# Evaluating Survival Prognosis in the Presence of Immortal Time Bias

for *a meeting of the research group of Prof. Sandrine Dudoit*,
given Tuesday 20th February, 2018

Kevin Benac and Nima Hejazi

Group in Biostatistics
University of California, Berkeley

slides: goo.gl/Cqd9ex

This slide deck is meant to serve as a template for building presentations conveniently from a reasonable baseline.

Slides: `https://goo.gl/Cqd9ex`
With notes: `https://goo.gl/PFeKRH`

## Data and Motivation

- Consider a data analysis scenario in which we are given survival times for patients recruited based on a first primary melanoma.
- Over the course of the observational study, an *a priori* unknown number of the patients ($n_2$) develop a second primary melanoma prior to death.
- **Question of interest:** *How does the occurrence of a second primary melanoma change the survival prognosis of a patient?*

# Preview: Summary

- Nonparametric estimators of survival (even the NP-MLE) displays bias under this data-generating mechanism.
- The Cox proportional hazards model provides a way to mitigate this bias but comes with assumptions that are difficult to verify in practice.
- Youlden provides an approach that appears intuitive but fails to approach the parameter of interest.
- Jewell provides a correction for employing the Kaplan–Meier estimator in this setting.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

# Survival Analysis

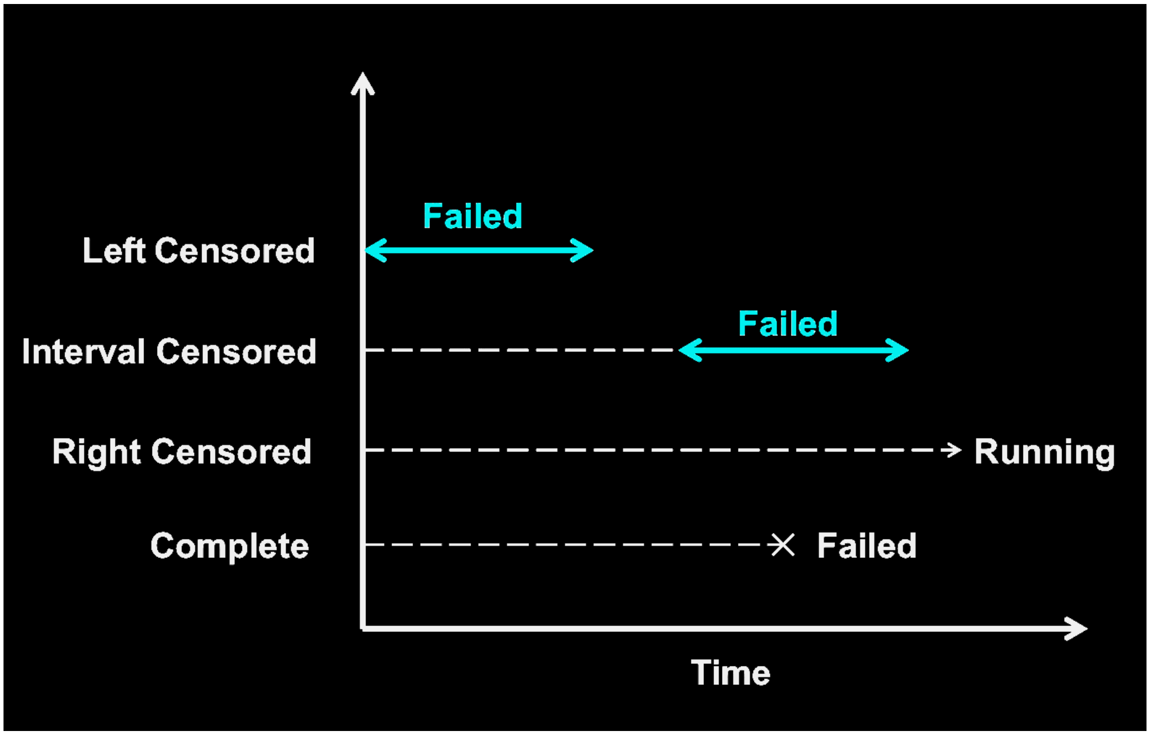
- Study of the distribution of a lifetime $T$, corresponding to the time from a **well-defined origin** until the occurrence of a **well-defined event** or **endpoint**.
- For historical reasons, the event is often referred to as a **failure**.
- An individual for whom an event has not occurred at time $t$ is said to be **at risk** at time $t$.
- Although the term *failure* is usually associated with *death*, especially in medical research, it has to be taken in the broad sense of a well-defined event.

# Survival Analysis

- $T$ is a non-negative random variable.
- If $T_1, \ldots, T_n \overset{iid}{\sim} T$ are observed, then we know that the empirical cumulative distribution function (eCDF) is the NP-MLE of $F_T(\cdot)$.
- *Problem:* In practice, we always have to deal with missing data (i.e., **censored observations**) in the context of survival data. Most of the time, this is merely **right-censoring**.

- A second problem that obviously makes this estimator way too simple is that, like usual, covariate information should be taken into account in the estimation of $F_T(\cdot)$.

# Survival Analysis



Caption: Different types of censoring. The dashed lines correspond to the period during which the individual is followed; red intervals correspond to time-intervals during which the individual is not followed and failed.

# The Kaplan–Meier Estimator

- Kaplan and Meier (1958) extensively studied the case where right-censored data are present in survival analysis.

- Let us denote the distinct ordered times of observed failures by
$$t^{(1)} < \cdots < t^{(m)},$$

| Time | $t^{(1)}$ | $t^{(2)}$ | $\ldots$ | $t^{(m)}$ |
|---|---|---|---|---|
| Failures | $d_1$ | $d_2$ | $\ldots$ | $d_m$ |
| At risk | $n_1 = n$ | $n_2$ | $\ldots$ | $n_m$ |

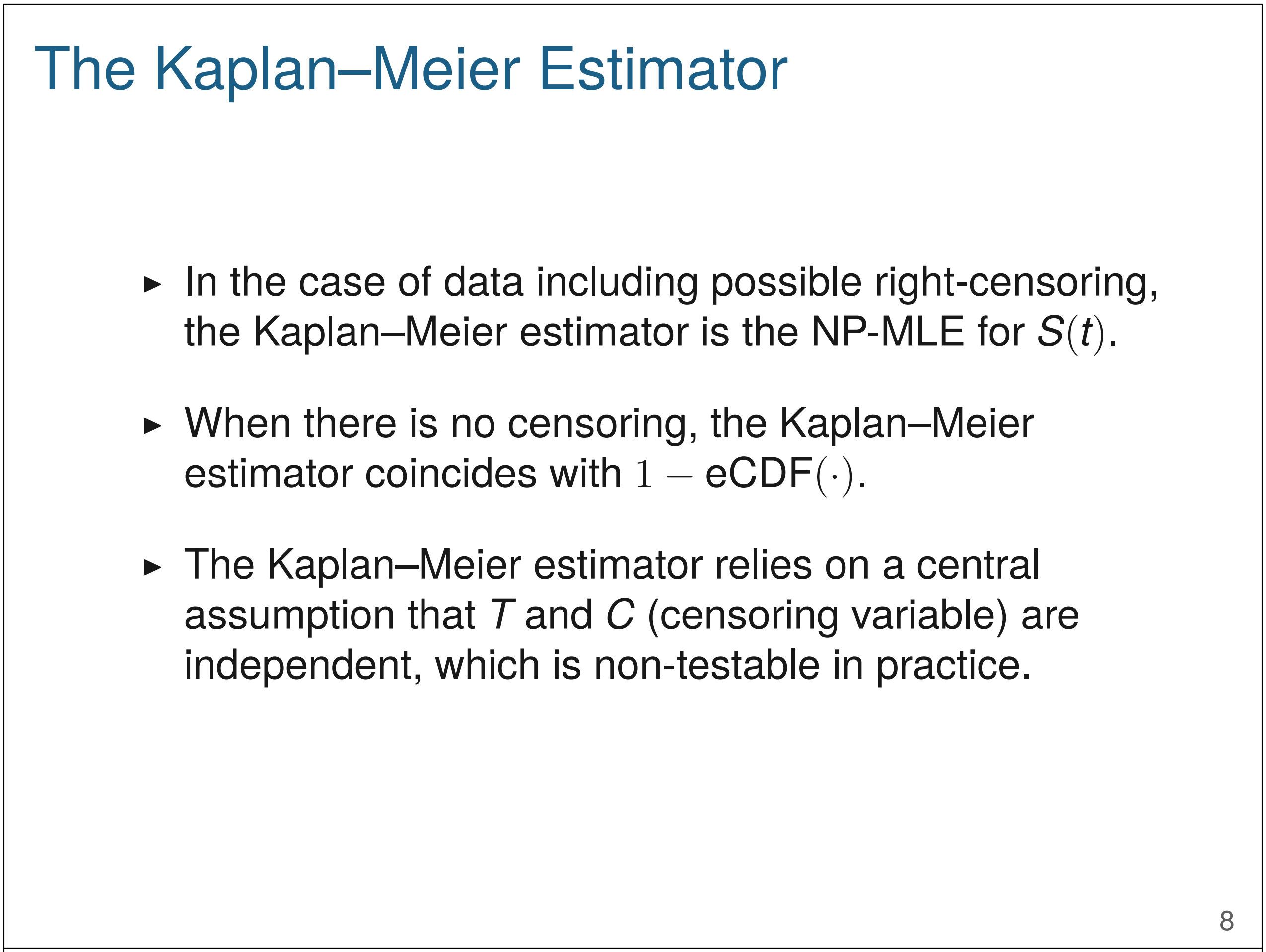
# The Kaplan–Meier Estimator

If $t > 0$, $t^{(i)} < t \leq t^{(i+1)}$ then we can decompose $S(t^{(i)})$ as

$$P\left\{T > t^{(1)}\right\} P\left\{T > t^{(2)} \mid T > t^{(1)}\right\} \cdots P\left\{T > t^{(i)} \mid T > t^{(i-1)}\right\}.$$

The Kaplan–Meier estimator is defined as

$$\widehat{S}(t) = \prod_{i:t(i)<t} \left(1 - \frac{d_i}{n_i}\right), \quad t \geq 0.$$



# The Kaplan–Meier Estimator

- In the case of data including possible right-censoring, the Kaplan–Meier estimator is the NP-MLE for $S(t)$.
- When there is no censoring, the Kaplan–Meier estimator coincides with $1-$ eCDF$(\cdot)$.
- The Kaplan–Meier estimator relies on a central assumption that $T$ and $C$ (censoring variable) are independent, which is non-testable in practice.

An important and problematic limitation of Kaplan–Meier is that it does not allow us to use covariate information.

# Hazard Function

The *hazard function* at time $t$ is defined by

$$\lambda(t) = \lim_{h \to 0} \frac{P\left(T < t + h \mid T \geq t\right)}{h} = \frac{f(t)}{S(t)}, \quad t > 0.$$

The hazard and the survival functions are related by

$$S(t) = \exp\left\{-\Lambda(t)\right\}, \quad t > 0,$$

where

$$\Lambda(t) = \int_0^t \lambda(s) ds, \quad t > 0$$

and is known as the **cumulative hazard function**.

# The Cox Proportional Hazards Model

- The most widely used regression model in survival analysis is Cox's *proportional hazards* model (of the hazard function):

$$\lambda\left(t; Z = z\right) = \lambda_0(t) \exp\left(\beta^T z\right), \quad t \geq 0.$$

- This is a *semiparametric* model: nonparametric in $\lambda_0(\cdot)$ but parametric in $\beta$.

Note that we refer to the Cox proportional hazards model as semiparamtric in the classical sense — that is, partly nonparametric and partly parametric. (This is different from the sense in which semiparametric is used in the missing data / causal inference communities (e.g., J.M. Robins, A. Tsiatis, M.J. van der Laan).

# Data and Motivation

- *Problem:* Efficiently estimate survival prognosis for a data structure exhibiting immortal time bias.
- *Why?* Efficient estimation under a time-dependent risks bias presents a novel challenge that has received meager attention in the literature.
- We employ and compare
  1. semiparametric estimators of survival: the Cox proportional hazards model (with time-varying covariates),
  2. Nonparametric estimators of survival: variations of the the Kaplan–Meier estimator.

We simulate data under the assumptions of the Cox proportional hazards model, evaluating the efficiency of each estimator so as to better develop an understanding of how to analyze the observed data we expect.

# The Cox Proportional Hazards Model

- In the case we consider for motivation, let $U$ be the time where a second event occurs, then define $Z(t) = I(t > U), \quad t \geq 0.$
- Fitting the Cox model with $Z(t)$ as a covariate enables us to estimate how the risk for the patient changes after the appearance of the second melanoma.

The Cox model may be extended to deal with time-dependent covariates, Martingale theory, etc.

## Methodology — Cox Regression

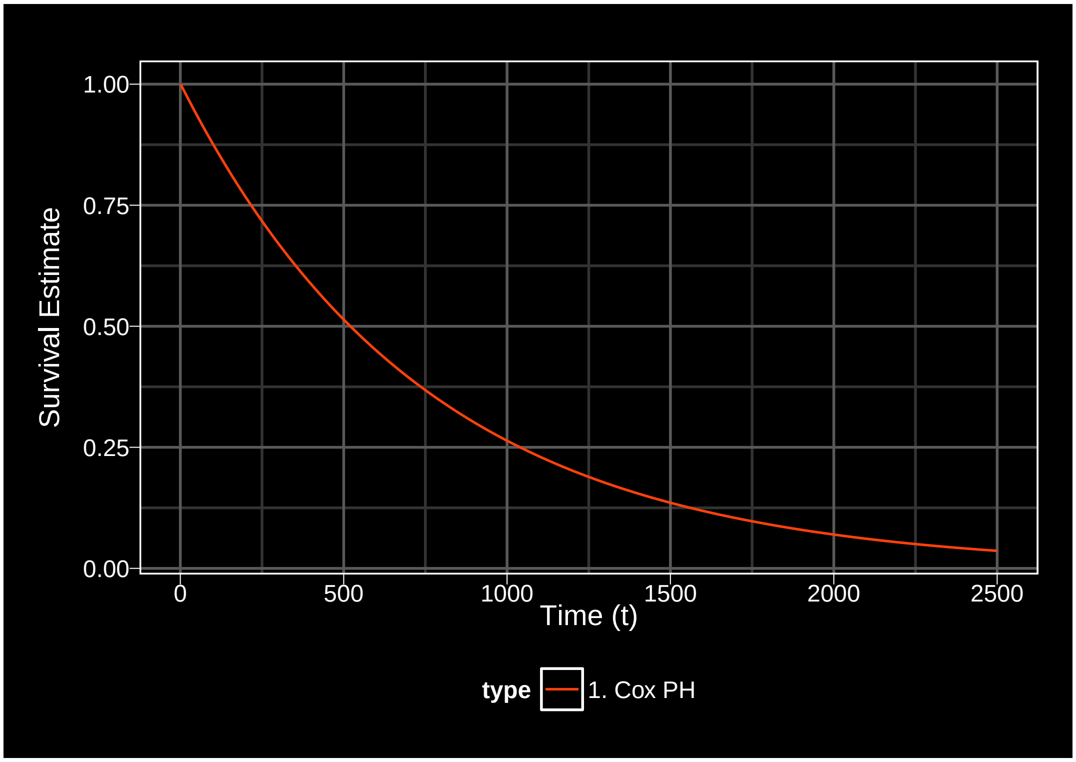
- We simulate observed data under the assumptions of the Cox model a total of $10,000$ times, averaging the estimated survival across all observed time points for each fit of the Cox proportional hazards regression.

- Recall that Cox regression estimates the hazard at a given time, assuming a simplistic relationship between hazards for events of interest.

- This borrows information across the two groups to estimate survival — that is, groups experiencing a single primary melanoma and those with two both inform estimation of survival.

- We account for transitions between the two groups by way of a *time-varying covariate*.

# Results — Cox Proportional Hazards

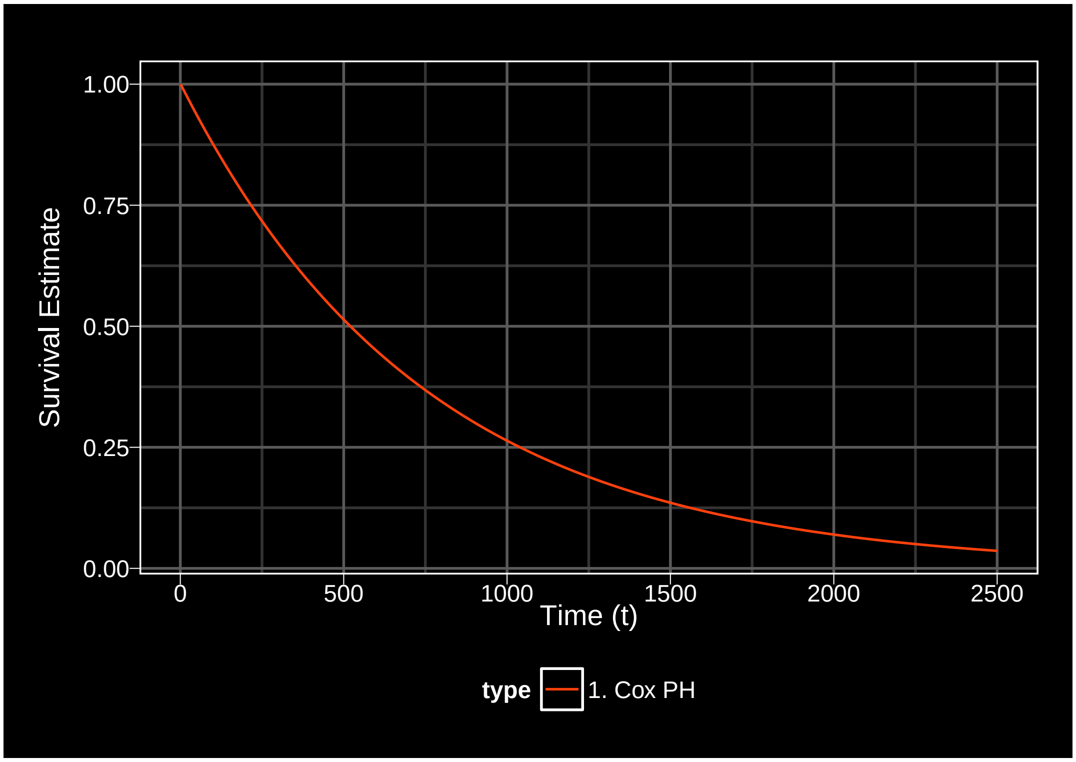


## Methdology — Kaplan–Meier (Youlden)

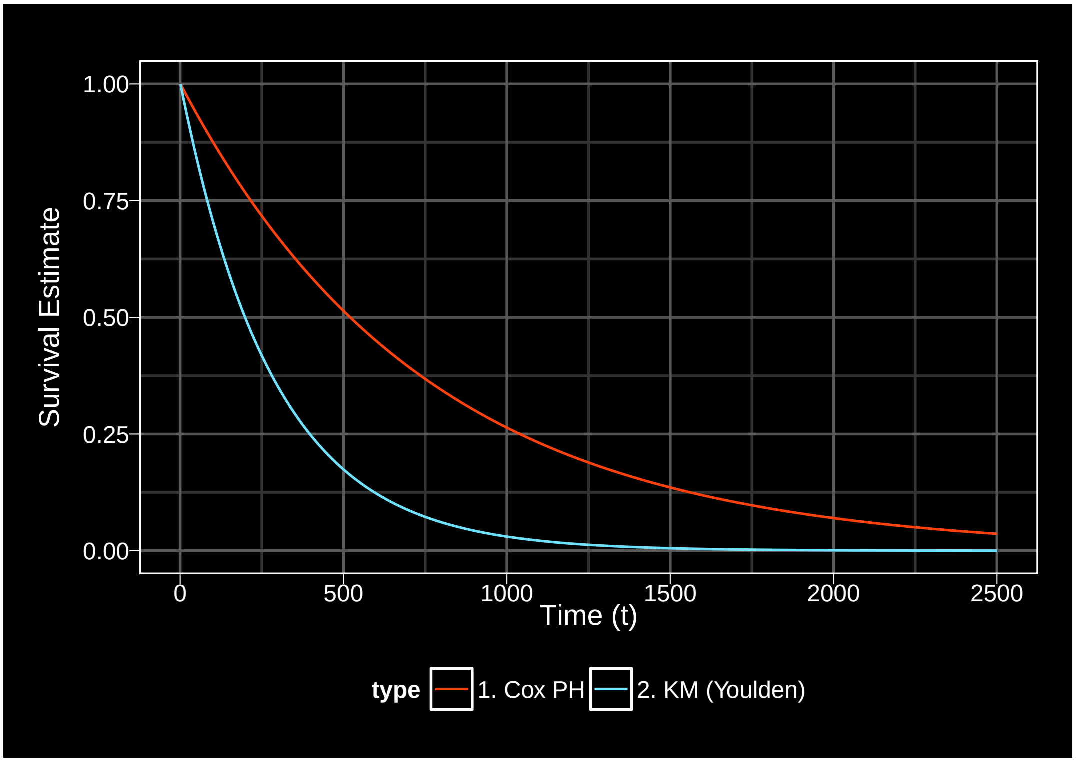
- In pratice it's impossible to know if assumptions of the Cox model hold for a given data-generating process we encounter “in the wild.”
- This makes the use of a nonparametric approach highly desirable, so, *how might we formulate a Kaplan–Meier estimator for this setting*?
- Recall that we cannot provide covariate information when fitting Kaplan–Meier estimators, let alone time-varying disease status.
- *Proposal:* Fit a Kaplan–Meier estimator for patients that experience only a single primary melanoma.

We expect that this will recover the survival function for the risk conferred by experiencing a single primary melanoma.

# Results — Kaplan–Meier (Youlden)



# Results — Kaplan–Meier (Youlden)

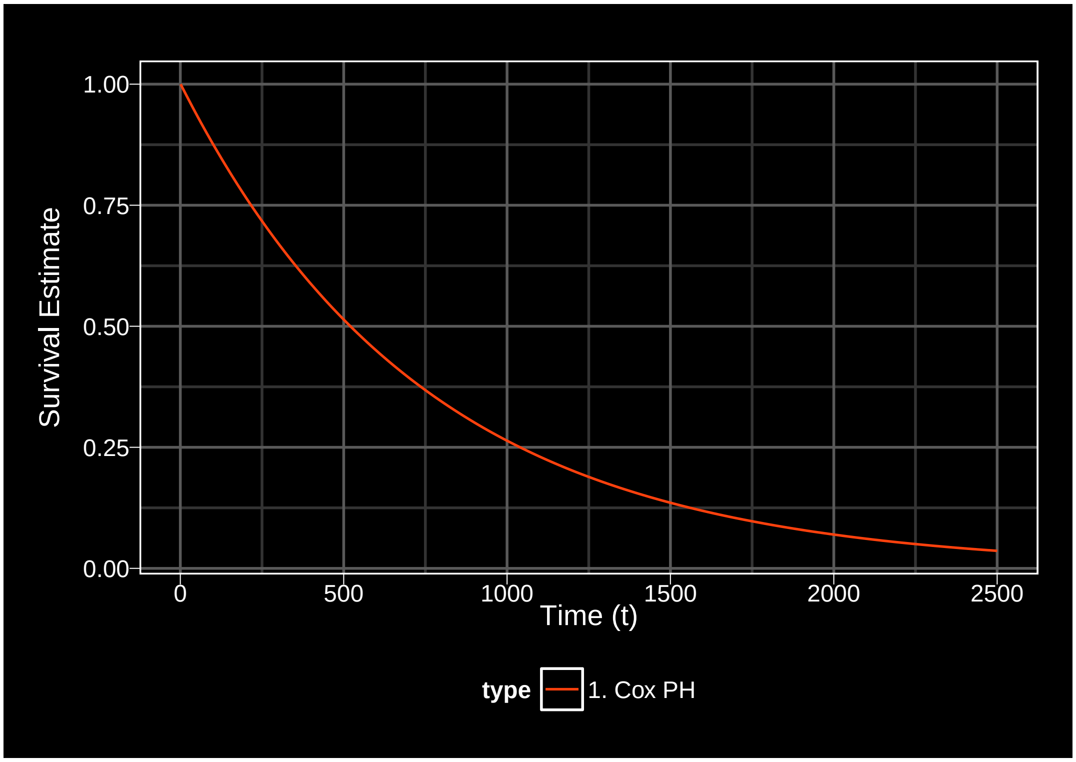


# Methdology — Kaplan–Meier (Jewell)

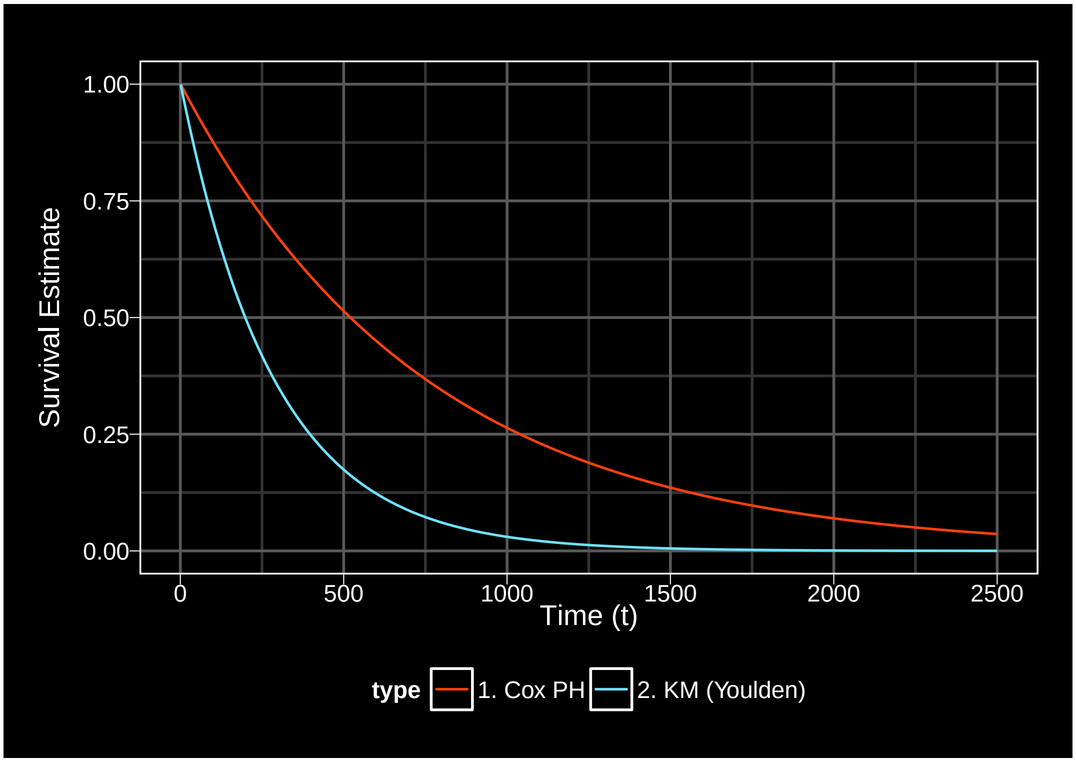
- The difference between the Kaplan–Meier and Cox regression estimates of survival is quite striking.
- Given that the assumptions of the Cox model hold, we can evaluate Kaplan–Meier relative to Cox — why is KM so strongly biased?
- Recall that we *chose* to discard the second group when fitting our chosen KM estimator.
- Including such observations when fitting our KM estimator should *de-bias* the early (in time) estimates.

- It appears that the Youlden approach displays a great degree of bias across all observed times.
- We need a better way in which to estimate survival — one that accounts for the time-varying group status.
- Importantly, all patients in the second group are guaranteed to survive until their occurrence of the second primary melanoma.

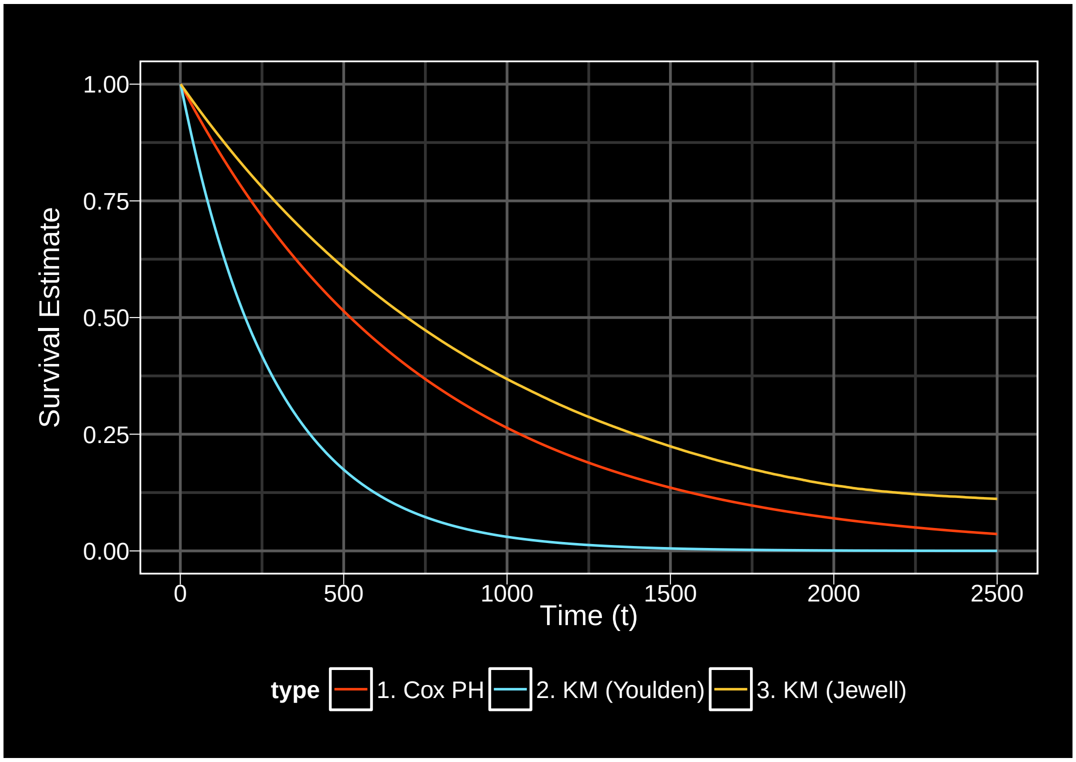
# Results — Kaplan–Meier (Jewell)



# Results — Kaplan–Meier (Jewell)



# Results — Kaplan–Meier (Jewell)



# Discussion

- Each of three estimators provides strikingly different estimates of survival across the observed times.
- Since the assumptions of the Cox model hold in our simulation, we can evaluate results from each of the nonparametric KM estimators relative to those from Cox PH.
- Further/ongoing investigation: *How does the relative performance of these estimators differ when the assumptions of the Cox model do not hold?*

# Discussion

- Our initial KM proposal exhibited strong bias, underestimating survival at all observed times, with a particularly noticeable bias early on.
- The corrected KM estimator, which draws on information from both groups, provides better estimates early in time, but displays a stronger positive bias at later times (overestimating survival).

## Review: Summary

- Nonparametric estimators of survival (even the NP-MLE) displays bias under this data-generating mechanism.
- The Cox proportional hazards model provides a way to mitigate this bias but comes with assumptions that are difficult to verify in practice.
- Youlden provides an approach that appears intuitive but fails to approach the parameter of interest.
- Jewell provides a correction for employing the Kaplan–Meier estimator in this setting.

It's always good to include a summary.

# References I

# Acknowledgments

Nicholas P. Jewell University of California, Berkeley

## Thank you. Questions?

- Nonparametric estimators of survival (even the NP-MLE) displays bias under this data-generating mechanism.
- The Cox proportional hazards model provides a way to mitigate this bias but comes with assumptions that are difficult to verify in practice.
- Youlden provides an approach that appears intuitive but fails to approach the parameter of interest.
- Jewell provides a correction for employing the Kaplan–Meier estimator in this setting.

It’s good to include a summary when taking questions.