,919,071	69.8
10–11	0.000081	99,167	8	99,162	6,819,900	68.8

1958 – the Kaplan Meier curve

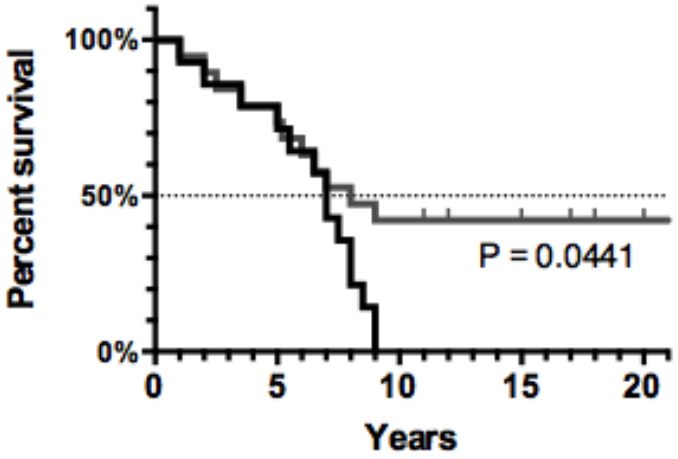
- You no longer have to bin your data
  - You present the data as a graph
  - The curve jumps at each event
  - You can use Greenwood for SEs
- You assume
  - Censoring unrelated to prognosis
  - Prob of survival independent of time of recruitment
- But, you can't formally test



Kaplan EL; Meier P (1958). Nonparametric estimation from incomplete observations. JASA 53: 457–481.

1966 – the log rank test

- As before
  - Assume censoring unrelated to prognosis
  - Assume prob of survival independent of time of recruitment
- You can assign a p-value to the curves
  - And it's optimal if the hazards are proportional
- But it has no summary statistics
  - Oncologists used the median, but that is only one point on the curve



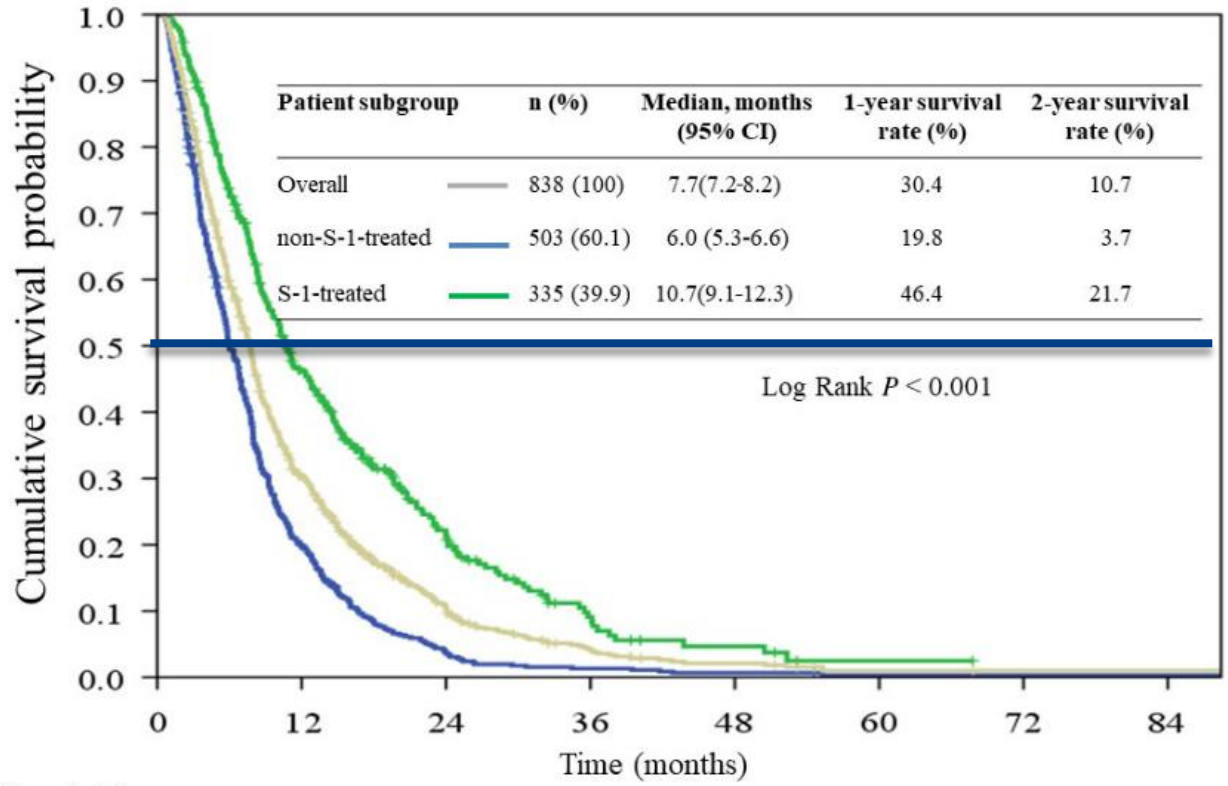
Mantel N (1966). Evaluation of survival data…. *Cancer Chemotherapy Reports*. **50**: 163–70

Why do we use medians anyhow?

- Answer – because usually the curve doesn't go down to 0
- If it did, the area under the curve would be mean survival

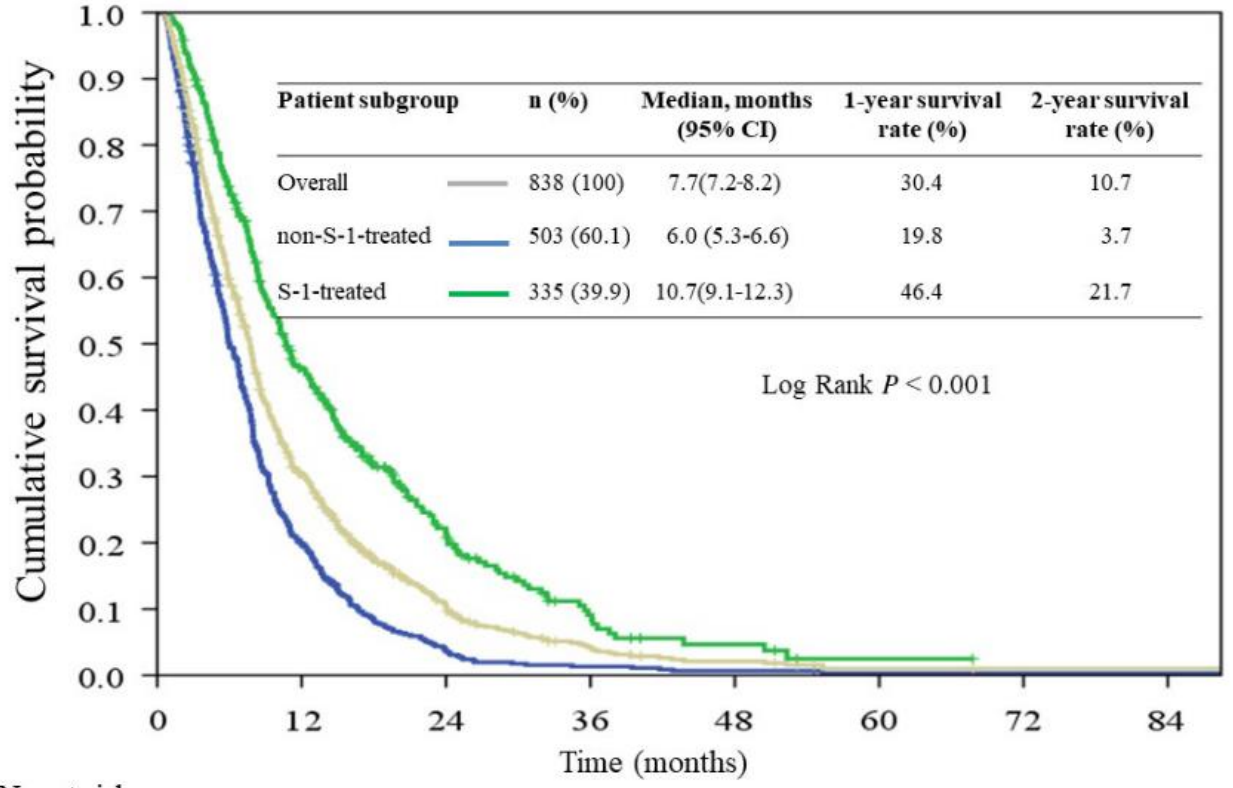
Pancreatic cancer

Observational study in Taiwan. S-1 is an oral 5-FU derivative.



Lai et al. (2019) Cancer Medicine 8:2085-94.

Pancreatic cancer



Note that the area under the curve is the mean survival

Why not assume a functional form for survival?

- We knew if we assumed an exponential or Weibull distribution…
  - we could get estimates of parameters
- But, consensus that such assumptions were too strong
- So we were stuck…
  - We could characterize survival visually
  - We could test whether the curves differed statistically
  - But we had no way of getting an estimate for the whole curve

1972 – Cox's proportional hazards model

Semi-parametric

- Survival curves have two parts…
  - Underlying baseline hazard function-
    - How hazard changes over time assuming baseline levels
  - How hazard changes over time as a function of covariates
    - For trials, the main covariate is treatment
- Assume changes over time are proportional
  - i.e., Hazard ratio is constant over time
  - No need to know the underlying hazard function
  - We can then estimate the hazard ratio

We now had all the tools we needed…

- We could draw our survival curves (Kaplan-Meier)
- We could calculate SEs (Greenwood) and p-value (Mantel)
- And we could summarize the curves with a single parameter that characterized the whole curve – hazard ratio (Cox)
- Further, if our data were in fact exponential or Weibull
  - We would have proportional hazards.

But what is a hazard ratio?

- National Cancer Institute -A measure of how often a particular event happens in one group compared to how often it happens in another group

But what is a hazard ratio?

- National Cancer Institute -A measure of how often a particular event happens in one group compared to how often it happens in another group, over time
- The hazard ratio describes the relative risk of the complication based on comparison of event rates. …The hazard ratio is the odds of a patient's healing faster under treatment but does not convey any information about how much faster this event may occur.
- Wikipedia: In survival analysis, the **hazard ratio** (**HR**) is the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable.„„ Hazard ratios differ from relative risks (RRs) and odds ratios (ORs) in that RRs and ORs are cumulative over an entire study, using a defined endpoint, while HRs represent instantaneous risk over the study time period, or some subset thereof.

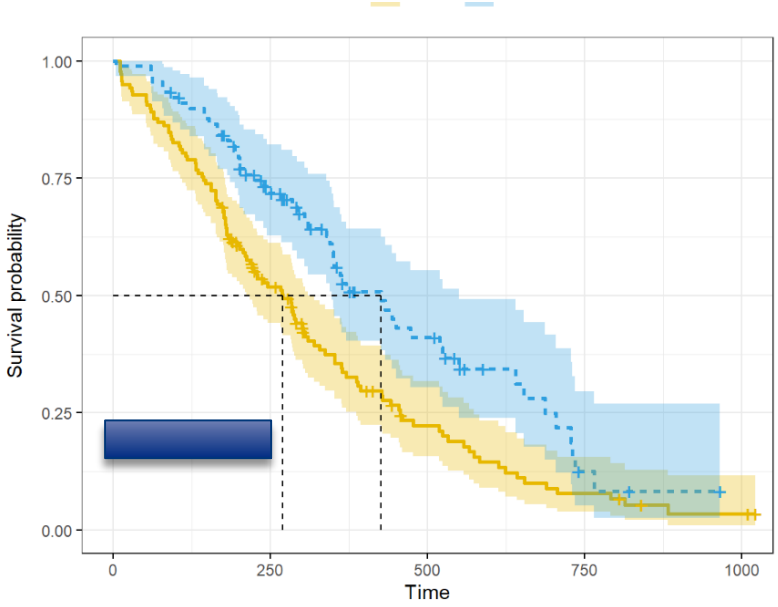
More definitions

- This brief communication will clarify the difference between a relative hazard and a relative risk. We highlight the importance of this difference, and demonstrate in practical terms that 1 minus the hazard ratio should not be interpreted as a risk reduction in the commonly understood sense of the term.

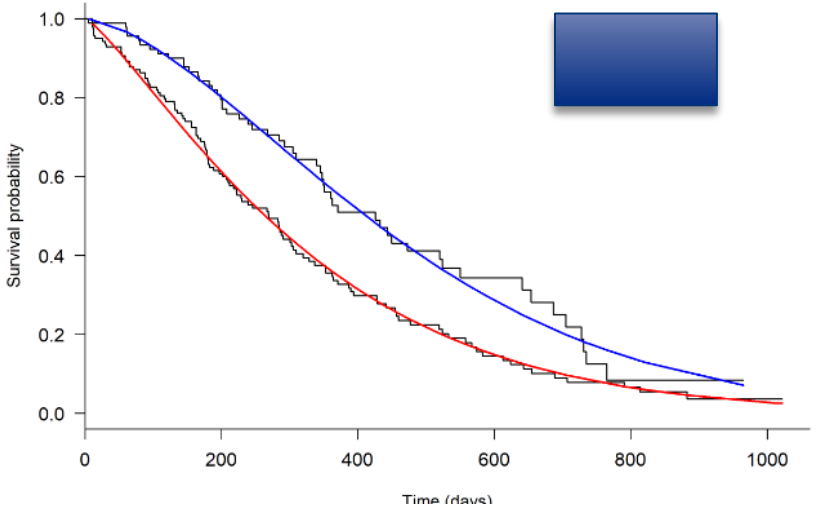
Sashegyi A, Ferry D. On the Interpretation of the Hazard Ratio and Communication of Survival Benefit. The Oncologist 2017 Apr; 22(4): 484–486.

Exponential and Weibull Kaplan-Meier curves

Exponential

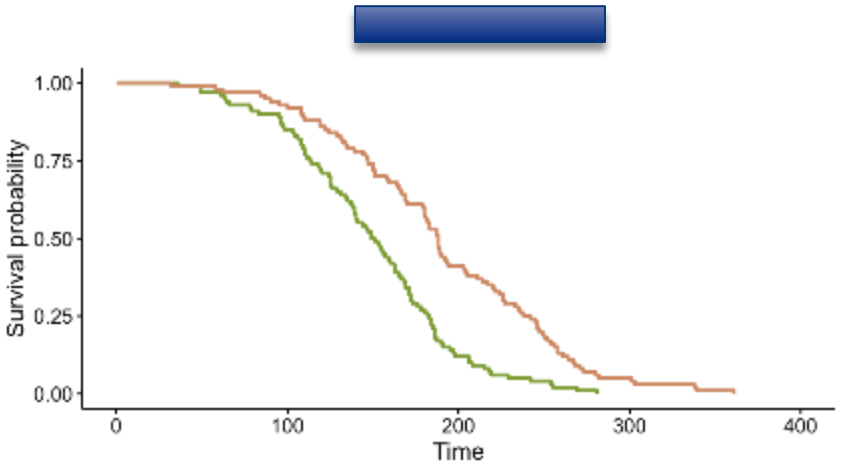


Weibull

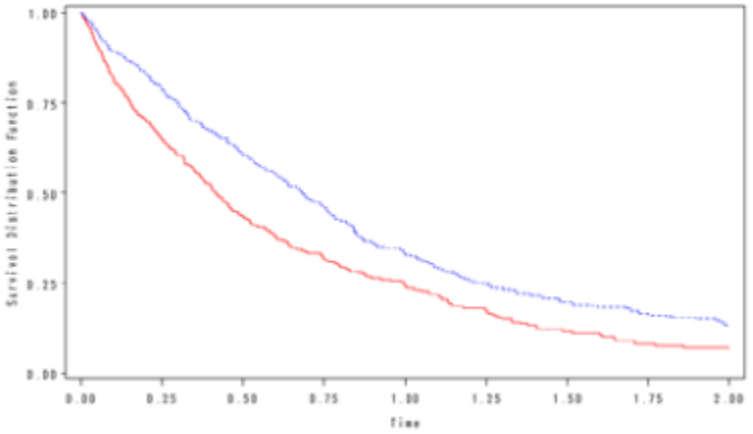


Even if hazards are not exactly proportional

Early on the curves stick together

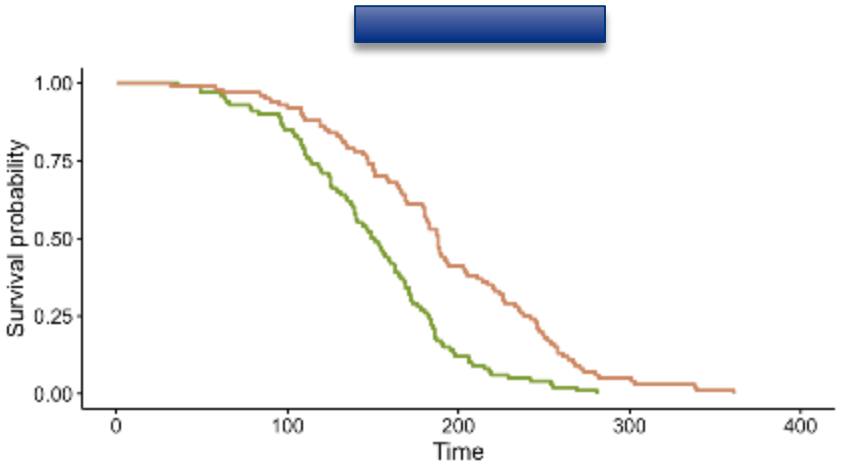


Hazards don't look proportional

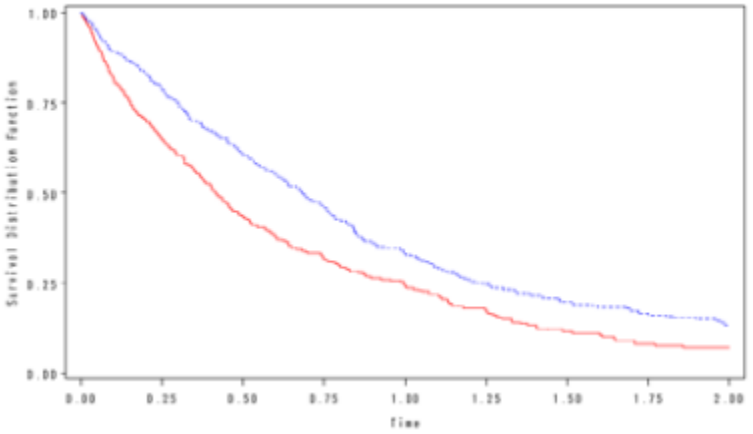


Even if hazards are not exactly proportional

Early on the curves stick together



Hazards don't look proportional



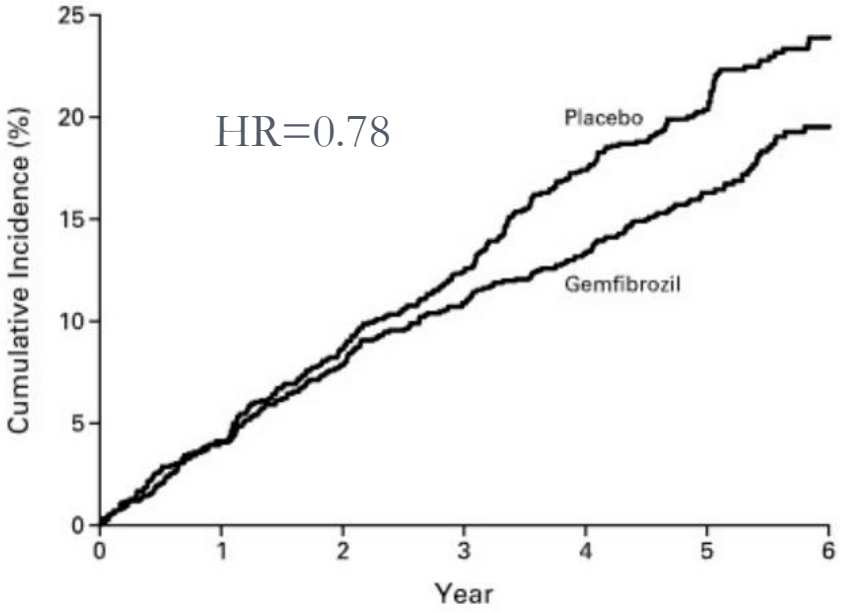
But these are close enough for most of us. Think of the hazard ratio as some sort of average hazard ratio over time.

VA-HIT

To determine if gemfibrozil can reduce CHD death and MI in patients whose primary lipid abnormality is a low level of HDL cholesterol.

Rubins HB et al. (1999) NEJM 341:410-418

But what if there is a delay in effect?



No. at Risk							
Placebo	1267	1200	1118	1040	962	666	466
Gemfibrozil	1264	1201	1128	1067	1005	706	498

We actually wrote: "reduction in relative risk was 22 percent (95 percent confidence interval, 7 to 35 percent; P=0.006)."

Rubins HB et al. (1999) NEJM 341:410-418

What about a 1 year trial

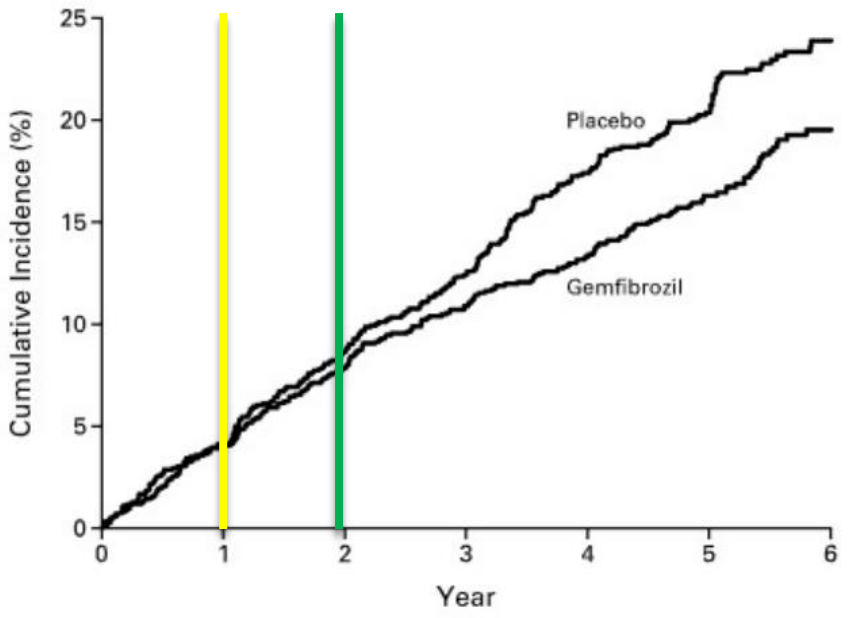
HR=1

No. at Risk

Placebo	1267	1200	1118	1040	962	666	466
Gemfibrozil	1264	1201	1128	1067	1005	706	498

What about a 1 year, 2 year trial

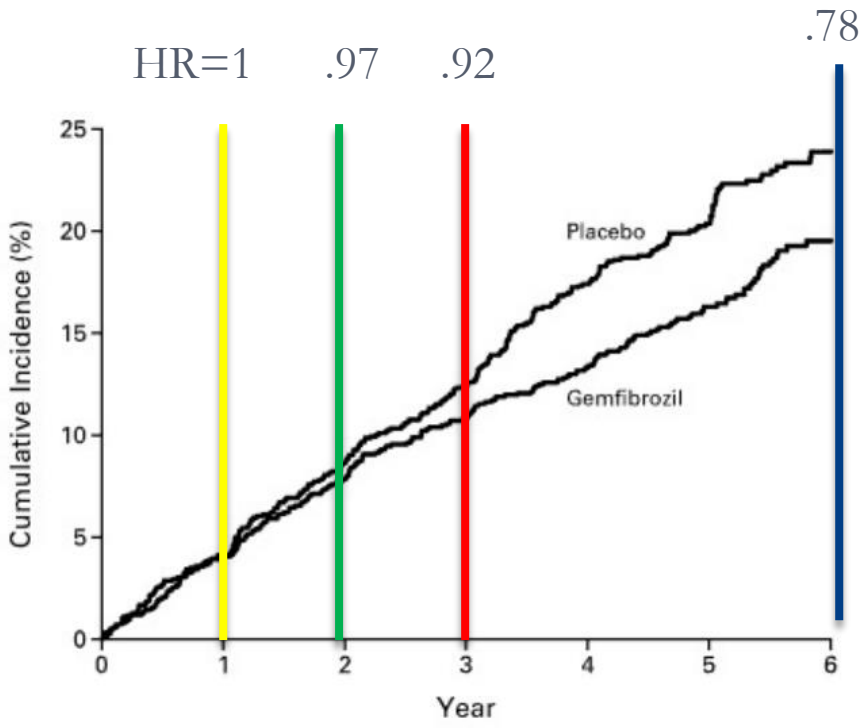
HR=1 .97



No. at Risk

Placebo	1267	1200	1118	1040	962	666	466
Gemfibrozil	1264	1201	1128	1067	1005	706	498

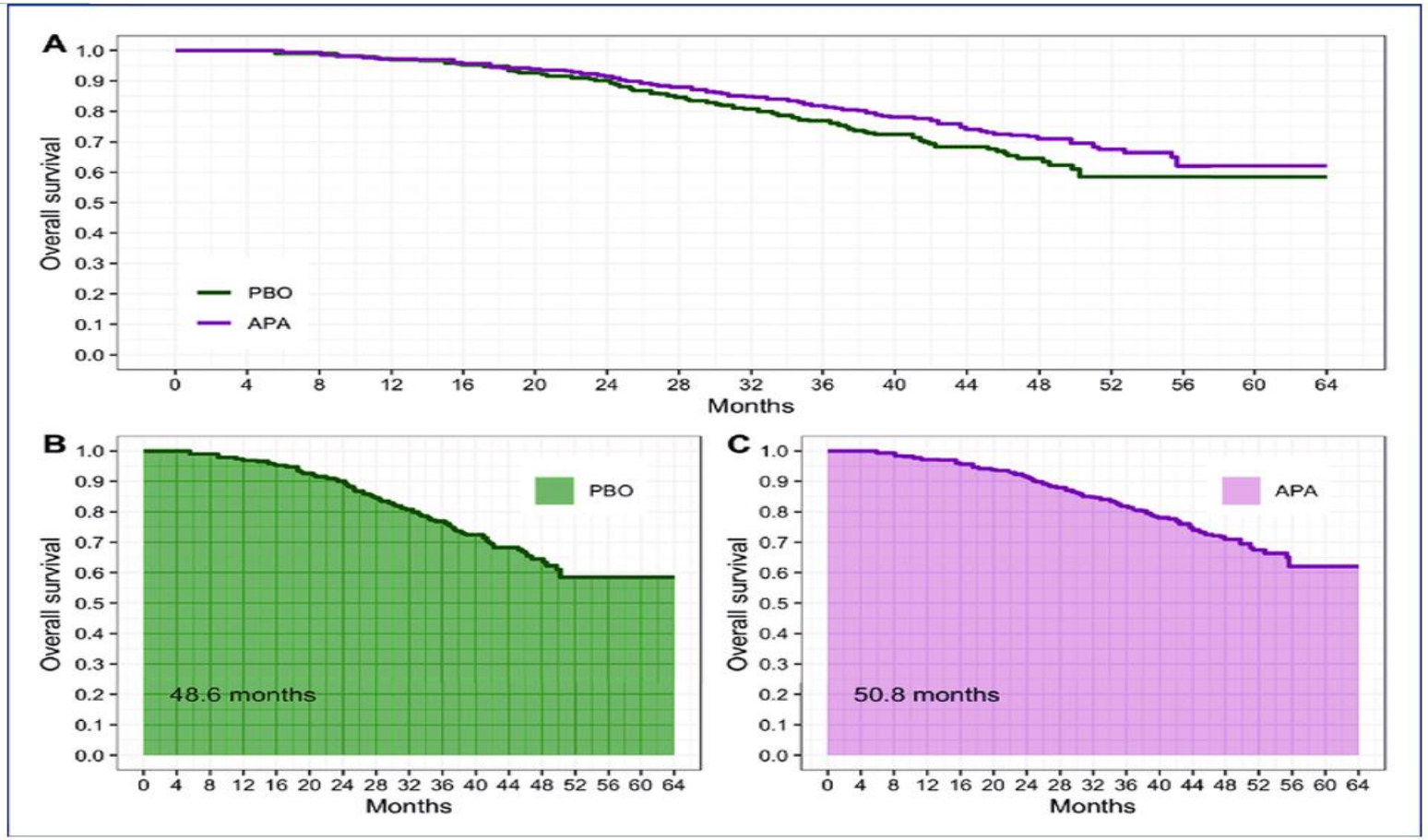
What about a 1 year, 2 year, 3 year trial



The longer the trial, the smaller our hazard ratio becomes.

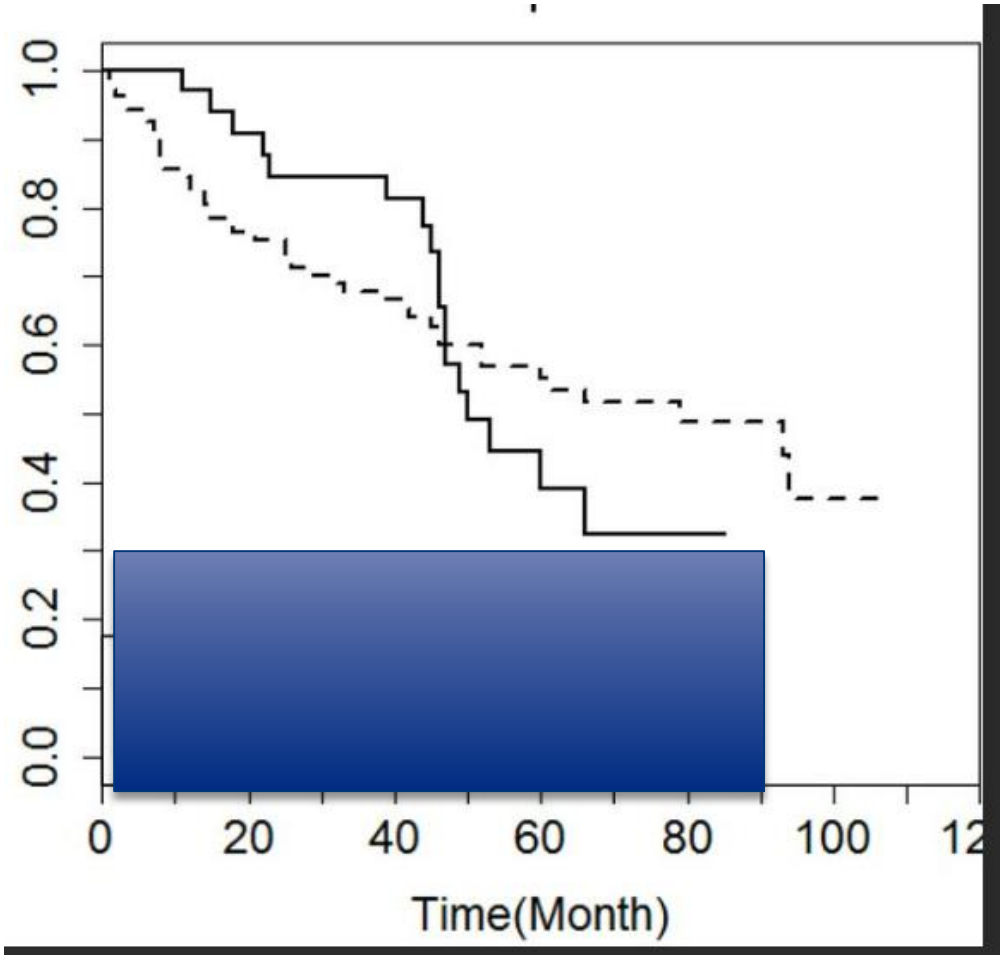
So the least we should do when we cite a HR is report the length of the trial.

Restricted mean survival time



McCaw, Wei, Ludmir (2020). Interpreting the Impact of Apalutamide on Overall Survival Among Patients with Non-metastatic Castration-resistant Prostate Cancer. Annals of Oncology.

What if the curves cross?



Most people would say, "don't use log rank or Cox". But RMST is still meaningful

What are our choices?

Criterion	HR	Median	RMST	AD
Easily interpretable	☹	☺	☺	☺
Reflects entire survival history	☺	☹	😐	☺
Is a measure of survival time	☹	☺	☺	☹
Can be used with all models	☹	☺	☺	☺
Can be calculated in any dataset	☺	☹	☺	☺
Does not require specifying a timepoint	☺	☺	☹	☹
Does not change with extended follow-up	☹	😐	☺	☺
Is routinely associated with a clinically meaningful timepoint	☹	☹	😐	☺
Does not assume proportional hazards	☹	☺	☺	☺

AD=absolute difference in proportions

Shamelessly cribbed from Royston P, Parmar MKD (2013) BMC Med Res Meth. Dec 7, 13:152

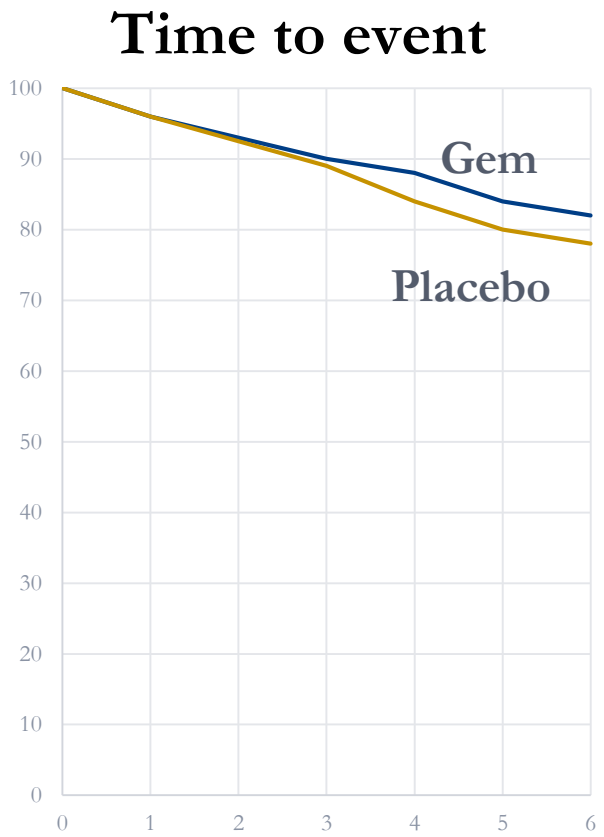
The four measures for VA-HIT

HR=0.78

No. at Risk							
Placebo	1267	1200	1118	1040	962	666	466
Gemfibrozil	1264	1201	1128	1067	1005	706	498

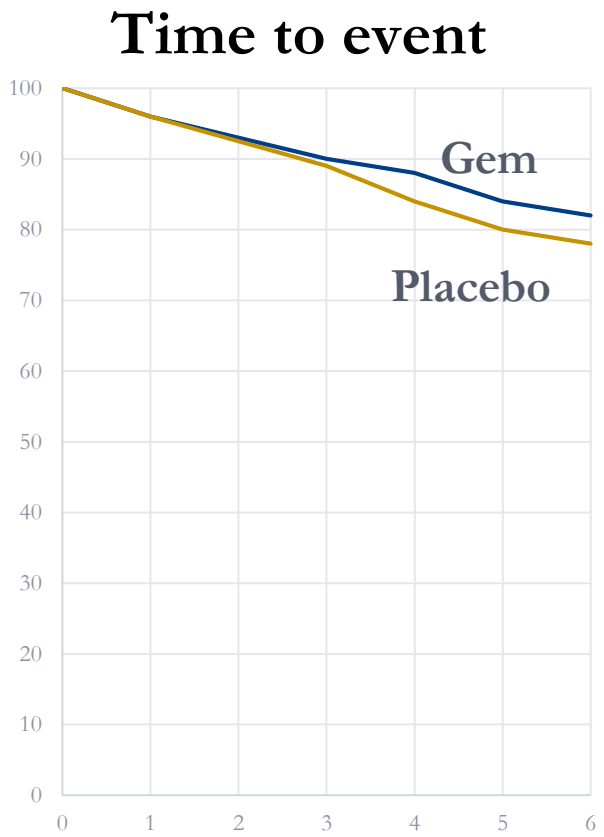
Rubins HB et al. (1999) NEJM 341:410-418

The four measures for VA-HIT: HR



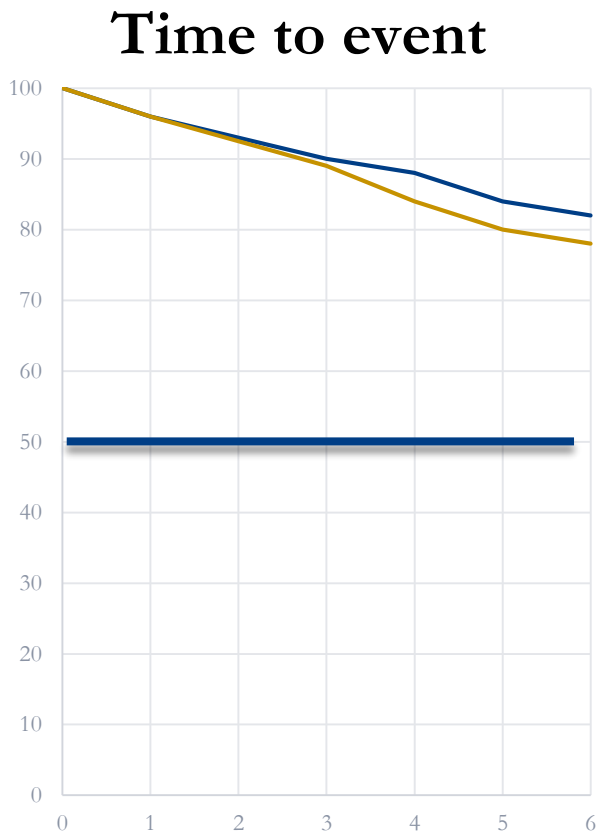
- HR = 0.78; p=0.006
  - 95% CI: (0.65, 0.93)
  - Reduction in rel risk: 22%; 95% CI, 7%- 35%

The four measures for VA-HIT: HR



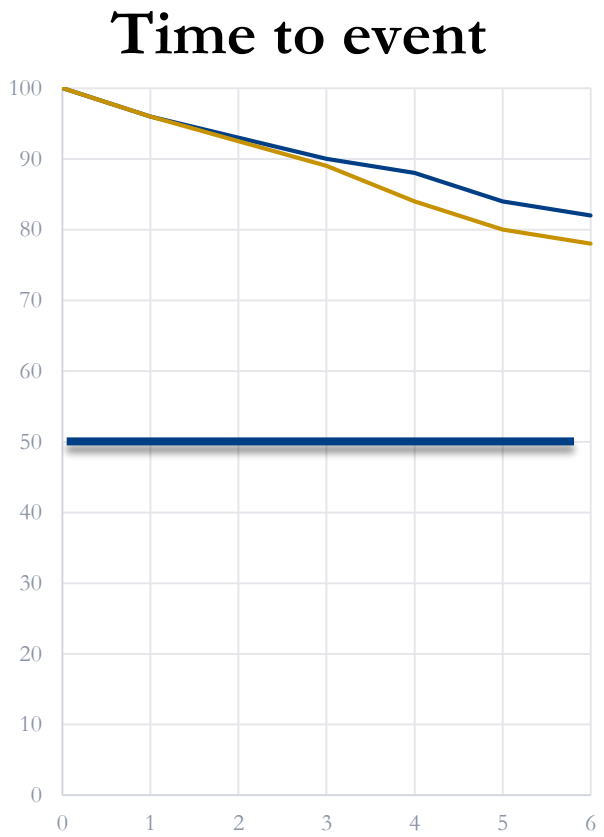
- HR = 0.78; p=0.006
  - 95% CI: (0.65, 0.93)
- What to tell patient
  - Reduction in rel risk: 22%; 95% CI, 7%- 35%

Median



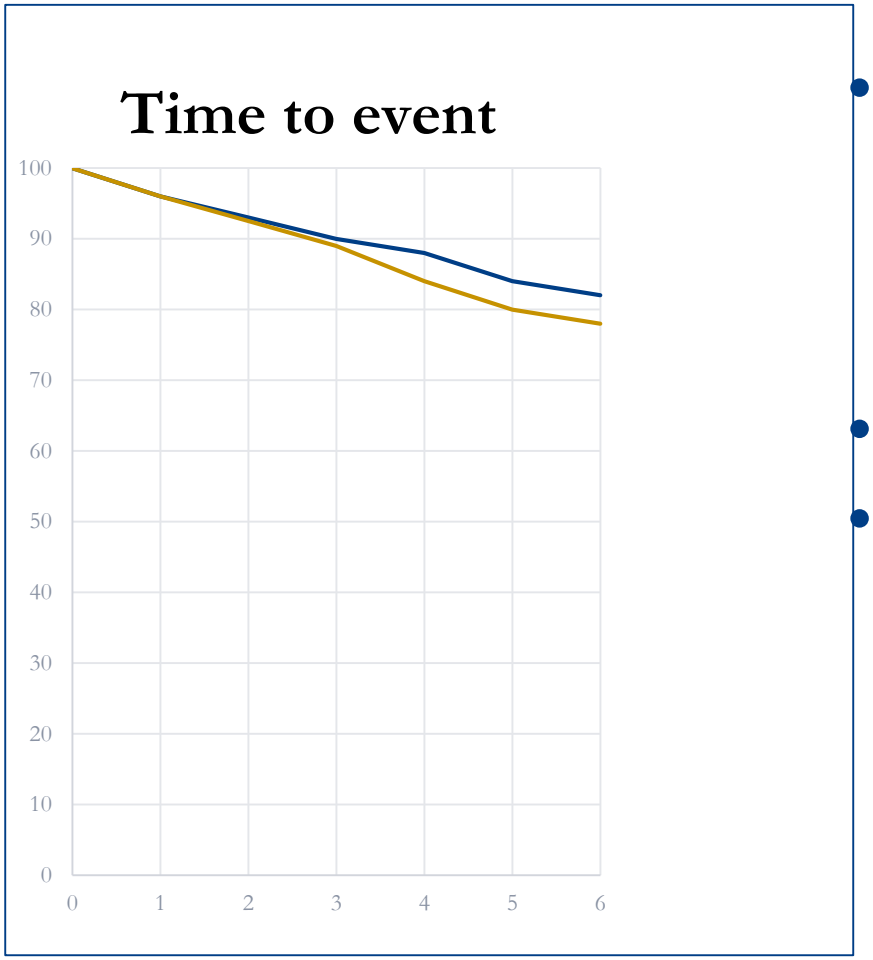
- HR = 0.78; p=0.006
  - 95% CI: (0.65, 0.93)
  - Reduction in rel risk: 22%; 95% CI, 7%- 35%
- Median – can't calculate

Median



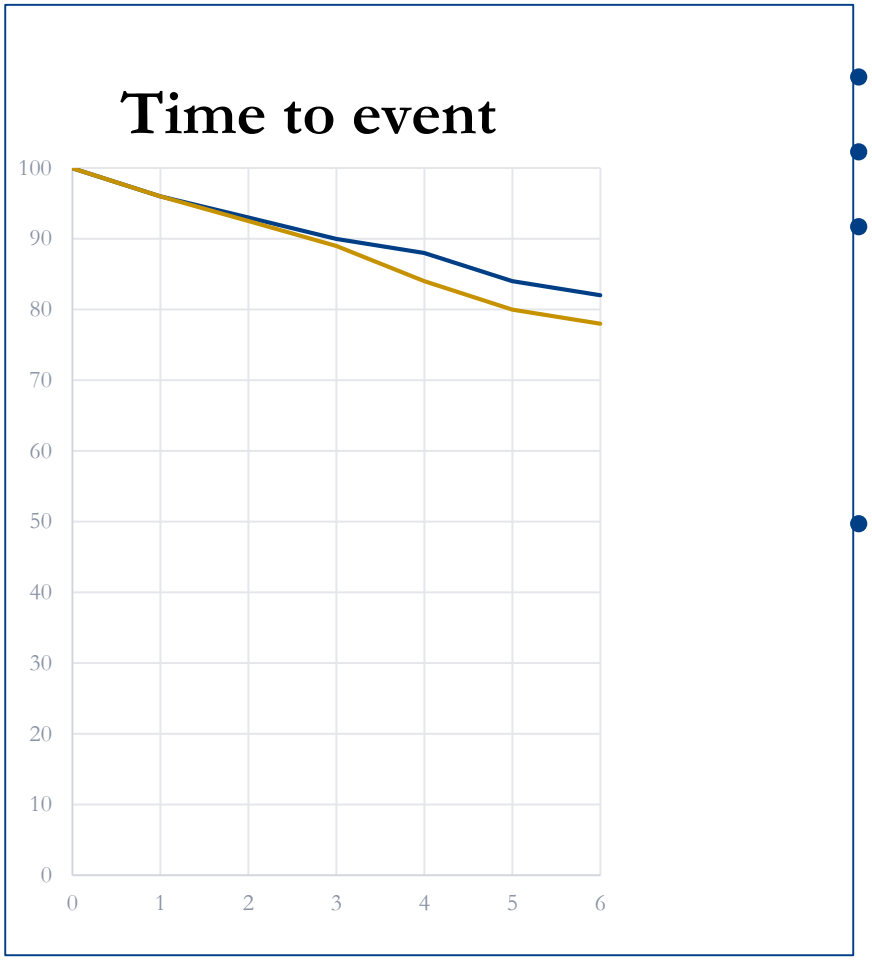
- HR = 0.78; p=0.006
- Median – can't calculate
- What to tell your patient
  - Prob of 5 or 6 year survival

6 year RMST



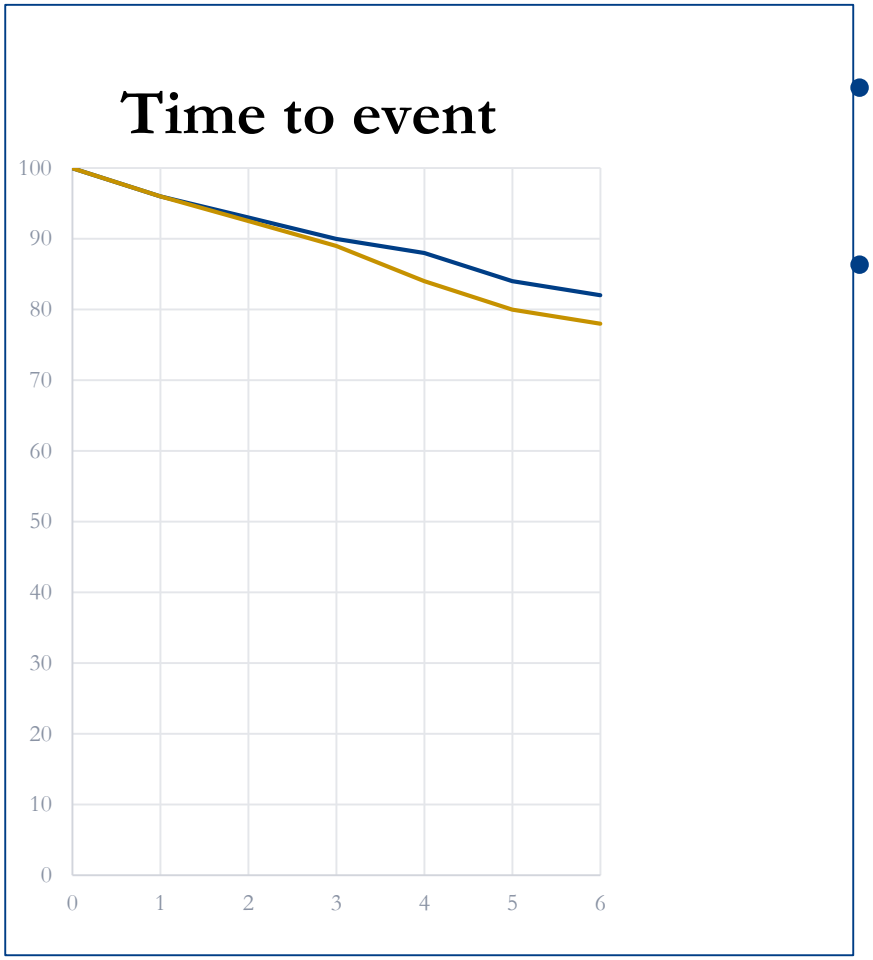
- HR = 0.78; p=0.006
  - 95% CI: (0.65, 0.93)
  - Reduction in rel risk: 22%; 95% CI, 7%- 35%
- Median – can't calculate
- **6 year RMST –**
  - **Gem: 4.9 years**
  - **Pbo: 4.6 years**
  - **Diff: 3.6 months**

6 year RMST



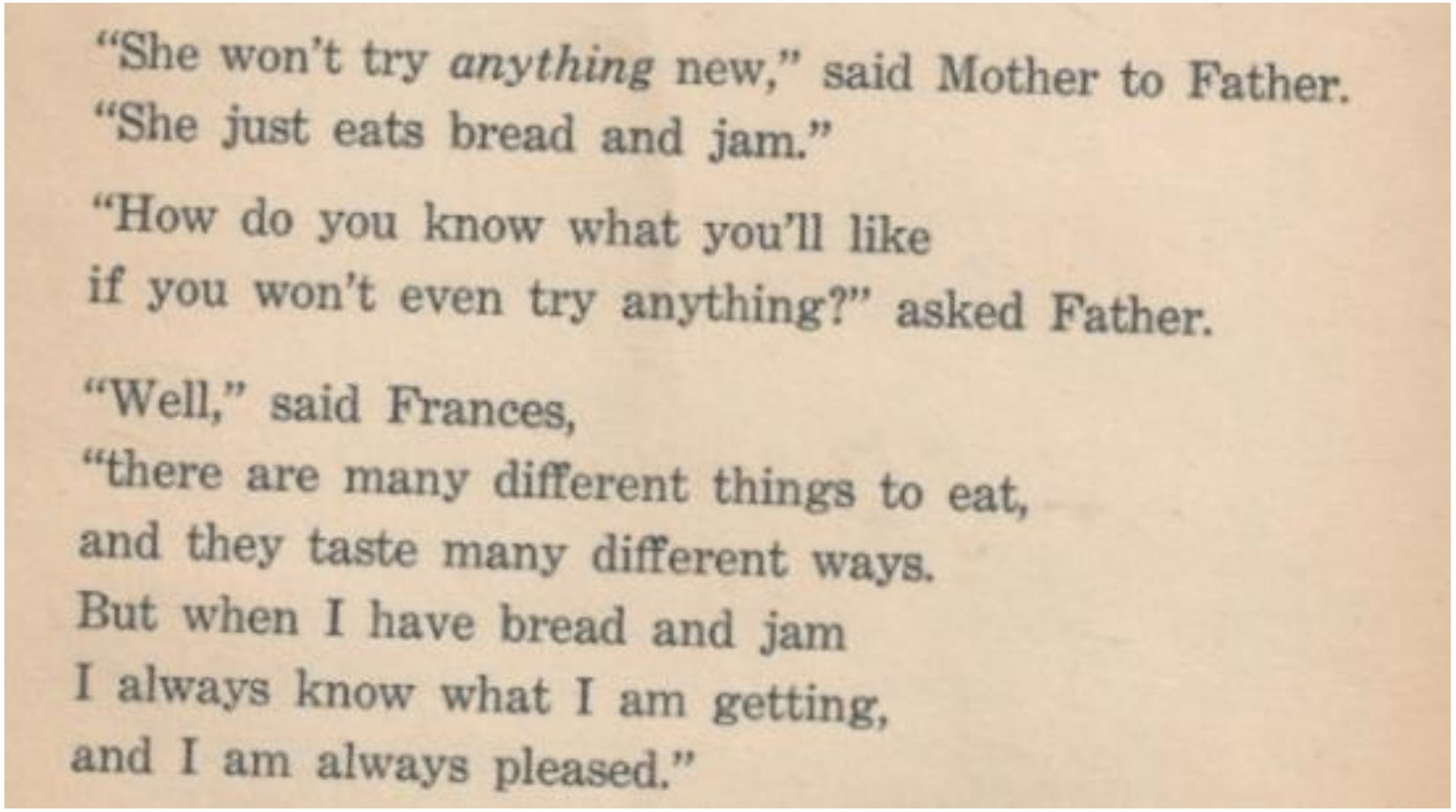
- HR = 0.78; p=0.006
- Median – can't calculate
- **6 year RMST –**
  - **Gem: 4.9 years**
  - **Pbo: 4.6 years**
  - **Diff: 3.6 months**
- **What to say to patient**
  - **Over 6 years, you are likely to be event-free for about 3 and a half months longer**
  - **My guess is that you would go to probabilities**

Absolute Difference in Probability (6 yr)



- 6 yr AD: 82% vs. 78%
  - Difference =4%
- Easy to talk about
  - Do people understand?

So if HR is the loser, why do we use it?



"She won't try *anything* new," said Mother to Father.
"She just eats bread and jam."

"How do you know what you'll like
if you won't even try anything?" asked Father.

"Well," said Frances,
"there are many different things to eat,
and they taste many different ways.
But when I have bread and jam
I always know what I am getting,
and I am always pleased."

Russell and Lillian Hoban. Bread and Jam for Frances. Scholastic.

So if HR is the loser, why do we use it?

- We think we understand it.
  - We think we know what a HR of 0.8, or 1.6, etc. means

- It has really nice mathematical properties
- We grew up with it
- And…

Once in love with Amy
Always in love with Amy
Ever and ever fascinated by her
Sets your heart on fire to stay

How to convert from your favorite

- For those who love medians, break up with them!
  - They tell you nothing about the long term
  - If you have low risk events, you can't calculate a median
- If you have nicely behaved curves (~proportional hazards)
  - Stick with HR augmented
- If you have curves that you expect will cross or splay
  - Think of RMST or AD
  - But you have to choose your primary up front
- No matter what you choose, look at the survival curves
  - …but don't take them too literally!

What about HR for recurrent events?

Topics

- Why we like time-to-first event
- Why analysis of recurrent events is attractive
- Some taxonomies
  - *Types* of recurrent events
  - Scientific *questions*
  - *Methods* of analysis
- How to match *types* with *questions* with *methods*

Why time-to-first event is attractive

- It's simple to
  - Analyze statistically
  - Interpret clinically
- We can graph it
  - Once people understand K-M curves, they can read them
- Doesn't require many assumptions
  - Except proportional hazards for HR

Time to first helps to answer:

What's the median survival time? – if you can measure median

What's the probability of survival at, say, 24 months?

Are events occurring earlier in one group than another?

Are trends similar, then split at a certain time?

The mean survival time (or the restricted mean survival time)

More reasons for time-to-first event

- It's simple to
  - Analyze statistically
  - Interpret clinically
- Doesn't require many assumptions
- We can graph it and one graph speaks 1000 words
- When an event occurs, the trajectory of the patient may change
  - Hard to disentangle effect of test drug from later treatments
  - (From an intent-to-treat point of view that's fine)
  - (Not a relevant problem in some conditions)

The big problem

- Time to first fails to address the total impact on the disease

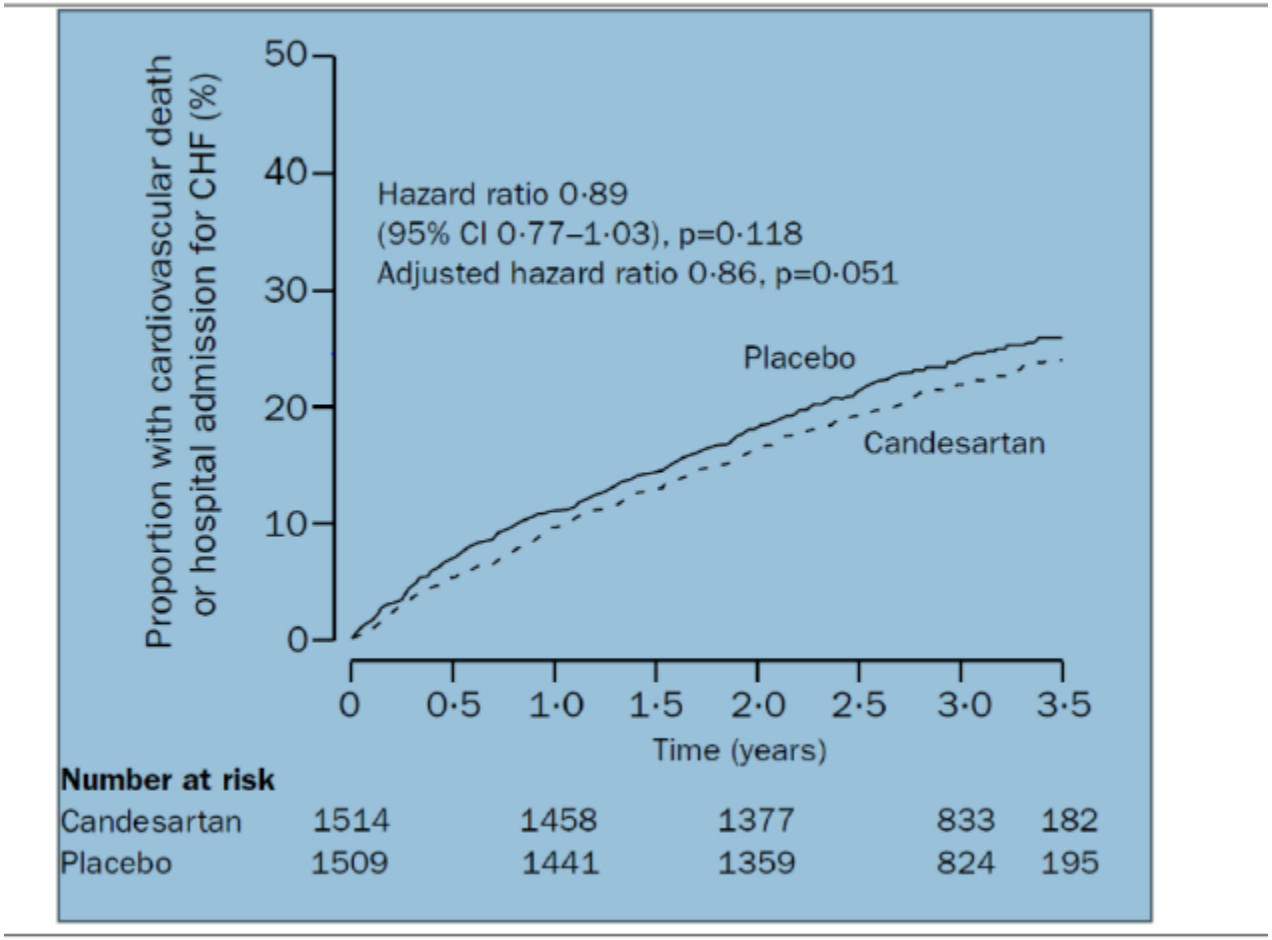
Used frequently in diseases characterized by exacerbations and flares

- Asthma
- COPD
- Cystic fibrosis
- Hemophilia
- Sickle cell disease
- MS
- Gout
- Epilepsy

Cardiology

- Heart failure – hospitalizations – analogous to "flare"
  - Complicated because of censoring by death
  - Disease severity worsens over time
- Coronary disease, hypertension, etc. use a composite outcome
  - MACE (MI and stroke and death) or variant
  - Not recurrent, so "counting" the number of events is harder

Example: CHARM-preserved



But that only counts first event

Heart Failure Hospitalizations	Candesartan (N=1513)	Placebo (N=1508)
At least one	230	278
All admissions	392	547
Admissions time-to-first ignores	162	269

If we count all hospitalizations….

Heart Failure Hospitalizations	Candesartan (N=1513)	Placebo (N=1508)	P/C
At least one	230	278	1.2
All admissions	392	547	1.4

Method	Hazard Ratio	95% CI	Width of CI	p-value
Time to first	0.89	(0.77, 1.03)	0.26	0.12
Three common recurrent event methods				
Poisson	0.71	(0.62, 0.81)	0.19	<0.001
Negative binomial	0.68	(0.54, 0.85)	0.31	<0.001
Counting process*	0.71	(0.57, 0.88)	0.31	0.002

*AKA: Anderson-Gill

If we count them all….

Hazard ratios for recurrent events – almost the same
But, the width of the confidence intervals are quite different.
What is going on?

Method	Hazard Ratio	95% CI	Width of CI	p-value
Time to first	0.89	(0.77, 1.03)	0.26	0.12
Recurrent event methods				
Poisson	0.71	(0.62, 0.81)	0.19	<0.001
Negative binomial	0.68	(0.54, 0.85)	0.31	<0.001
Counting process	0.71	(0.57, 0.88)	0.31	0.002

Taxonomy #1 – types of recurrent events

Types of recurrent event
All people have same rate; events occur randomly at constant rate (analogy: lump of uranium)
People have different rates; events occur randomly at constant rate (e.g.: uranium & radium mix)
All people have same rate but events within person are not random*
People have different rates and events within person are not random
All of the above types but events are not the same (e.g., MACE) – not really "recurrent"

* Or we can stratify to get different rates.

Taxonomy #2 - questions

Types of recurrent event	Questions
All people have same rate; events occur randomly	Is the number of events in the study period different in the intervention and control groups?
People have different rates; events occur randomly	Compared to the control, what is the average number of events the intervention prevents?
All people have same rate; events within person are not random	What is the effect of the intervention on events after the first event? (e.g., how many second events are prevented? 3rd?)
All of the above; events are not the same (e.g., MACE)	How does the intervention affect the probability of an event among those who have a first event?
	Does intervention affect total disease "burden"?
	What trajectories do people prefer?

Taxonomy #3 – Types of analyses

Types of recurrent event	Questions	Analyses
All people have same rate; events occur randomly	Number of events	Poisson regression
People have different rates; events occur randomly	Average number prevented	Negative binomial regression
All people have same rate; events within person not random	Effect on later events	Counting process
All of the above but events are not the same	Effect on later events among those who have a first event?	Days alive and not sick (e.g., or not in hospital)
	Effect on disease "burden"?	Finkelstein-Schoenfeld; General methods of ordering trajectories; Win-ratio
	Preferences	

How do we match columns?

Types of recurrent event	Questions	Analyses
All people have same rate; events occur randomly	Number of events	Poisson regression
People have different rates; events occur randomly	Average number prevented	Negative binomial regression
All people have same rate; events within person not random	Effect on later events	Counting process
All of the above but events are not the same	Effect on later events among those who have a first event?	Days alive and not sick (e.g., or not in hospital)
	Effect on disease "burden"?	Finkelstein-Schoenfeld; General methods of ordering trajectories; Win-ratio
	Preferences	

How do we match columns?

Types of recurrent event	Questions	Analyses
All people have same rate; events occur randomly	Number of events	Poisson regression
People have different rates; events occur randomly	Average number prevented	Negative binomial regression
All people have same rate; events within person not random	Effect on later events	Counting process
All of the above but events are not the same	Effect on later events among those who have a first event?	Days alive and not sick (e.g., or not in hospital)
	Effect on disease "burden"?	Finkelstein-Schoenfeld; General methods of ordering trajectories; Win-ratio
	Preferences	

How do we match columns?

Types of recurrent event	Questions	Analyses
All people have same rate; events occur randomly	Number of events	Poisson regression
People have different rates; events occur randomly	Average number prevented	Negative binomial regression
All people have same rate; events within person not random	Effect on later events	Counting process
All of the above but events are not the same	Effect on later events among those who have a first event?	Days alive and not sick (e.g., or not in hospital)
	Effect on disease "burden"?	Finkelstein-Schoenfeld; General methods of ordering trajectories; Win-ratio
	Preferences	

Matching columns

Types of recurrent event	Questions	Analyses
All people have same rate; events occur randomly	Number of events	Poisson regression
People have different rates; events occur randomly	Average number prevented	Negative binomial regression
All people have same rate; events within person not random	Effect on later events	Counting process (but watch this one – not protected by randomization)
All of the above but events are not the same	Effect on later events among those who have a first event?	Days alive and not sick (e.g., or not in hospital)
	Effect on disease "burden"?	Finkelstein-Schoenfeld; General methods of ordering trajectories; Win-ratio
	Preferences	

Matching columns

Types of recurrent event	Questions	Analyses
All people have same rate; events occur randomly	Number of events	Poisson regression
People have different rates; events occur randomly	Average number prevented	Negative binomial regression
All people have same rate; events within person not random	Effect on later events	Counting process
All of the above but events are not the same	Effect on later events among those who have a first event?	Days alive and not sick (e.g., or not in hospital)
	Effect on disease "burden"?	Finkelstein-Schoenfeld and Win-ratio (prespecified preferences); General methods of ordering trajectories (Follman/Wittes; Armstrong et al.)
	Preferences	

Measuring the effect of the intervention

- Total number of events over a fixed time period
  - Study with fixed follow-up period
  - Count the average number of events in treated and control
  - Makes no assumption about rate or constancy
- Rate of events: rate per unit time
  - Accounts for different time per person
  - Assumes rate is constant
- Times to successive events – "how long do I wait for an event"
- Times between successive events –"how long am I event-free?"

If we have trouble understanding HR in time to first…

- Recurrent events even harder outside Poisson and negative binomial
- Assumes subjects are in the risk set for the k^{th} event from the time of the $(k-1)^{the}$ event
- Therefore, does not respect the randomisation
- So the HRs for 2^{nd}, 3^{rd},… events are not interpretable causally.
- To arrive at a single HR estimate one has to average treatment effects over occurrences or assume they are the same
- Some models assume all subjects are at risk for the 1^{st}, 2^{nd}, 3^{rd} … event at the same time. That preserves randomization but does not seem sensible

Why I am perplexed

- (and I think you should be too)
- I want to
  - Preserve the randomization
  - Summarize the survival curve in one number
  - Count all events
  - But I can't have my cake and eat it
- So, before jumping in to choosing a method, think and discuss