Survival Analysis

Survival Analysis

- Survival analysis concerns a special kind of outcome variable: the *time until an event occurs*.

Survival Analysis

- Survival analysis concerns a special kind of outcome variable: the *time until an event occurs*.
- For example, suppose that we have conducted a five-year medical study, in which patients have been treated for cancer.

Survival Analysis

- Survival analysis concerns a special kind of outcome variable: the *time until an event occurs*.
- For example, suppose that we have conducted a five-year medical study, in which patients have been treated for cancer.
- We would like to fit a model to predict patient survival time, using features such as baseline health measurements or type of treatment.

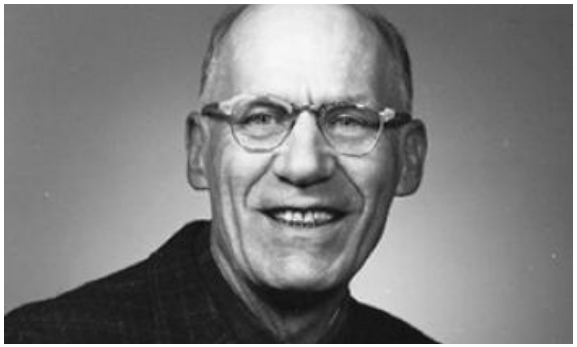
Survival Analysis

- Survival analysis concerns a special kind of outcome variable: the *time until an event occurs*.
- For example, suppose that we have conducted a five-year medical study, in which patients have been treated for cancer.
- We would like to fit a model to predict patient survival time, using features such as baseline health measurements or type of treatment.
- Sounds like a *regression problem*. But there is an important complication: some of the patients have survived until the end of the study. Such a patient's survival time is said to be *censored*.

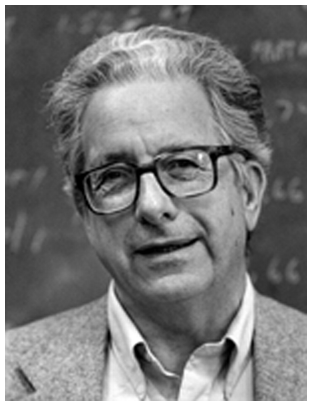
Survival Analysis

- Survival analysis concerns a special kind of outcome variable: the *time until an event occurs*.
- For example, suppose that we have conducted a five-year medical study, in which patients have been treated for cancer.
- We would like to fit a model to predict patient survival time, using features such as baseline health measurements or type of treatment.
- Sounds like a *regression problem*. But there is an important complication: some of the patients have survived until the end of the study. Such a patient's survival time is said to be *censored*.
- We do not want to discard this subset of surviving patients, since the fact that they survived at least five years amounts to valuable information.

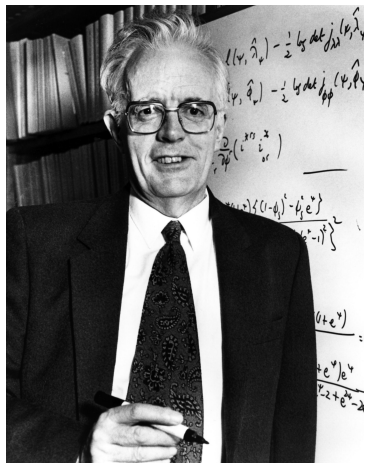
Some of the big names in this field



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

William Haenszel

(***log rank test***)

Terry Therneau

(***author of Survival package in R)***

Non-medical Examples

Non-medical Examples

- The applications of survival analysis extend far beyond medicine. For example, consider a company that wishes to model *churn*, the event when customers cancel subscription to a service.

Non-medical Examples

- The applications of survival analysis extend far beyond medicine. For example, consider a company that wishes to model *churn*, the event when customers cancel subscription to a service.
- The company might collect data on customers over some time period, in order to predict each customer's time to cancellation.

Non-medical Examples

- The applications of survival analysis extend far beyond medicine. For example, consider a company that wishes to model *churn*, the event when customers cancel subscription to a service.
- The company might collect data on customers over some time period, in order to predict each customer's time to cancellation.
- However, presumably not all customers will have cancelled their subscription by the end of this time period; for such customers, the time to cancellation is censored.

Non-medical Examples

- The applications of survival analysis extend far beyond medicine. For example, consider a company that wishes to model *churn*, the event when customers cancel subscription to a service.
- The company might collect data on customers over some time period, in order to predict each customer's time to cancellation.
- However, presumably not all customers will have cancelled their subscription by the end of this time period; for such customers, the time to cancellation is censored.
- Survival analysis is a very well-studied topic within statistics. However, it has received relatively little attention in the machine learning community.

Survival and Censoring Times

Survival and Censoring Times

- For each individual, we suppose that there is a true *failure* or *event* time T, as well as a true censoring time C.

Survival and Censoring Times

- For each individual, we suppose that there is a true *failure* or *event* time T, as well as a true censoring time C.
- The survival time represents the time at which the event of interest occurs (such as death).

Survival and Censoring Times

- For each individual, we suppose that there is a true *failure* or *event* time T, as well as a true censoring time C.
- The survival time represents the time at which the event of interest occurs (such as death).
- By contrast, the *censoring* is the time at which censoring occurs: for example, the time at which the patient drops out of the study or the study ends.

Survival and Censoring Times — Continued

Survival and Censoring Times — Continued

- We observe either the survival time T or else the censoring time C. Specifically, we observe the random variable

$$Y = \min(T, C).$$

Survival and Censoring Times — Continued

- We observe either the survival time T or else the censoring time C. Specifically, we observe the random variable

$$Y = \min(T, C).$$

- If the event occurs before censoring (i.e. $T < C$) then we observe the true survival time T; if censoring occurs before the event ($T > C$) then we observe the censoring time. We also observe a status indicator,

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C. \end{cases}$$

Survival and Censoring Times — Continued

- We observe either the survival time T or else the censoring time C. Specifically, we observe the random variable

$$Y = \min(T, C).$$

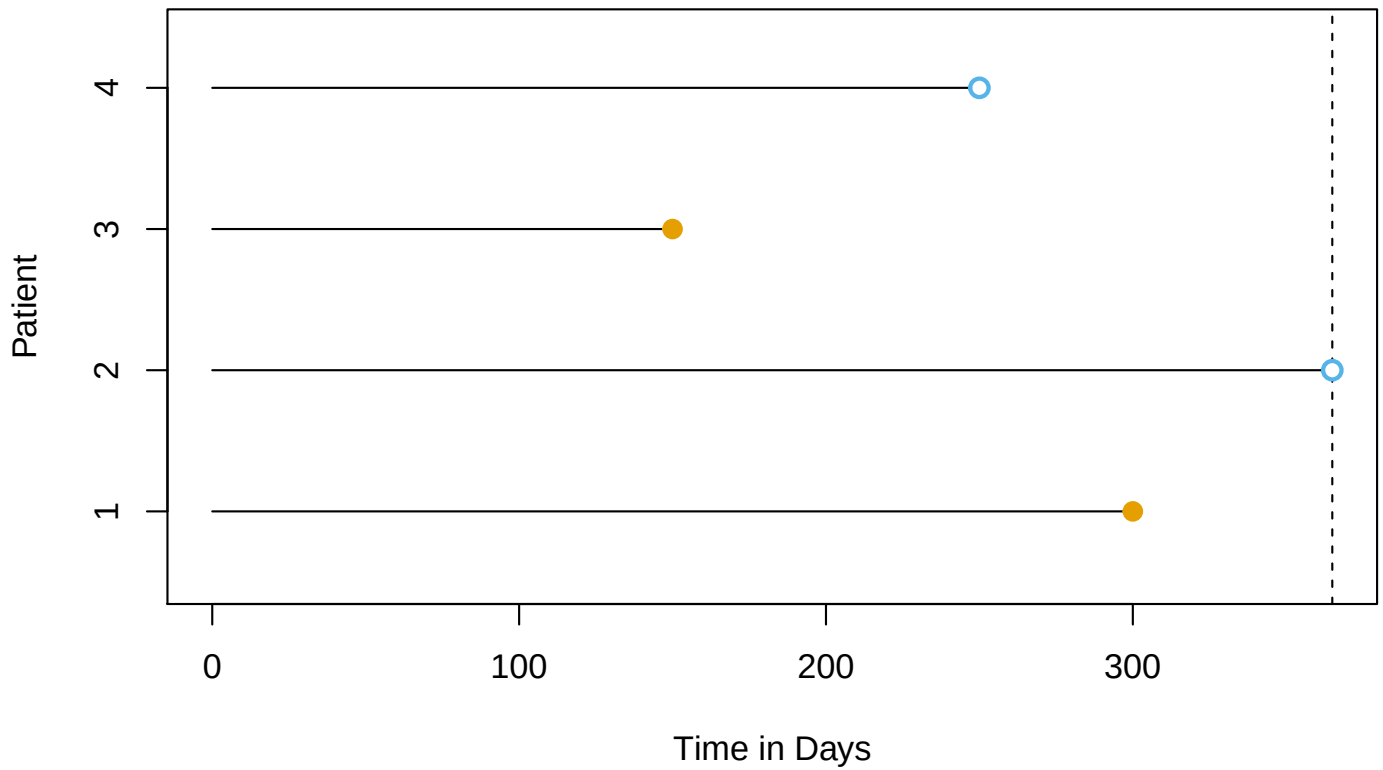
- If the event occurs before censoring (i.e. $T < C$) then we observe the true survival time T; if censoring occurs before the event ($T > C$) then we observe the censoring time. We also observe a status indicator,

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C. \end{cases}$$

- Finally, in our dataset we observe n pairs (Y, δ), which we denote as $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

Illustration

Here is an illustration of censored survival data. For patients 1 and 3, the event was observed. Patient 2 was alive when the study ended. Patient 4 dropped out of the study.



A Closer Look at Censoring

A Closer Look at Censoring

- Suppose that a number of patients drop out of a cancer study early because they are very sick.

A Closer Look at Censoring

- Suppose that a number of patients drop out of a cancer study early because they are very sick.
- An analysis that does not take into consideration the reason why the patients dropped out will likely overestimate the true average survival time.

A Closer Look at Censoring

- Suppose that a number of patients drop out of a cancer study early because they are very sick.
- An analysis that does not take into consideration the reason why the patients dropped out will likely overestimate the true average survival time.
- Similarly, suppose that males who are very sick are more likely to drop out of the study than females who are very sick. Then a comparison of male and female survival times may wrongly suggest that males survive longer than females.

A Closer Look at Censoring

- Suppose that a number of patients drop out of a cancer study early because they are very sick.
- An analysis that does not take into consideration the reason why the patients dropped out will likely overestimate the true average survival time.
- Similarly, suppose that males who are very sick are more likely to drop out of the study than females who are very sick. Then a comparison of male and female survival times may wrongly suggest that males survive longer than females.
- In general, we need to assume that, conditional on the features, the event time T is *independent* of the censoring time C. The two examples above violate the assumption of independent censoring.

The Survival Curve

The Survival Curve

- The survival function (or curve) is defined as

$$S(t) = \Pr(T > t).$$

The Survival Curve

- The survival function (or curve) is defined as

$$S(t) = \Pr(T > t).$$

- This decreasing function quantifies the probability of surviving past time t.

The Survival Curve

- The survival function (or curve) is defined as

$$S(t) = \Pr(T > t).$$

- This decreasing function quantifies the probability of surviving past time t.
- For example, suppose that a company is interested in modeling customer churn. Let T represent the time that a customer cancels a subscription to the company's service.

The Survival Curve

- The survival function (or curve) is defined as

$$S(t) = \Pr(T > t).$$

- This decreasing function quantifies the probability of surviving past time t.
- For example, suppose that a company is interested in modeling customer churn. Let T represent the time that a customer cancels a subscription to the company's service.
- Then $S(t)$ represents the probability that a customer cancels later than time t. The larger the value of $S(t)$, the less likely that the customer will cancel before time t.

Estimating the Survival Curve

Estimating the Survival Curve

- Consider the `BrainCancer` dataset, which contains the survival times for patients with primary brain tumors undergoing treatment with stereotactic radiation methods.

Estimating the Survival Curve

- Consider the `BrainCancer` dataset, which contains the survival times for patients with primary brain tumors undergoing treatment with stereotactic radiation methods.
- The predictors are `gtv` (gross tumor volume, in cubic centimeters); `sex` (male or female); `diagnosis` (meningioma, LG glioma, HG glioma, or other); `loc` (the tumor location: either infratentorial or supratentorial); `ki` (Karnofsky index); and `stereo` (stereotactic method).

Estimating the Survival Curve

- Consider the `BrainCancer` dataset, which contains the survival times for patients with primary brain tumors undergoing treatment with stereotactic radiation methods.
- The predictors are `gtv` (gross tumor volume, in cubic centimeters); `sex` (male or female); `diagnosis` (meningioma, LG glioma, HG glioma, or other); `loc` (the tumor location: either infratentorial or supratentorial); `ki` (Karnofsky index); and `stereo` (stereotactic method).
- Only 53 of the 88 patients were still alive at the end of the study.

Estimating the Survival Curve — Continued

Estimating the Survival Curve — Continued

- Suppose we'd like to estimate $S(20) = \Pr(T > 20)$, the probability that a patient survives for at least 20 months,

Estimating the Survival Curve — Continued

- Suppose we'd like to estimate $S(20) = \Pr(T > 20)$, the probability that a patient survives for at least 20 months,
- It is tempting to simply compute the proportion of patients who are known to have survived past 20 months, that is, the proportion of patients for whom $Y > 20$.

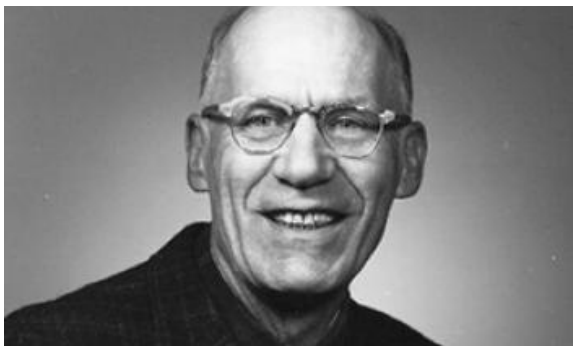
Estimating the Survival Curve — Continued

- Suppose we'd like to estimate $S(20) = \Pr(T > 20)$, the probability that a patient survives for at least 20 months,
- It is tempting to simply compute the proportion of patients who are known to have survived past 20 months, that is, the proportion of patients for whom $Y > 20$.
- This turns out to be 48/88, or approximately 55%.

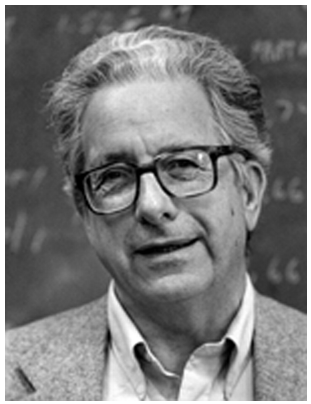
Estimating the Survival Curve — Continued

- Suppose we'd like to estimate $S(20) = \Pr(T > 20)$, the probability that a patient survives for at least 20 months,
- It is tempting to simply compute the proportion of patients who are known to have survived past 20 months, that is, the proportion of patients for whom $Y > 20$.
- This turns out to be 48/88, or approximately 55%.
- However, this does not seem quite right: 17 of the 40 patients who did not survive to 20 months were actually censored, and this analysis implicitly assumes they died before 20 months. Hence it is probably an underestimate.

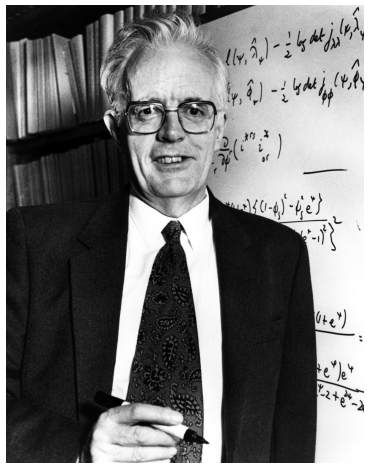
Those big names again



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

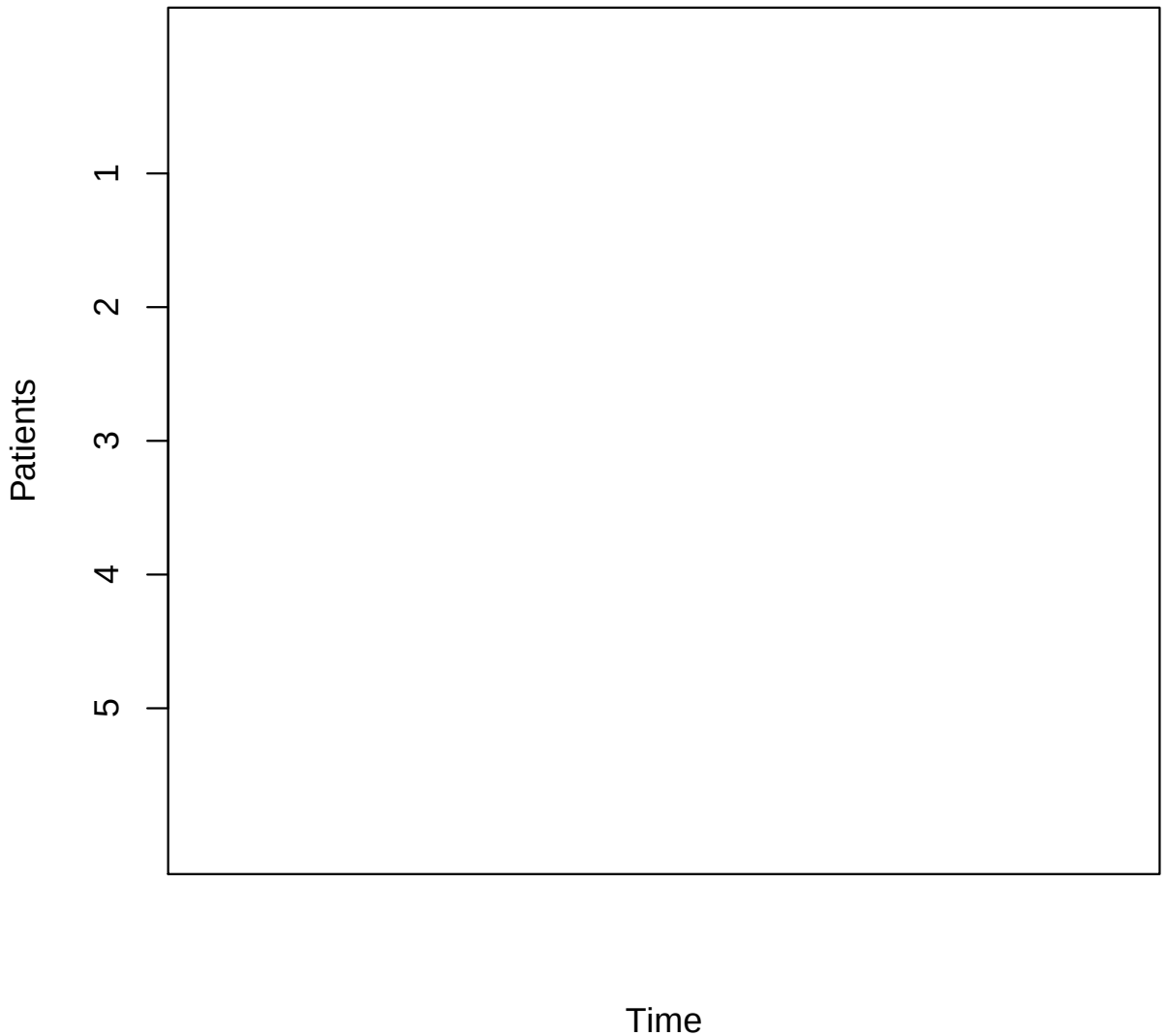
William Haenszel

(***log rank test***)

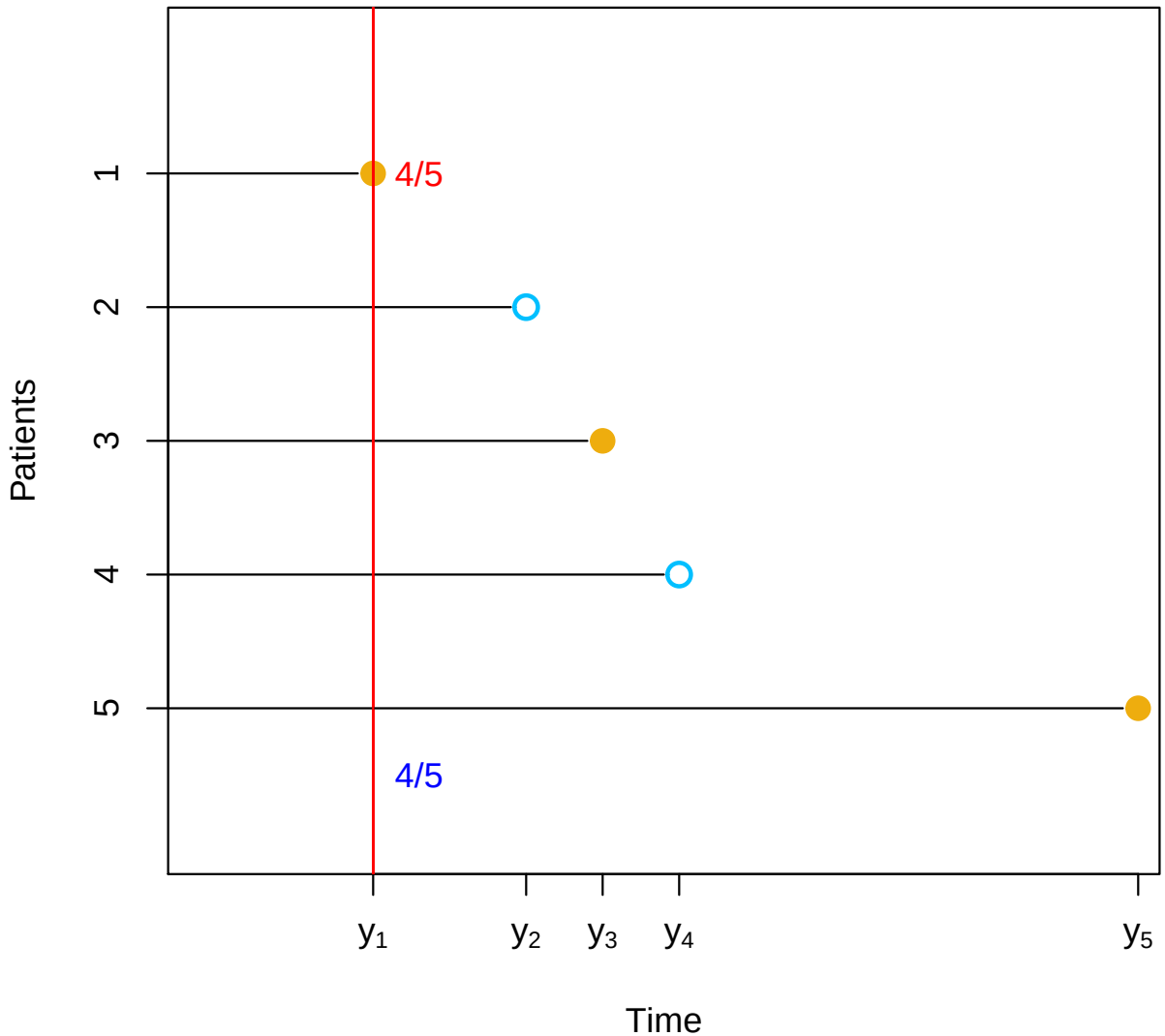
Terry Therneau

(***author of Survival package in R)***

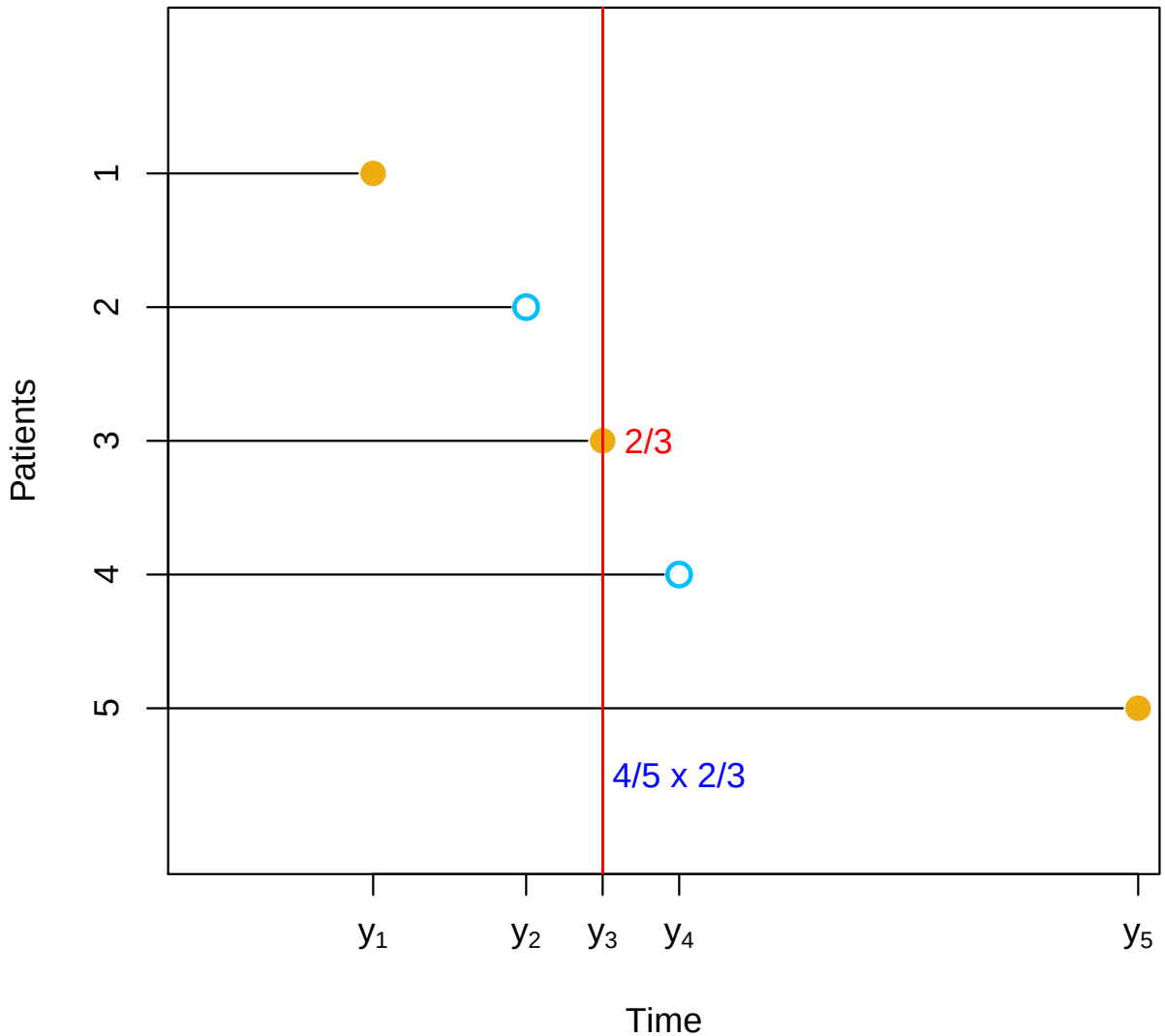
The Kaplan-Meier Estimate: Example



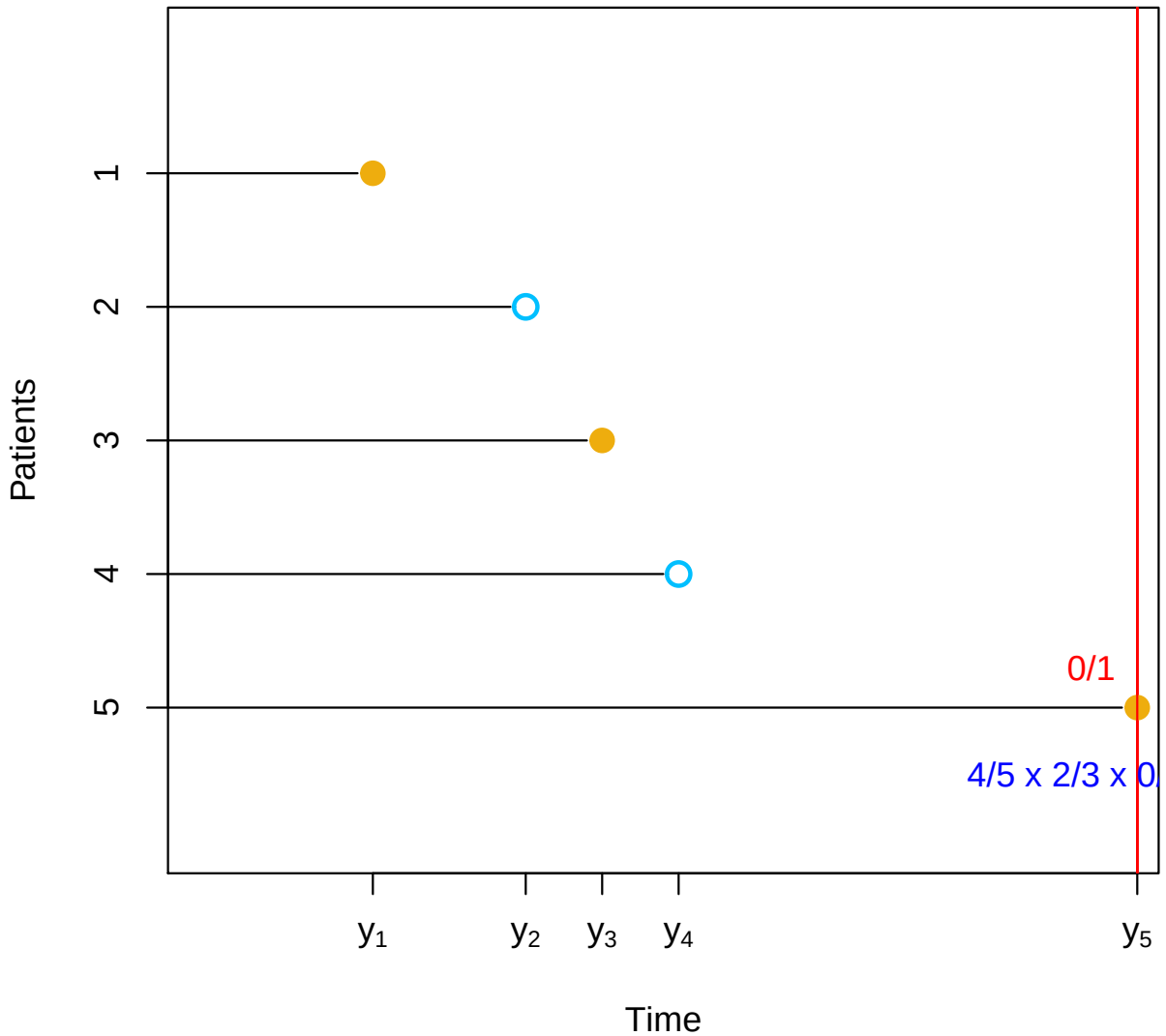
First Failure



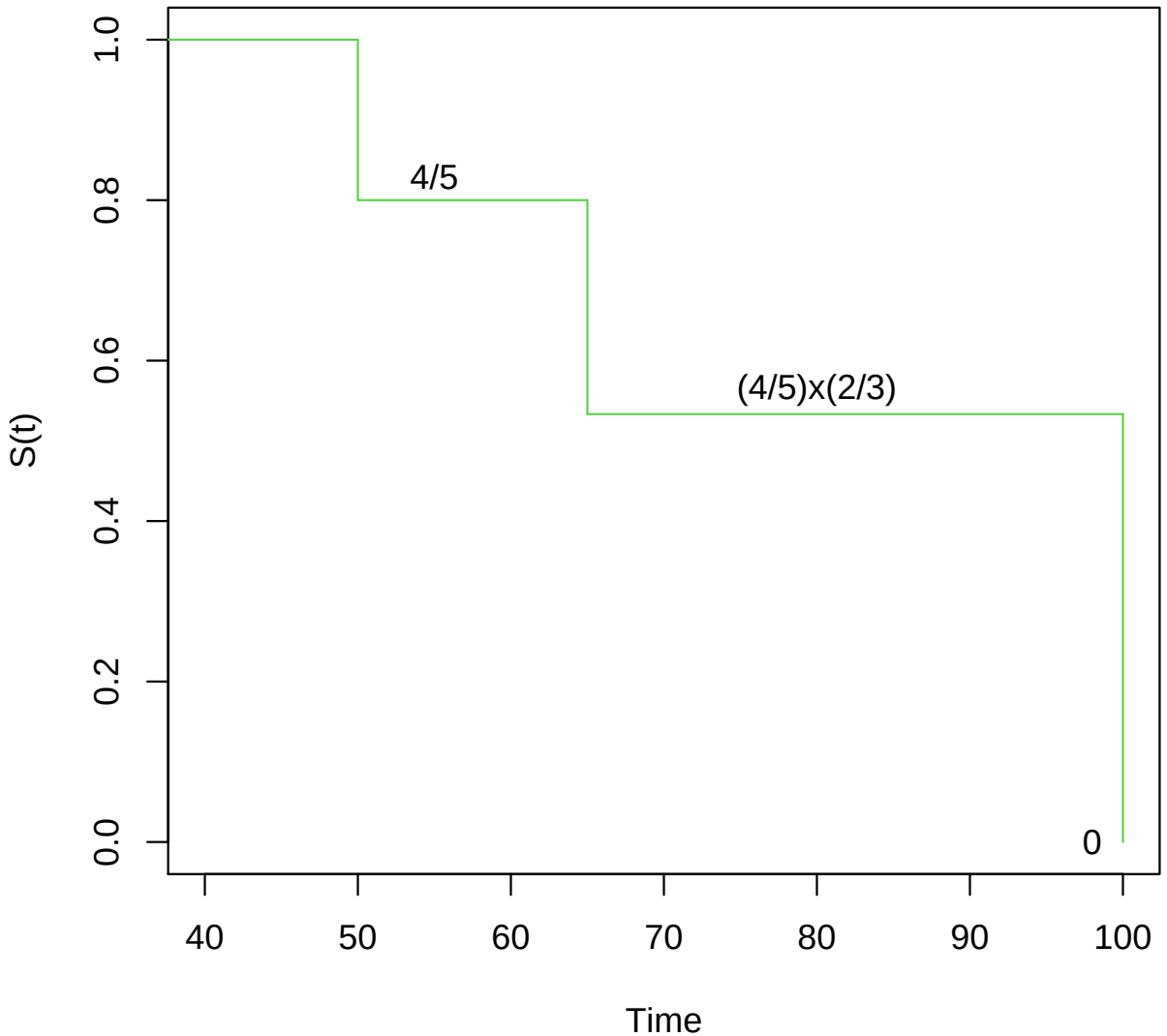
Second Failure



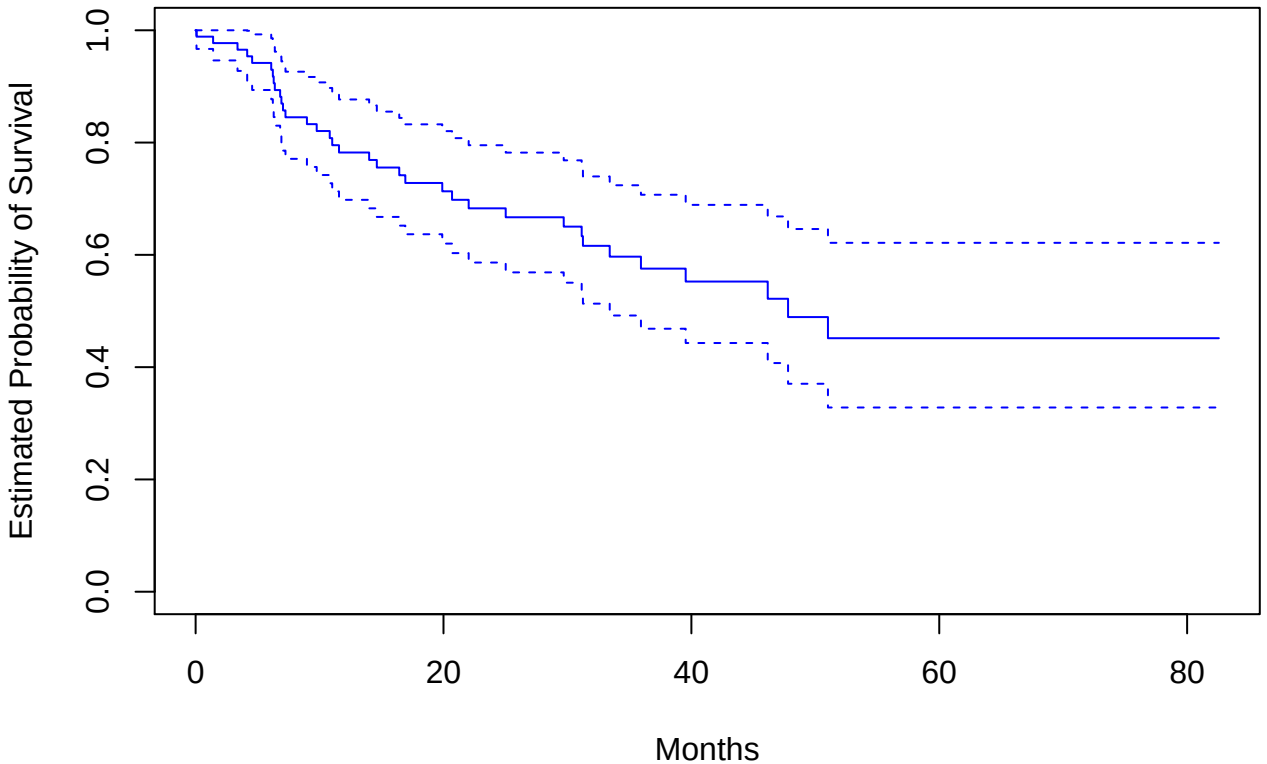
Third Failure



Resulting KM Survival Curve



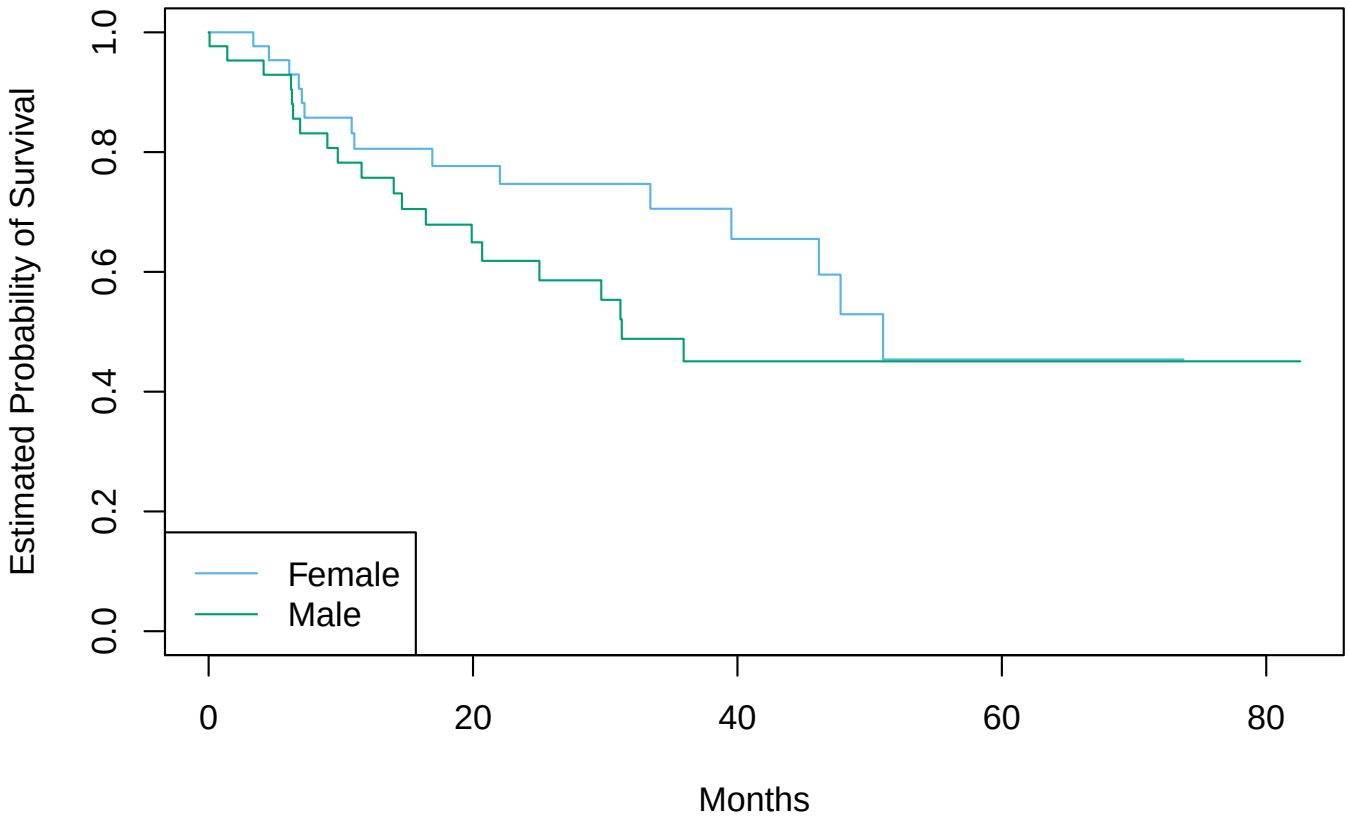
Kaplan-Meier Survival Curve for the BrainCancer Data



Each point in the solid step-like curve shows the estimated probability of surviving past the time indicated on the horizontal axis.

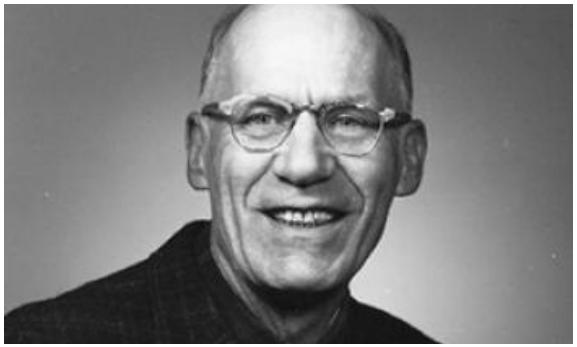
The estimated probability of survival past 20 months is 71%, which is quite a bit higher than the naive estimate of 55% presented earlier.

The Log-Rank Test

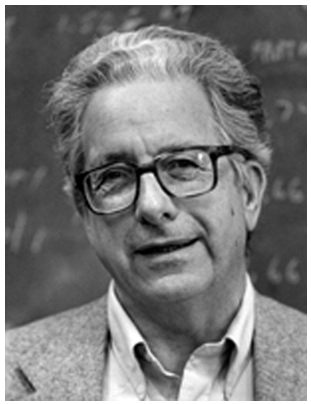


We wish to compare the survival of males to that of females. Shown are the Kaplan-Meier survival curves for the two groups.

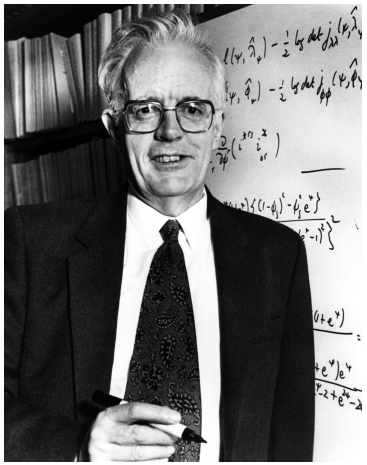
Those big names again



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel **William Haenszel**

(***log rank test***)

Terry Therneau

(author of Survival package in R)

The Log-Rank Test — Continued

The Log-Rank Test — Continued

- Females seem to fare a little better up to about 50 months, but then the two curves both level off to about 50%. How can we carry out a formal test of equality of the two survival curves?

The Log-Rank Test — Continued

- Females seem to fare a little better up to about 50 months, but then the two curves both level off to about 50%. How can we carry out a formal test of equality of the two survival curves?
- At first glance, a two-sample t-test seems like an obvious choice: but the presence of censoring again creates a complication.

The Log-Rank Test — Continued

- Females seem to fare a little better up to about 50 months, but then the two curves both level off to about 50%. How can we carry out a formal test of equality of the two survival curves?
- At first glance, a two-sample t-test seems like an obvious choice: but the presence of censoring again creates a complication.
- To overcome this challenge, we will conduct a log-rank test.

The Log-Rank Test — Continued

The Log-Rank Test — Continued

- Recall that $d_1 < d_2 < \cdots < d_K$ are the unique death times among the non-censored patients, r_k is the number of patients at risk at time d_k, and q_k is the number of patients who died at time d_k.

The Log-Rank Test — Continued

- Recall that $d_1 < d_2 < \cdots < d_K$ are the unique death times among the non-censored patients, r_k is the number of patients at risk at time d_k, and q_k is the number of patients who died at time d_k.
- We further define r_{1k} and r_{2k} to be the number of patients in groups 1 and 2, respectively, who are at risk at time d_k.

The Log-Rank Test — Continued

- Recall that $d_1 < d_2 < \cdots < d_K$ are the unique death times among the non-censored patients, r_k is the number of patients at risk at time d_k, and q_k is the number of patients who died at time d_k.
- We further define r_{1k} and r_{2k} to be the number of patients in groups 1 and 2, respectively, who are at risk at time d_k.
- Similarly, we define q_{1k} and q_{2k} to be the number of patients in groups 1 and 2, respectively, who died at time d_k. Note that $r_{1k} + r_{2k} = r_k$ and $q_{1k} + q_{2k} = q_k$.

Details of the Test Statistic

	Group 1	Group 2	Total
Died	q_{1k}	q_{2k}	q_k
Survived	$r_{1k} - q_{1k}$	$r_{2k} - q_{2k}$	$r_k - q_k$
Total	r_{1k}	r_{2k}	r_k

At each death time d_k, we construct a 2×2 table of counts of the form shown above.

Note that if the death times are unique (i.e. no two individuals die at the same time), then one of q_{1k} and q_{2k} equals one, and the other equals zero.

Log Rank Test: the Main Idea

Log Rank Test: the Main Idea

- To test $H_0 : \mathrm{E}(X) = 0$ for some random variable X, one approach is to construct a test statistic of the form

$$W = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}},$$

where $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ are the expectation and variance, respectively, of X under H_0.

Log Rank Test: the Main Idea

- To test $H_0 : \mathrm{E}(X) = 0$ for some random variable X, one approach is to construct a test statistic of the form

$$W = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}},$$

  where $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ are the expectation and variance, respectively, of X under H_0.
- In order to construct the log-rank test statistic, we compute a quantity that takes exactly the form above, with $X = \sum_{k=1}^{K} q_{1k}$, where q_{1k} is given in the top left of the table above.

The Final Result

The resulting formula for the log-rank test statistic is

$$W = \frac{\sum_{k=1}^{K} \left(q_{1k} - \mathrm{E}(q_{1k})\right)}{\sqrt{\sum_{k=1}^{K} \mathrm{Var}\left(q_{1k}\right)}} = \frac{\sum_{k=1}^{K} \left(q_{1k} - \frac{q_k}{r_k} r_{1k}\right)}{\sqrt{\sum_{k=1}^{K} \frac{q_k (r_{1k}/r_k)(1-r_{1k}/r_k)(r_k-q_k)}{r_k-1}}}.$$

When the sample size is large, the log-rank test statistic W has approximately a standard normal distribution.

This can be used to compute a p-value for the null hypothesis that there is no difference between the survival curves in the two groups.

Application to the Brain Cancer Dataset

Application to the Brain Cancer Dataset

- Comparing the survival times of females and males on the `BrainCancer` data gives a log-rank test statistic of $W = 1.2$, which corresponds to a two-sided p-value of 0.2.

Application to the Brain Cancer Dataset

- Comparing the survival times of females and males on the `BrainCancer` data gives a log-rank test statistic of $W = 1.2$, which corresponds to a two-sided p-value of 0.2.
- Thus, we cannot reject the null hypothesis of no difference in survival curves between females and males.

Application to the Brain Cancer Dataset

- Comparing the survival times of females and males on the `BrainCancer` data gives a log-rank test statistic of $W = 1.2$, which corresponds to a two-sided p-value of 0.2.
- Thus, we cannot reject the null hypothesis of no difference in survival curves between females and males.
- The log-rank test is closely related to Cox's proportional hazards model, which we discuss next.

Regression Models with a Survival Response

Regression Models with a Survival Response

- We now consider the task of fitting a regression model to survival data.

Regression Models with a Survival Response

- We now consider the task of fitting a regression model to survival data.
- We wish to predict the true survival time T.
  Since the observed quantity $Y = \min(T, C)$ is positive and may have a long right tail, we might be tempted to fit a linear regression of $\log(Y)$ on X. But *censoring again creates a problem*.

Regression Models with a Survival Response

- We now consider the task of fitting a regression model to survival data.
- We wish to predict the true survival time T.
  Since the observed quantity $Y = \min(T, C)$ is positive and may have a long right tail, we might be tempted to fit a linear regression of $\log(Y)$ on X. But *censoring again creates a problem*.
- To overcome this difficulty, we instead make use of a sequential construction, similar to the idea used for the Kaplan-Meier survival curve.

The Hazard Function

The *hazard function* or *hazard rate* — also known as the *force of mortality* — is formally defined as

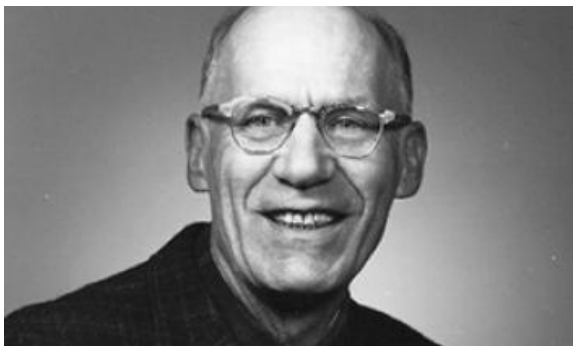
$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t | T > t)}{\Delta t},$$

where T is the (true) survival time.

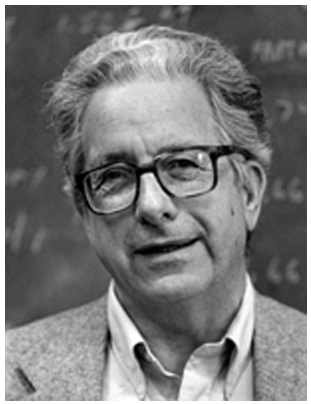
It is the death rate in the instant after time t, given survival up to that time.

The hazard function is the basis for the *Proportional Hazards Model*, discussed next.

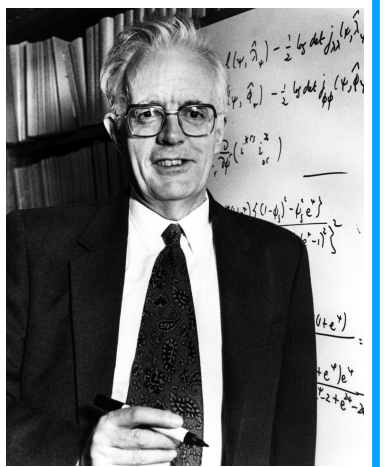
Bringing in the covariates: those big names again



Edward Kaplan



Paul Meier

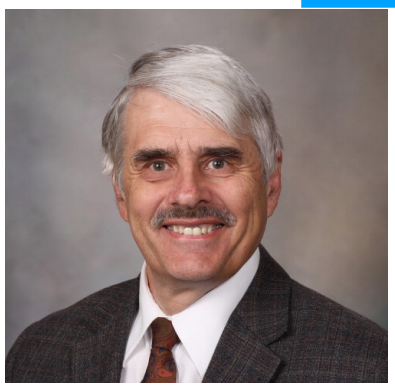


David Cox

Nathan Mantel

William Haenszel

(log rank test)



Terry Therneau

(author of Survival package in R)

The Proportional Hazards Model

The Proportional Hazards Model

- The proportional hazards assumption states that

$$h(t|x_i) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right),$$

where $h_0(t) \geq 0$ is an unspecified function, known as the *baseline hazard*. It is the hazard function for an individual with features $x_{i1} = \cdots = x_{ip} = 0$.

The Proportional Hazards Model

- The proportional hazards assumption states that

$$h(t|x_i) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right),$$

  where $h_0(t) \geq 0$ is an unspecified function, known as the *baseline hazard*. It is the hazard function for an individual with features $x_{i1} = \cdots = x_{ip} = 0$.

- The name *proportional hazards* arises from the fact that the hazard function for an individual with feature vector x_i is some unknown function $h_0(t)$ times the factor $\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$. The quantity $\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$ is called the *relative risk* for the feature vector $x_i = (x_{i1}, \ldots, x_{ip})$, relative to that for the feature vector $x_i = (0, \ldots, 0)$.

Proportional Hazards Model— Continued

Proportional Hazards Model— Continued

- What does it mean that the baseline hazard function $h_0(t)$ is unspecified?

Proportional Hazards Model— Continued

- What does it mean that the baseline hazard function $h_0(t)$ is unspecified?
- Basically, we make *no assumptions about its functional form*. We allow the instantaneous probability of death at time t, given that one has survived at least until time t, to take any form.

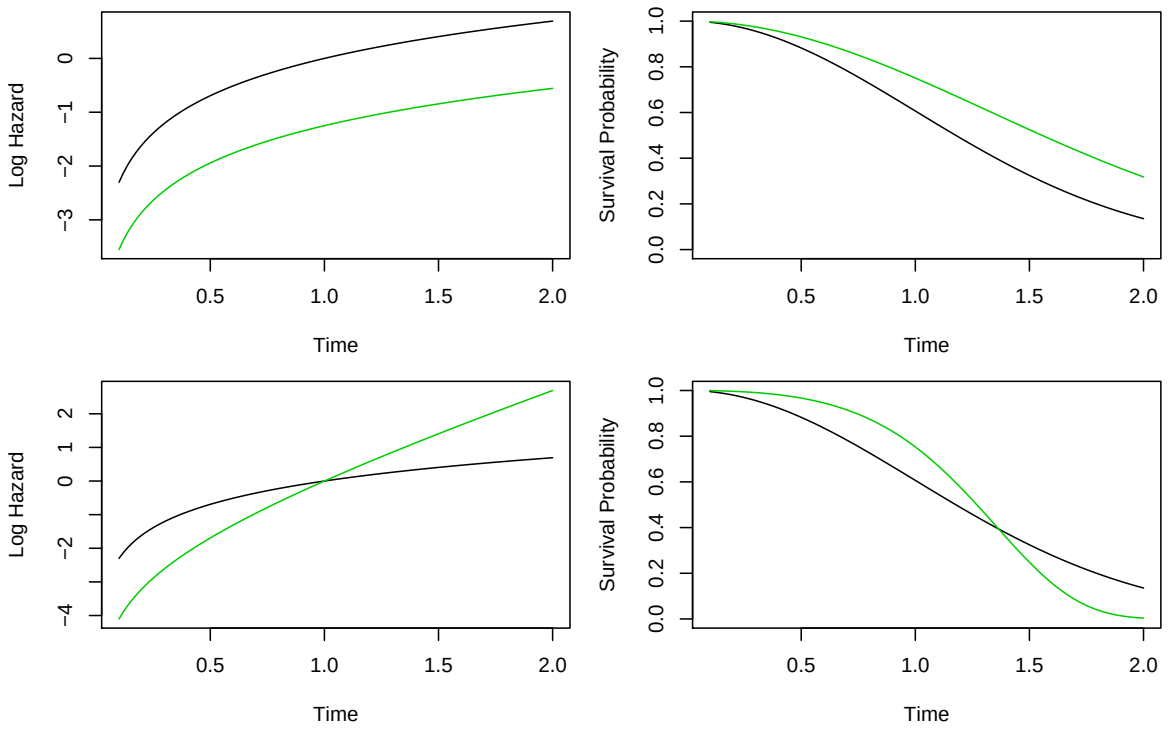
Proportional Hazards Model— Continued

- What does it mean that the baseline hazard function $h_0(t)$ is unspecified?
- Basically, we make *no assumptions about its functional form*. We allow the instantaneous probability of death at time t, given that one has survived at least until time t, to take any form.
- This means that the hazard function is very flexible and can model a wide range of relationships between the covariates and survival time.

Proportional Hazards Model— Continued

- What does it mean that the baseline hazard function $h_0(t)$ is unspecified?
- Basically, we make *no assumptions about its functional form*. We allow the instantaneous probability of death at time t, given that one has survived at least until time t, to take any form.
- This means that the hazard function is very flexible and can model a wide range of relationships between the covariates and survival time.
- Our only assumption is that a one-unit increase in x_{ij} corresponds to an increase in $h(t|x_i)$ by a factor of $\exp(\beta_j)$.

An Example



Here is an example with $p = 1$ and a binary covariate $x_i \in \{0, 1\}$.

Top row: the log hazard and the survival function under the model are shown (green for $x_i = 0$ and black for $x_i = 1$). Because of the proportional hazards assumption, the log hazard functions differ by a constant, and the survival functions do not cross.

Bottom row: the proportional hazards assumption does not hold.

Partial Likelihood

Partial Likelihood

- Because the form of the baseline hazard is unknown, we cannot simply plug $h(t|x_i)$ into the likelihood and then estimate $\beta = (\beta_1, \ldots, \beta_p)^T$ by maximum likelihood.

Partial Likelihood

- Because the form of the baseline hazard is unknown, we cannot simply plug $h(t|x_i)$ into the likelihood and then estimate $\beta = (\beta_1, \ldots, \beta_p)^T$ by maximum likelihood.
- The magic of Cox's proportional hazards model lies in the fact that it is in fact possible to estimate β *without having to specify the form of* $h_0(t)$.

Partial Likelihood

- Because the form of the baseline hazard is unknown, we cannot simply plug $h(t|x_i)$ into the likelihood and then estimate $\beta = (\beta_1, \ldots, \beta_p)^T$ by maximum likelihood.
- The magic of Cox's proportional hazards model lies in the fact that it is in fact possible to estimate β *without having to specify the form of* $h_0(t)$.
- To accomplish this, we make use of the same "sequential in time" logic that we used to derive the Kaplan-Meier survival curve and the log-rank test. Then the total hazard at failure time y_i for the at-risk observations is

$$\sum_{i':y_{i'} \geq y_i} h_0(y_i) \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right).$$

Partial Likelihood — Continued

Partial Likelihood — Continued

- Therefore, the probability that the ith observation is the one to fail at time y_i (as opposed to one of the other observations in the risk set) is

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

Partial Likelihood — Continued

- Therefore, the probability that the ith observation is the one to fail at time y_i (as opposed to one of the other observations in the risk set) is

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

- Notice that the unspecified baseline hazard function $h_0(y_i)$ cancels out of the numerator and denominator!

Partial Likelihood — Continued

Partial Likelihood — Continued

- The partial likelihood is simply the product of these probabilities over all of the uncensored observations,

$$PL(\beta) = \prod_{i:\delta_i=1} \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

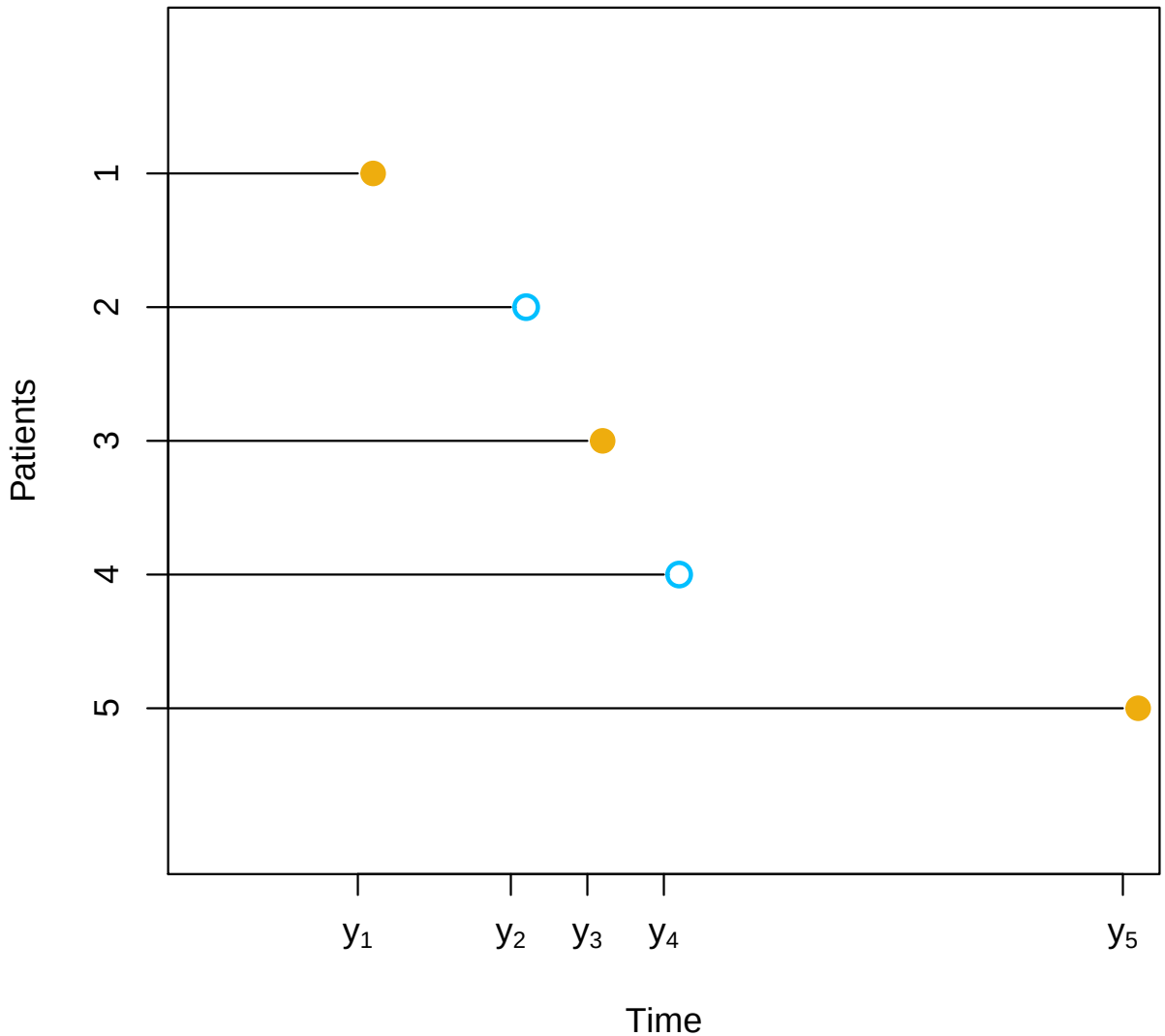
Partial Likelihood — Continued

- The partial likelihood is simply the product of these probabilities over all of the uncensored observations,

$$PL(\beta) = \prod_{i:\delta_i=1} \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

- Critically, the partial likelihood is valid regardless of the true value of $h_0(t)$, making the model very flexible and robust.

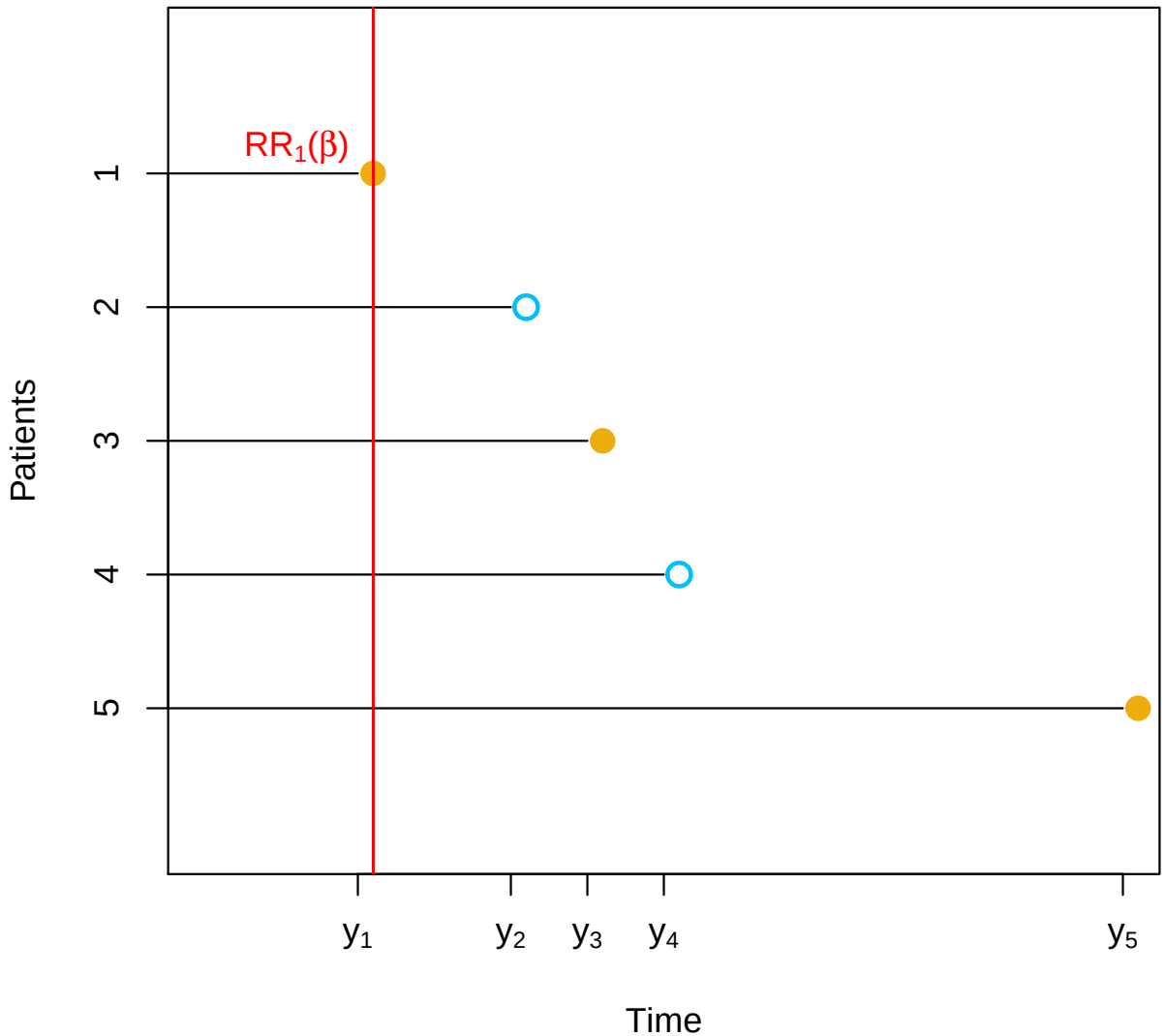
The Partial Likelihood: Example



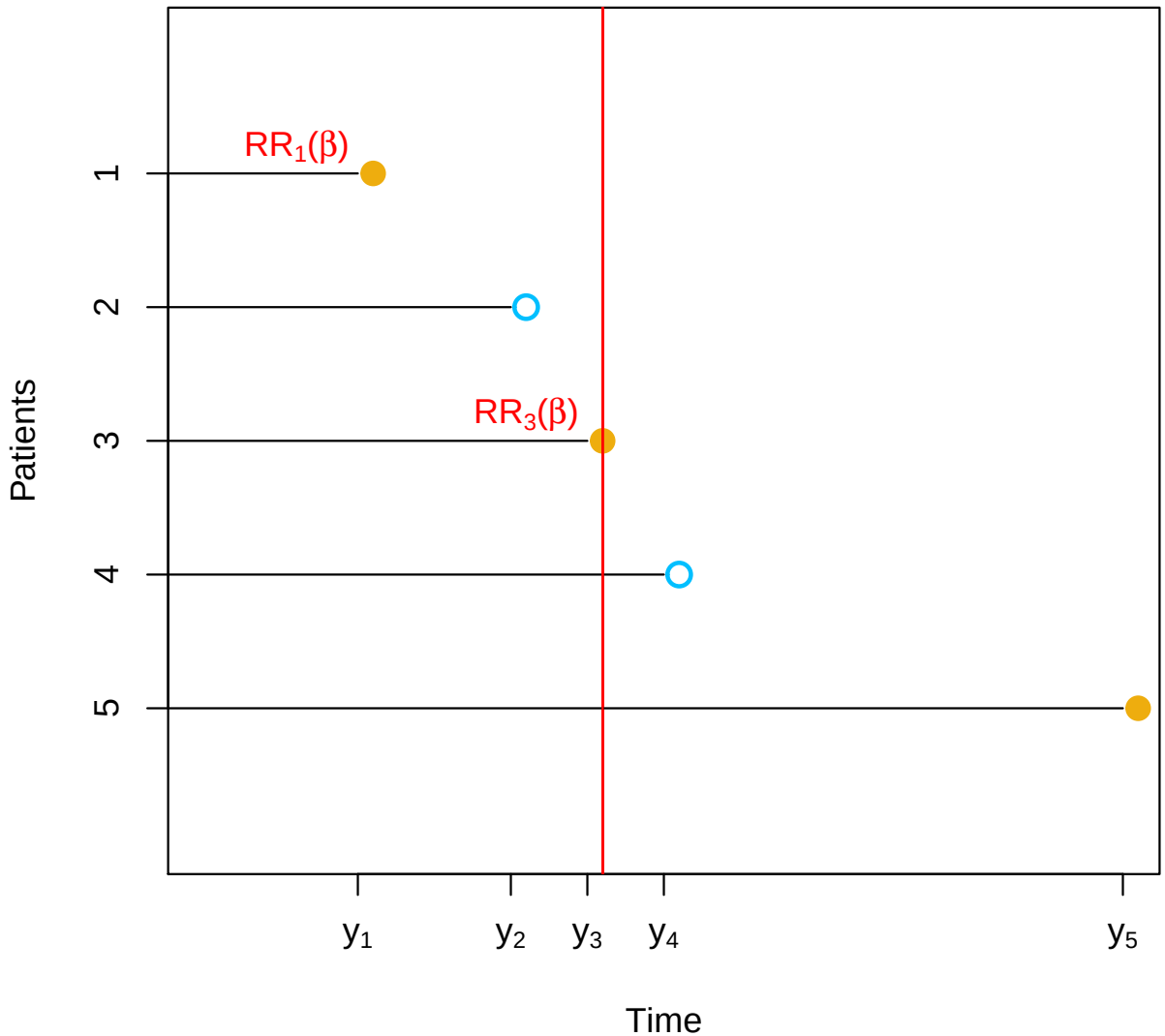
Relative Risk Functions at each Failure Time

$$
\begin{aligned}
RR_1(\beta) &= \frac{\exp\left(\sum_{j=1}^{p} x_{1j}\beta_j\right)}{\sum_{i':y_{i'}\geq y_1}\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} \\
RR_3(\beta) &= \frac{\exp\left(\sum_{j=1}^{p} x_{3j}\beta_j\right)}{\sum_{i':y_{i'}\geq y_3}\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} \\
RR_5(\beta) &= \frac{\exp\left(\sum_{j=1}^{p} x_{5j}\beta_j\right)}{\sum_{i':y_{i'}\geq y_5}\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}
\end{aligned}
$$

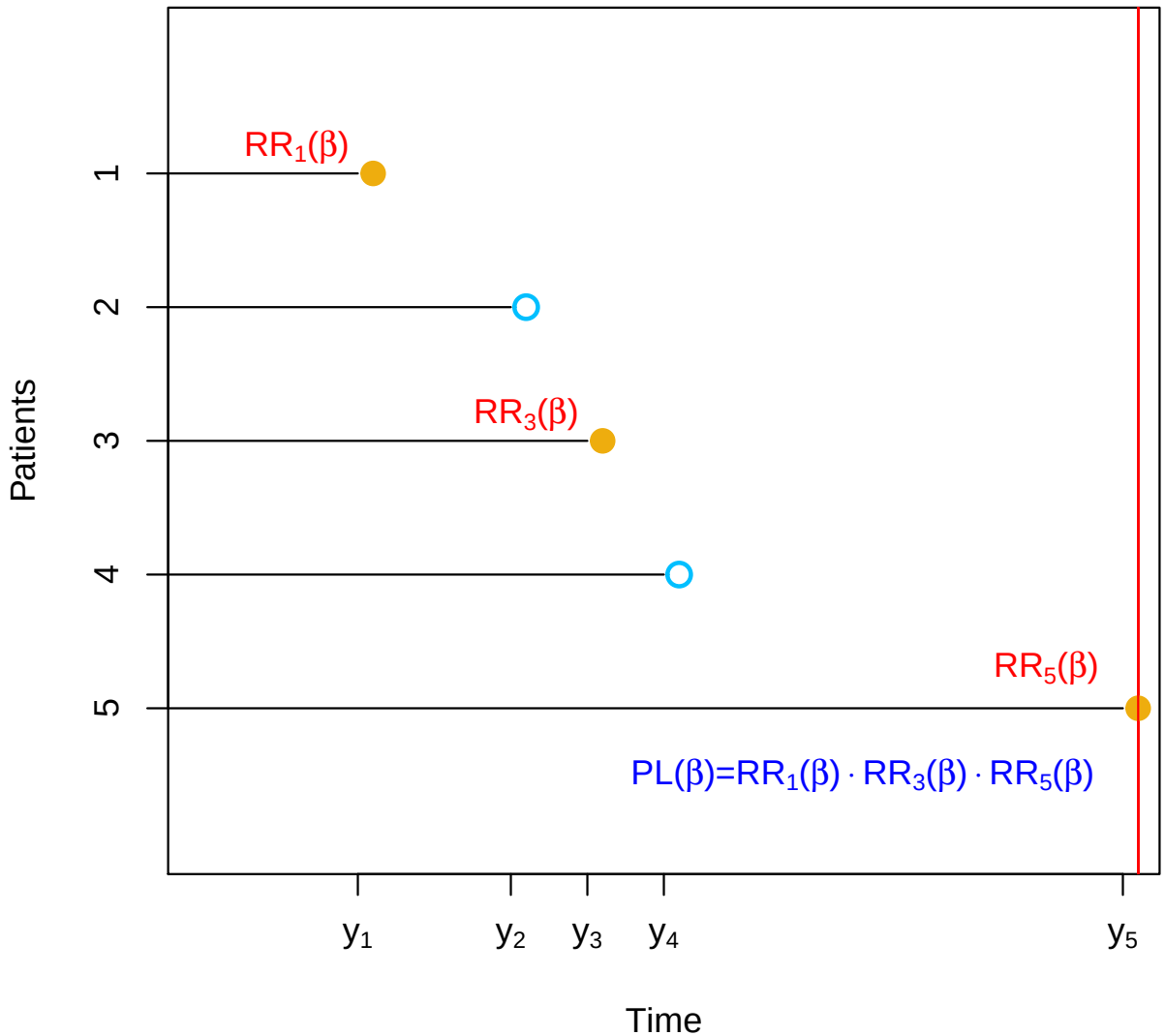
First Failure



Second Failure



Third Failure



Partial Likelihood — Computation

Partial Likelihood — Computation

- To estimate β, we simply maximize the partial likelihood with respect to β. As is the case for logistic regression, no closed-form solution is available, and so iterative algorithms are required.

Partial Likelihood — Computation

- To estimate β, we simply maximize the partial likelihood with respect to β. As is the case for logistic regression, no closed-form solution is available, and so iterative algorithms are required.
- In addition to estimating β, we can also obtain other model outputs, like those in least squares regression and logistic regression.

Partial Likelihood — Computation

- To estimate β, we simply maximize the partial likelihood with respect to β. As is the case for logistic regression, no closed-form solution is available, and so iterative algorithms are required.
- In addition to estimating β, we can also obtain other model outputs, like those in least squares regression and logistic regression.
- For example, we can obtain *p-values* corresponding to particular null hypotheses (e.g. $H_0 : \beta_j = 0$), as well as *estimated standard errors* and *confidence intervals* associated with the coefficients.

Connection with the Log-Rank Test

Connection with the Log-Rank Test

- Suppose that we have just a single predictor ($p = 1$) with $x_i \in \{0, 1\}$. To test whether there is a difference between the survival times of the observations in the two groups, we can consider taking two possible approaches:

Connection with the Log-Rank Test

- Suppose that we have just a single predictor ($p = 1$) with $x_i \in \{0, 1\}$. To test whether there is a difference between the survival times of the observations in the two groups, we can consider taking two possible approaches:
    1. Fit a Cox proportional hazards model, and test the null hypothesis $H_0 : \beta = 0$. (Since $p = 1$, β is a scalar.)

Connection with the Log-Rank Test

- Suppose that we have just a single predictor ($p = 1$) with $x_i \in \{0, 1\}$. To test whether there is a difference between the survival times of the observations in the two groups, we can consider taking two possible approaches:
  1. Fit a Cox proportional hazards model, and test the null hypothesis $H_0 : \beta = 0$. (Since $p = 1$, β is a scalar.)
  2. Perform a log-rank test to compare the two groups.

Connection with the Log-Rank Test

- Suppose that we have just a single predictor ($p = 1$) with $x_i \in \{0, 1\}$. To test whether there is a difference between the survival times of the observations in the two groups, we can consider taking two possible approaches:
  1. Fit a Cox proportional hazards model, and test the null hypothesis $H_0 : \beta = 0$. (Since $p = 1$, β is a scalar.)
  2. Perform a log-rank test to compare the two groups.
- Now when taking approach #1, there are a number of possible ways to test H_0. One way is known as a *score test*.

Connection with the Log-Rank Test

- Suppose that we have just a single predictor ($p = 1$) with $x_i \in \{0, 1\}$. To test whether there is a difference between the survival times of the observations in the two groups, we can consider taking two possible approaches:
  1. Fit a Cox proportional hazards model, and test the null hypothesis $H_0 : \beta = 0$. (Since $p = 1$, β is a scalar.)
  2. Perform a log-rank test to compare the two groups.
- Now when taking approach #1, there are a number of possible ways to test H_0. One way is known as a *score test*.
- It turns out that in the case of a single binary covariate, the score test for $H_0 : \beta = 0$ in Cox's proportional hazards model is *exactly equal to the log-rank test*.

The Proportional Hazards Model— Additional Details

The discussion of the proportional hazards model glossed over a few subtleties:

The Proportional Hazards Model— Additional Details

The discussion of the proportional hazards model glossed over a few subtleties:

- There is *no intercept* in the proportional hazards model because an intercept can be absorbed into the baseline hazard $h_0(t)$.

The Proportional Hazards Model— Additional Details

The discussion of the proportional hazards model glossed over a few subtleties:

- There is *no intercept* in the proportional hazards model because an intercept can be absorbed into the baseline hazard $h_0(t)$.
- We have assumed that there are *no tied failure times*. In the case of ties, the exact form of the partial likelihood is more complicated, and a number of computational approximations must be used.

The Proportional Hazards Model— Additional Details

The discussion of the proportional hazards model glossed over a few subtleties:

- There is *no intercept* in the proportional hazards model because an intercept can be absorbed into the baseline hazard $h_0(t)$.
- We have assumed that there are *no tied failure times*. In the case of ties, the exact form of the partial likelihood is more complicated, and a number of computational approximations must be used.
- The *partial* likelihood gets its name because it is not exactly a likelihood. However, it is a very good approximation.

The Proportional Hazards Model— Additional Details

The discussion of the proportional hazards model glossed over a few subtleties:

- There is *no intercept* in the proportional hazards model because an intercept can be absorbed into the baseline hazard $h_0(t)$.
- We have assumed that there are *no tied failure times*. In the case of ties, the exact form of the partial likelihood is more complicated, and a number of computational approximations must be used.
- The *partial* likelihood gets its name because it is not exactly a likelihood. However, it is a very good approximation.
- We have focused only on estimation of the coefficients β. However, we may also wish to estimate the baseline hazard $h_0(t)$, for instance so that we can estimate the survival curve $S(t|x)$. These are implemented in the `survival` package in `R`.

Example: Brain Cancer Data

	Coefficient	Std. error	z-statistic	p-value
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	$<$0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

Example: Brain Cancer Data

	Coefficient	Std. error	z-statistic	p-value
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	$<$0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

- This table shows the result of fitting the proportional hazards model to the `BrainCancer` data.

Example: Brain Cancer Data

	Coefficient	Std. error	z-statistic	p-value
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	<0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

- This table shows the result of fitting the proportional hazards model to the `BrainCancer` data.
- We see for example that each one-unit increase in the Karnofsky index corresponds to a multiplier of $\exp(-0.05) = 0.95$ in the instantaneous chance of dying.

Example: Brain Cancer Data

	Coefficient	Std. error	z-statistic	p-value
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	<0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

- This table shows the result of fitting the proportional hazards model to the `BrainCancer` data.
- We see for example that each one-unit increase in the Karnofsky index corresponds to a multiplier of $\exp(-0.05) = 0.95$ in the instantaneous chance of dying.
- In other words, the higher the Karnofsky index, the lower the chance of dying at any given point in time. This effect is highly significant, with a p-value of 0.0027.

Example: Publication Data

Next, we consider the `Publication` dataset involving the time to publication of journal papers reporting the results of clinical trials funded by the National Heart, Lung, and Blood Institute.

Example: Publication Data

Next, we consider the `Publication` dataset involving the time to publication of journal papers reporting the results of clinical trials funded by the National Heart, Lung, and Blood Institute.

- For 244 trials, the time in months until publication is recorded. Of the 244 trials, only 156 were published during the study period; the remaining studies were censored.

Example: Publication Data

Next, we consider the `Publication` dataset involving the time to publication of journal papers reporting the results of clinical trials funded by the National Heart, Lung, and Blood Institute.

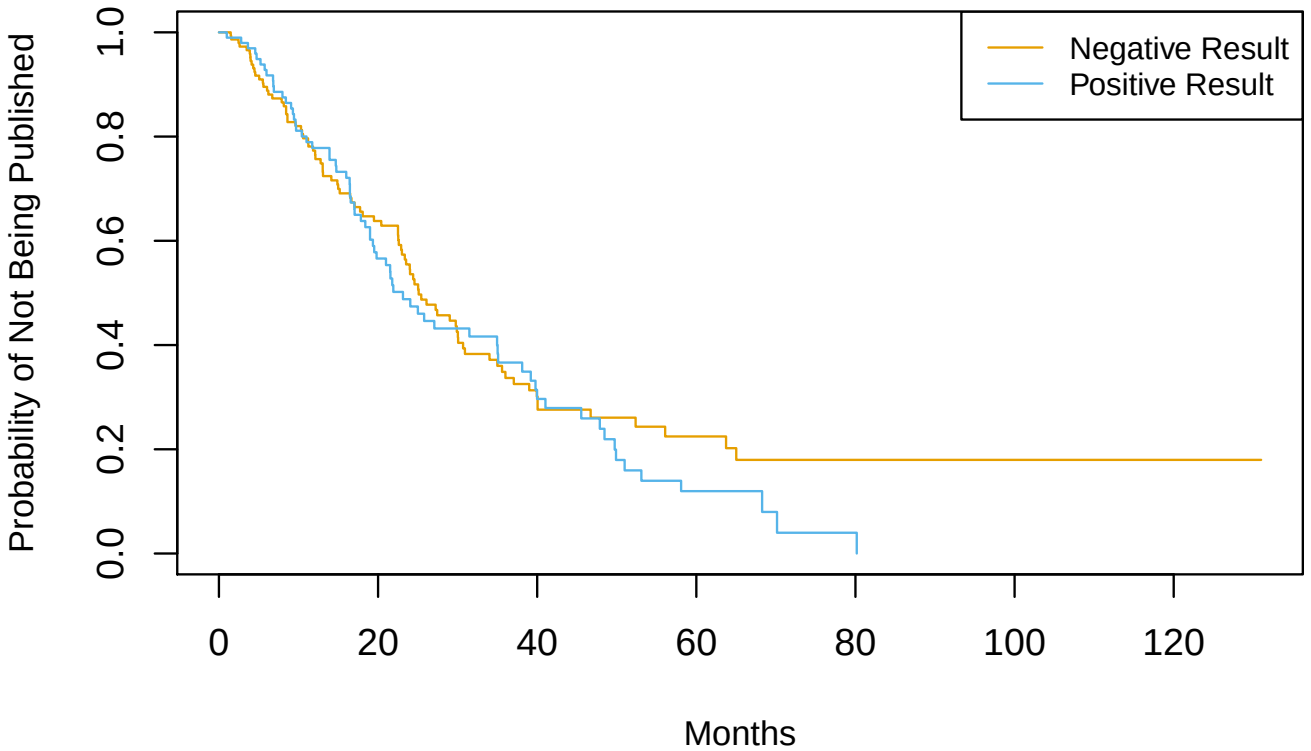
- For 244 trials, the time in months until publication is recorded. Of the 244 trials, only 156 were published during the study period; the remaining studies were censored.
- The covariates include whether the trial focused on a clinical endpoint (`clinend`), whether the trial involved multiple centers (`multi`), the funding mechanism within the National Institutes of Health (`mech`), trial sample size (`sampsize`), budget (`budget`), impact (`impact`, related to the number of citations), and whether the trial produced a positive (significant) result (`posres`).

Example: Publication Data

Next, we consider the `Publication` dataset involving the time to publication of journal papers reporting the results of clinical trials funded by the National Heart, Lung, and Blood Institute.

- For 244 trials, the time in months until publication is recorded. Of the 244 trials, only 156 were published during the study period; the remaining studies were censored.
- The covariates include whether the trial focused on a clinical endpoint (`clinend`), whether the trial involved multiple centers (`multi`), the funding mechanism within the National Institutes of Health (`mech`), trial sample size (`sampsize`), budget (`budget`), impact (`impact`, related to the number of citations), and whether the trial produced a positive (significant) result (`posres`).
- The last covariate is particularly interesting, as a number of studies have suggested that positive trials have a higher publication rate.

Publication Data — Continued



- The figure above shows the Kaplan-Meier curves for the time until publication, stratified by whether or not the study produced a positive result.
- We see slight evidence that time until publication is lower for studies with a positive result. However, the log-rank test yields a very unimpressive p-value of 0.36.

Publication Data: Multivariate Analysis

	Coefficient	Std. error	z-statistic	p-value
`posres[Yes]`	0.55	0.18	3.02	0.00
`multi[Yes]`	0.15	0.31	0.47	0.64
`clinend[Yes]`	0.51	0.27	1.89	0.06
`mech[K01]`	1.05	1.06	1.00	0.32
many mech lines omitted				
`sampsize`	0.00	0.00	0.19	0.85
`budget`	0.00	0.00	1.67	0.09
`impact`	0.06	0.01	8.23	0.00

- The results of fitting Cox's proportional hazards model using all of the available features are shown above.
- We find that the chance of publication of a study with a positive result is $e^{0.55} = 1.74$ times higher than that of a negative result at any point in time, holding all other covariates fixed.
- The very small p-value associated with `posres` indicates that this result is highly significant.

Digging Deeper

In order to gain more insight into this result, on the next slide we display estimates of the survival curves associated with positive and negative results, adjusting for the other predictors.

Digging Deeper

In order to gain more insight into this result, on the next slide we display estimates of the survival curves associated with positive and negative results, adjusting for the other predictors.

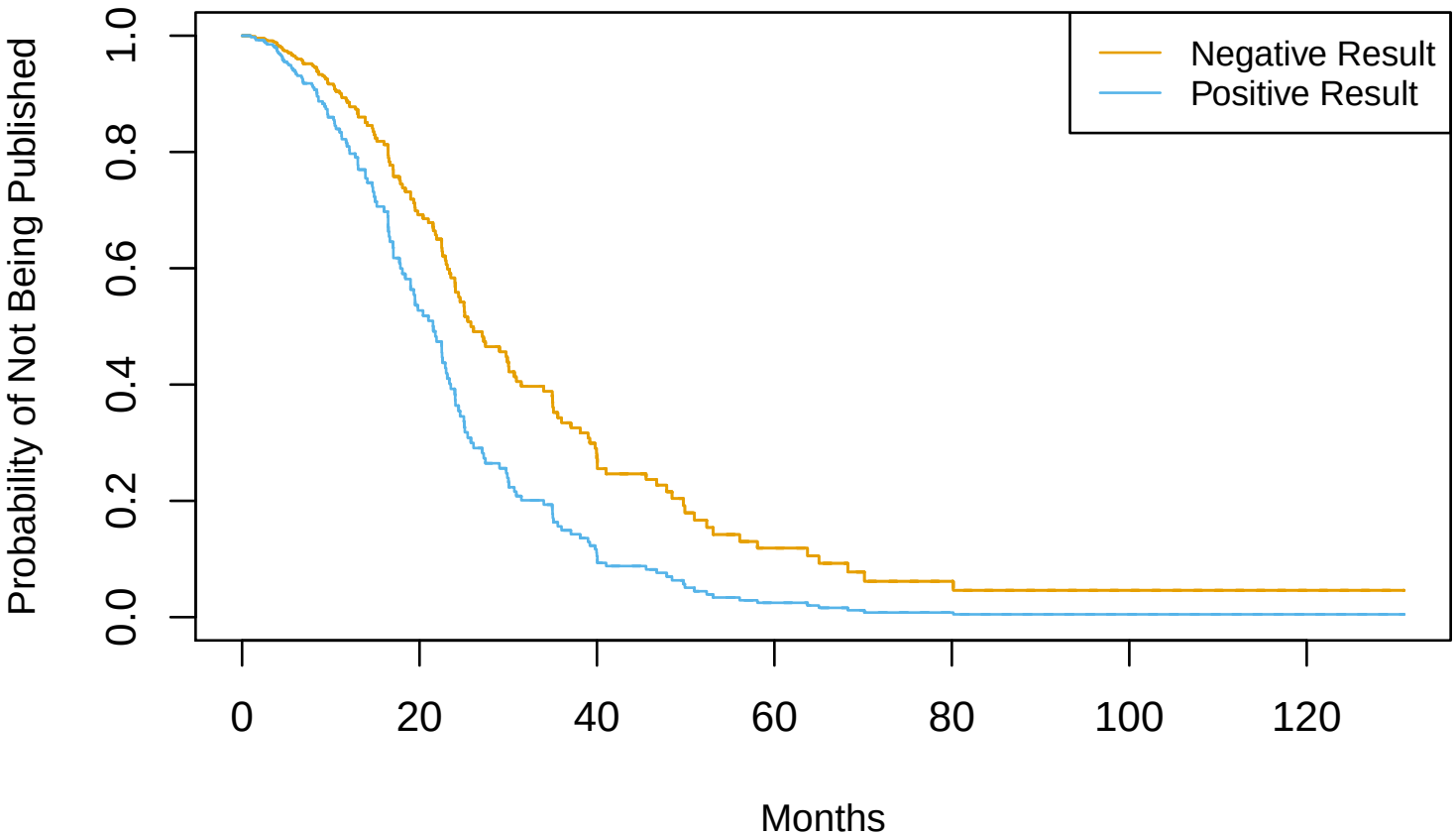
- To produce these survival curves, we estimated the underlying baseline hazard $h_0(t)$: this is implemented in the `survival` package in `R`, although the details are beyond the scope of this course.

Digging Deeper

In order to gain more insight into this result, on the next slide we display estimates of the survival curves associated with positive and negative results, adjusting for the other predictors.

- To produce these survival curves, we estimated the underlying baseline hazard $h_0(t)$: this is implemented in the `survival` package in `R`, although the details are beyond the scope of this course.
- We also needed to select representative values for the other predictors; we used the mean value for each predictor, except for the categorical predictor `mech`, for which we used the most prevalent category (`R01`).

Adjusted Survival Curves



Adjusting for the other predictors, we now see a clear difference in the survival curves between studies with positive versus negative results. *[What has happened?]*

AUC for Survival Analysis: the C-index

AUC for Survival Analysis: the C-index

- This is an appealing method for assessing a fitted Cox model on a test set.

AUC for Survival Analysis: the C-index

- This is an appealing method for assessing a fitted Cox model on a test set.
- For each observation, we calculate the estimated risk score, $\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$, for $i = 1, \ldots, n$, using the estimated Cox model coefficients.

AUC for Survival Analysis: the C-index

- This is an appealing method for assessing a fitted Cox model on a test set.
- For each observation, we calculate the estimated risk score, $\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$, for $i = 1, \ldots, n$, using the estimated Cox model coefficients.
- Then Harrell's concordance index (or *C-index*) computes the proportion of observation pairs for which $\hat{\eta}_{i'} > \hat{\eta}_i$ and $y_i > y_{i'}$:

$$C = \frac{\sum_{i,i':y_i > y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i > y_{i'}} \delta_{i'}}.$$

AUC for Survival Analysis: the C-index

- This is an appealing method for assessing a fitted Cox model on a test set.
- For each observation, we calculate the estimated risk score, $\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$, for $i = 1, \ldots, n$, using the estimated Cox model coefficients.
- Then Harrell's concordance index (or *C-index*) computes the proportion of observation pairs for which $\hat{\eta}_{i'} > \hat{\eta}_i$ and $y_i > y_{i'}$:

$$C = \frac{\sum_{i,i':y_i > y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i > y_{i'}} \delta_{i'}}.$$

- This is the proportion of pairs for which the model correctly predicts the relative survival time, among all pairs for which this can be determined

C-index: Example

We fit a Cox proportional hazards model on the training set of the `Publication` data, and computed the C-index on the test set.

This yielded $C = 0.733$. Roughly speaking, given two random papers from the test set, the model can predict with 73.3% accuracy which will be published first.

Additional Topics

Here are some additional topics that are covered in the text:

Additional Topics

Here are some additional topics that are covered in the text:

- Other types of censoring: left and interval censoring.

Additional Topics

Here are some additional topics that are covered in the text:

- Other types of censoring: left and interval censoring.
- The choice of time scale, e.g calendar time or age?

Additional Topics

Here are some additional topics that are covered in the text:

- Other types of censoring: left and interval censoring.
- The choice of time scale, e.g calendar time or age?
- *Time-dependent covariates* — where we measure a feature (like blood pressure) at different time points

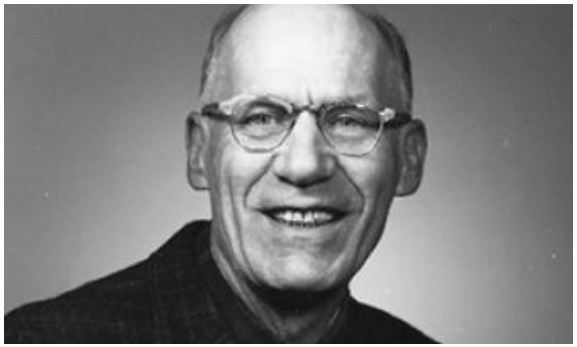
Additional Topics

Here are some additional topics that are covered in the text:

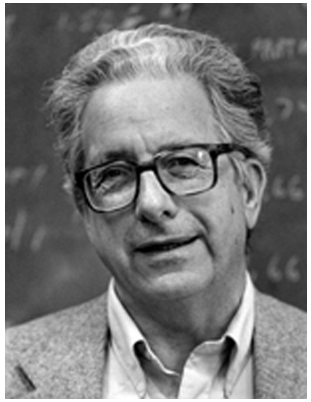
- Other types of censoring: left and interval censoring.
- The choice of time scale, e.g calendar time or age?
- *Time-dependent covariates* — where we measure a feature (like blood pressure) at different time points
- Methods for checking the proportional hazards assumption

There are also approaches for modeling survival data using other machine learning methods such as *random forests, boosting and neural networks*. Some of these avoid the proportional hazards assumption.

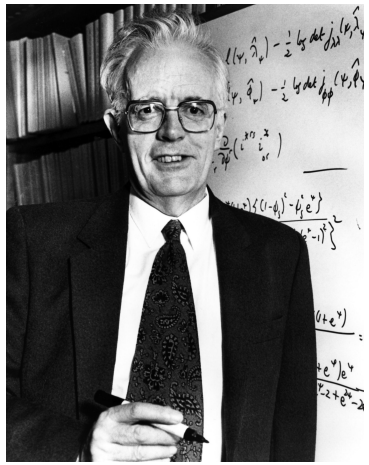
Those big names one last time



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

William Haenszel

(***log rank test***)

Terry Therneau

(author of Survival package in R)

Software for Survival Analysis

Software for Survival Analysis

- The examples in this lecture were creating using the `survival` and `glmnet` packages in `R`.
- Both packages can handle time-dependent covariates and general forms of censoring.

Software for Survival Analysis

- The examples in this lecture were creating using the `survival` and `glmnet` packages in `R`.
- Both packages can handle time-dependent covariates and general forms of censoring.
- Software for other machine learning approaches can be found both the `R` repository and the `scikit-survival` Python collection.