Lecture 13: Introduction to Survival Analysis

Lin Chen

Department of Public Health Sciences
The University of Chicago

What is Survival Analysis?

- Survival analysis is a body of statistical methods that deals with analysis of data where the outcome variable measures the time from a well defined *time origin* until the occurrence of some particular event, time-to-event data.
  - For example, in medical research, the *time origin* will often correspond to the recruitment of an individual into a study. It might also be time of diagnosis. In a social science problem studying school drop-out, it might be beginning of high school.
  - When the *event* is the death, then the resulting data are literally *survival times*; for other *events* that represent the *endpoint* of interest, e.g. occurrence of a disease, relief of pain, etc., the observations are more generally referred to as *time to event* data.
  - The time to event can be measured in any time scale: days, weeks, years, etc, depending on context.

Objectives of Survival Analysis

Quantify and make inference about the time, t, from the origin to the event of interest.

1. **Estimate the time to event distribution for a group of subjects**, such as the lifetimes (from birth) of a cohort of individuals from a certain birth year interval (life tables).
2. **Compare time to event experience between two or more groups**, such as determining whether the addition of an anti-angiogenic agent (Avastin) to radiation and chemotherapy in glioblastoma.
3. **Assess the relationship of covariates/predictor variables to time to event outcomes**: how do age, size of tumor, other personal and disease features influence survival times of men with prostate cancer?

Special Features of Survival Data

Survival time data are

- non-negative random variables, as measured from some calendar origin
- generally not symmetrically distributed (often positively skewed, more large values than small)
- in real life, subject to other practical constraints, i.e., within normal range of lifetimes for given situation, etc
- The main feature of survival data that renders many standard methods inappropriate is that survival times are frequently incomplete, that is, the failure event may not be observed. These survival times are referred to as censored.

Censoring

- The survival time of an individual is said to be *censored* when the endpoint of interest has not been observed for that individual so that the information about their survival time is incomplete.
- Why might censoring occur?
  - An individual does not experience the event before *the study ends*.
  - At any given analysis time, an individual may not have failed yet (due to limited follow-up time).
  - An individual is *lost to follow-up* during the study period.
  - Special case: A survival time for death due to a specific cause may be regarded as censored when the death is from another cause (*competing risks*)

Forms of Censoring

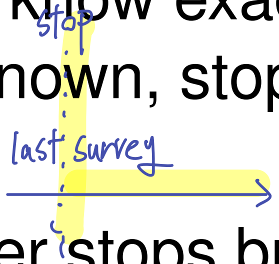
- There are three main forms of censoring:
  - *right censoring* (most common)
  - *left censoring*
  - *interval censoring*.
- Consider a study administering a survey to mothers every other month asking if they are still breastfeeding, where the event/endpoint of interest is the duration of breastfeeding.

Forms of Censoring

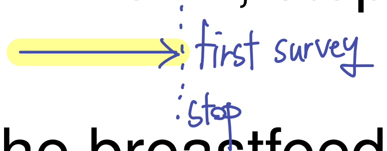
lost to account

Breastfeeding example:

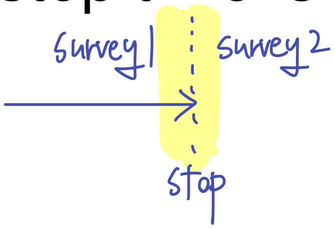
1. *Right censoring:* occurs when mothers are still breastfeeding after the last survey, and we do not know exactly how long they will continue. (The actual, but unknown, stop time is longer than the right-censored observed time)



2. *Left censoring:* occurs if mother stops breastfeeding before the first survey. (The actual, but unknown, stop time is less than the left-censored observed time)



3. *Interval censoring:* occurs if the breastfeeding ended between two successive surveys in which case one can only say that breastfeeding ended somewhere within the past two months. (The actual, but unknown, stop time is known to be within an observed interval of time)



Censoring Mechanism Matters

- Most important assumption, *non-informative censoring*: random distribution of censoring times *is independent of* survival times. This says that the actual survival time of an individual, t, is independent of any mechanism that causes that individual's survival time to be censored at time c (notice that our right censoring assumption says that $c < t$).
- This assumption cannot be made if, for example, a patient drops out of the study because he/she is much sicker and thus potentially has a shorter survival time. This type of censoring is known as *informative censoring*. If informative censoring is present and is not properly addressed, then the analysis will be biased. Unfortunately, informative censoring hard to detect.
- The methods presented in this course assume non-informative right censoring (or a minimal amount of informative censoring).

Why Struggle to Accommodate Censoring?

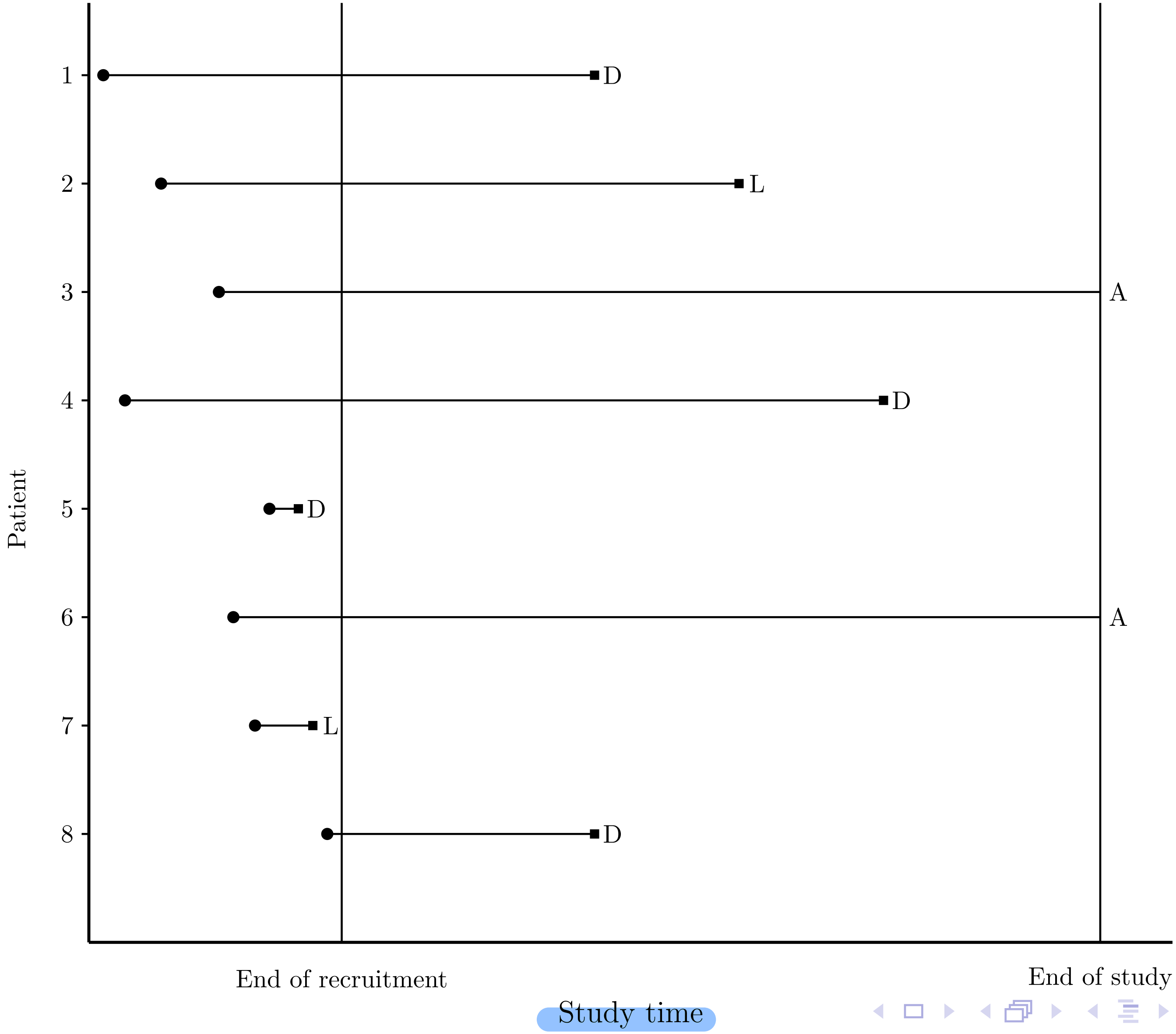
- Example - Estimate 5-year survival in a cohort of 300 patients: 100 died before 5 years, 100 were event-free at 5 years, and 100 were lost to follow-up before 5 years.
- **Simple options for handling censoring have undesirable properties**
  - If we just omit those lost to follow-up before 5 years (censored observations), estimation is highly inefficient at best (losing 100 samples) and likely highly biased downward (if most of those 100 died before 5 years).
  - If we assume those lost are event-free at 5 years, then estimate is biased upward (we are adding fictitious follow-up time).
- Alternatively, censoring can be handled as a 'missing data' problem, with some sort of unbiased imputation applied
- More straightforwardly, we make assumptions about censoring and used estimates that can handle the incomplete observations with maximum efficiency

Study time and *Patient time*

- In highly controlled experiments, the investigators are able to ‘start the clock’ at the same calendar time for all subjects
- In a typical clinical study, patients are not all recruited at exactly the same time, but accrue over a period of months or even years.
- After recruitment, patients are followed up until they fail, or until the end of study.
- Although the actual survival times will be observed for a number of patients, some patients may be lost to follow-up, while others will still be alive at the end of the study (or at any given analysis time).
- The calendar time period in which an individual is in the study is known as the *study time*. total time

Study time and *Patient time* (continued)

This diagram shows the study time for eight patients in a survival analysis. D=dead, A=alive, L=lost to follow-up.



Study time and *Patient time* (continued)

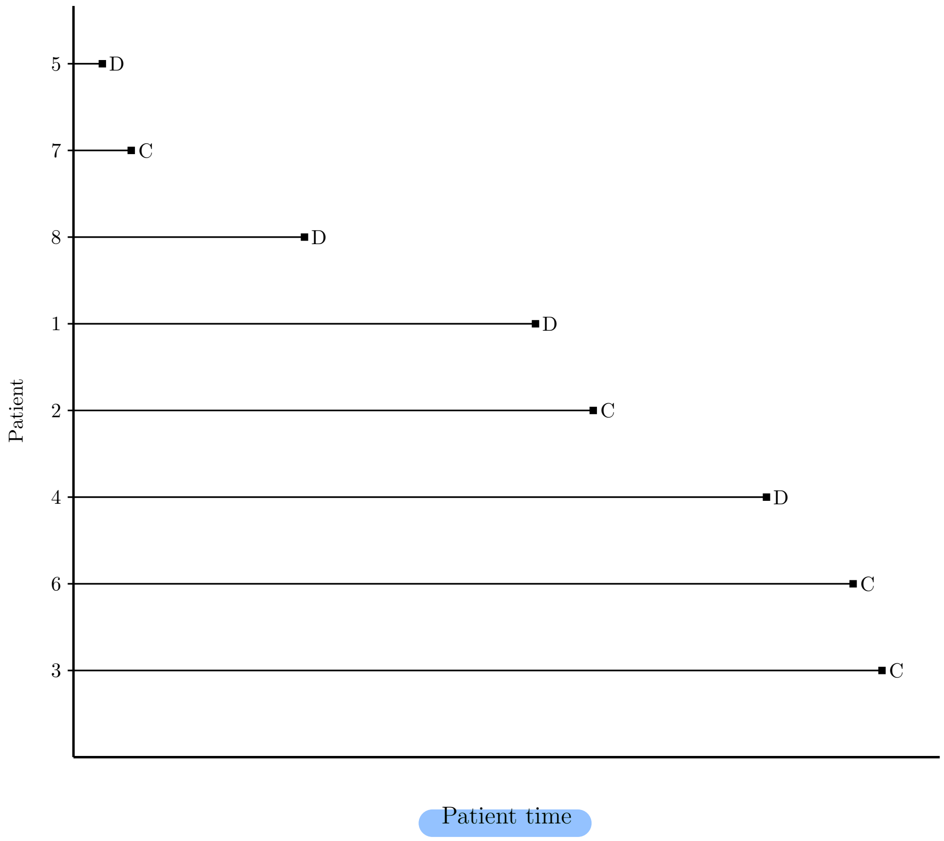
- The period of time that a patient spends in the study, measured from that patient's time origin, is often referred to as *patient time*.

  start → endpoint

- In practice, the data recorded will be in form of *study time*, that is the date on which each individual enters the study, and the date on which each individual fails or was last known to be free of the endpoint. Then *patient time* can be calculated accordingly. The patient/person-time under observation is of most relevance for analysis.

Study time and *Patient time* (continued)

This diagram shows the sorted patient time for eight patients in a survival analysis. D=dead. C=(right-)censored.



The period of time from time origin to death is the survival time for patient 1,4,5,8. The survival time of patient 2,3,6,7 are right-censored. Survival time is also called event time or failure time.

Key Quantities in Survival Analysis

- Define T to be the random variable that measures the survival time (from the time origin to the event). T is
  - Continuous
  - Non-negative
- Examples:
  - time from birth to death (age at death)
  - time from diagnosis of breast cancer to disease recurrence after treatment
  - time from HIV infection to onset of AIDS
  - time from parole to re-incarceration

Key Quantities in Survival Analysis (continued)

- There are several equivalent ways to characterize the probability distribution of T. We will focus on the following terms:
  - The distribution function and its complement, the survivor function $S(t)$
  - The density function $f(t)$
  - The hazard function $h(t)$
  - The cumulative hazard function $H(t)$

Key Quantities in Survival Analysis (continued)

- The *cumulative distribution function* of T is given by

  $$F(t) = P(T < t) \tag{1}$$

  which represents the probability that the survival time is less than some value t.

- The survivor function, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t, thus

  $$S(t) = P(T \geq t) = 1 - F(t) \tag{2}$$

- Given $F(t) (or S(t))$, the probability density function for T is given by

  $$f(t) = \frac{\mathrm{d}}{\mathrm{d}t} F(t) = -\frac{\mathrm{d}}{\mathrm{d}t} S(t) \tag{3}$$

Key Quantities in Survival Analysis (continued)

- The *hazard function* measured the instantaneous failure rate **conditional on survival to time t**:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)} \quad (4)$$

- Comparisons between $f(t)$ and $h(t)$:
  - $f(t)\mathrm{d}t = P($*event occurs in* $[t, t+\Delta t))$.
    Intuitively, one may think of $f(t)\mathrm{d}t$ as being the probability of survival time T falling within the infinitesimal interval $[t, t+\Delta t))$.
  - $h(t)\mathrm{d}t = P($*event occurs in* $[t, t+\Delta t) \mid T \in [t, \infty))$
    $h(t)$ is a conditional failure function, conditioning on a person has actually survived time t.

$$h(t) = \frac{f(t)}{S(t)} = -\frac{-\frac{\mathrm{d}}{\mathrm{d}t}S(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}\log S(t) \quad (5)$$

- The function $h(t)$ is also referred to as the *hazard rate*, the *instantaneous failure rate*, the *force of mortality*, or *conditional mortality rate* (in some contexts, for mortality data).

Key Quantities in Survival Analysis (continued)

- The *cumulative hazard function* is defined to be

$$H(t) = \int_0^t h(u)\mathrm{d}u \tag{6}$$

  thus

$$H(t) = \int_0^t h(u)\mathrm{d}u = -\int_0^t \frac{\mathrm{d}}{\mathrm{d}u}\log S(u)\mathrm{d}u = -\log S(t) \tag{7}$$

  and

$$S(t) = \exp(-H(t)) = \exp(-\int_0^t h(u)\mathrm{d}u) \tag{8}$$

- Thus, given any of $h(t)$, $f(t)$, or $S(t)$, one would be able to estimate the other two functions.
- The cumulative hazard function is not easy to interpret but it is mathematically useful to connect hazard function and survivor function.

Statistical Methods in Survival Analysis

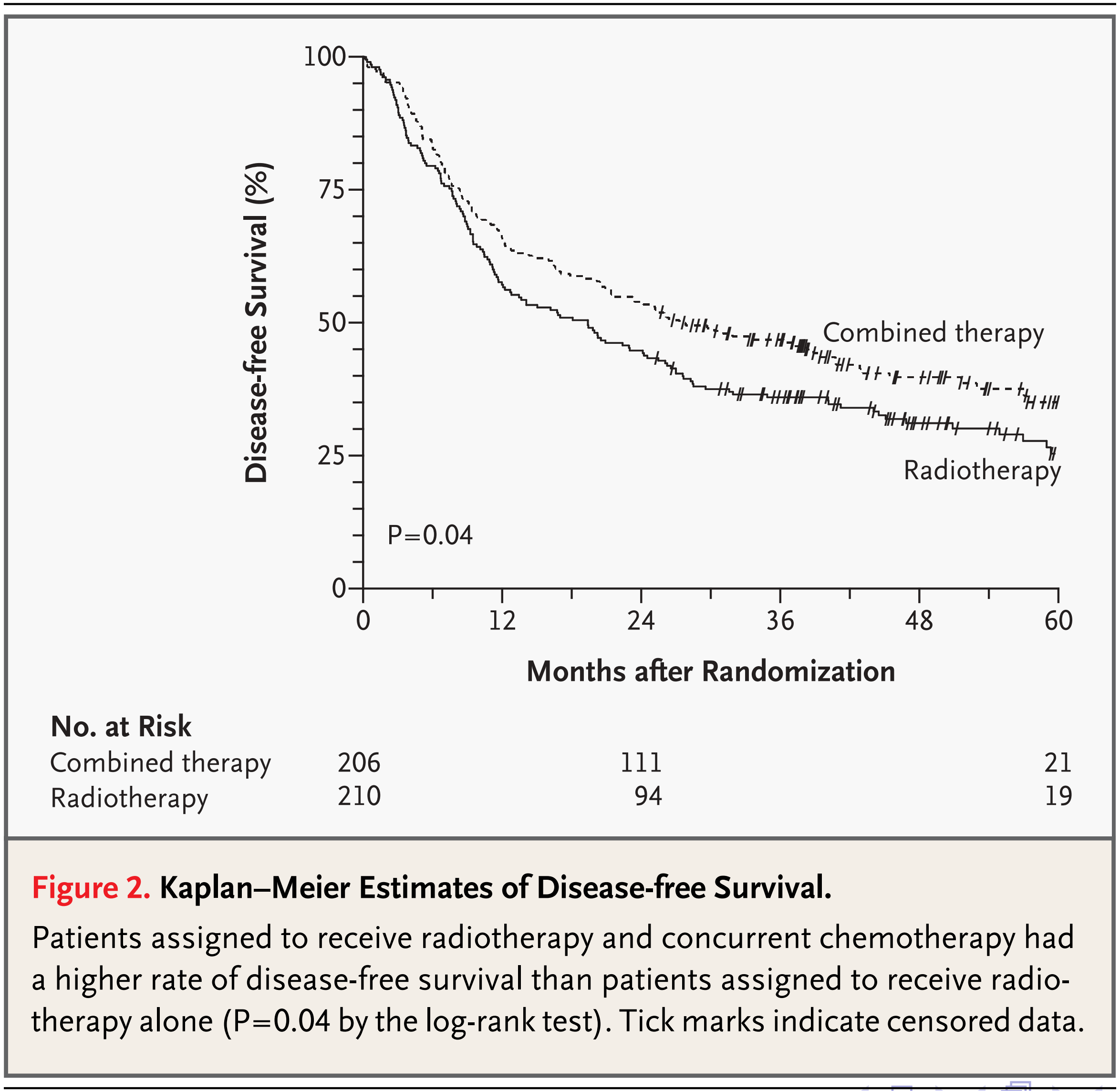
We can estimate the survival/hazard function in three ways:

1. Non-parametric methods (making no assumptions about the form of the survival/hazard functions): e.g. Kaplan-Meier estimator, Life-table estimator, Nelson-Aalen estimator $S(t)$
2. Parametric methods (restricting the form of survival/hazard functions to a specified family of distributions): e.g. exponential , Weibull , or Log-logistic model modeling
3. Semi-parametric methods (having both parametric and non-parametric components): e.g. Cox proportional hazards model. modeling

- (1) is used to estimate $S(t)$, which is of more interest in survival analysis than many other areas of statistics. (2) and (3) are used more in modeling (relating covariates to hazard/survivor function)

Examples of Survivor Functions

Cooper et al NEJM 2004: Head and Neck Cancer (pharynx, larynx, oral cavity)



No. at Risk

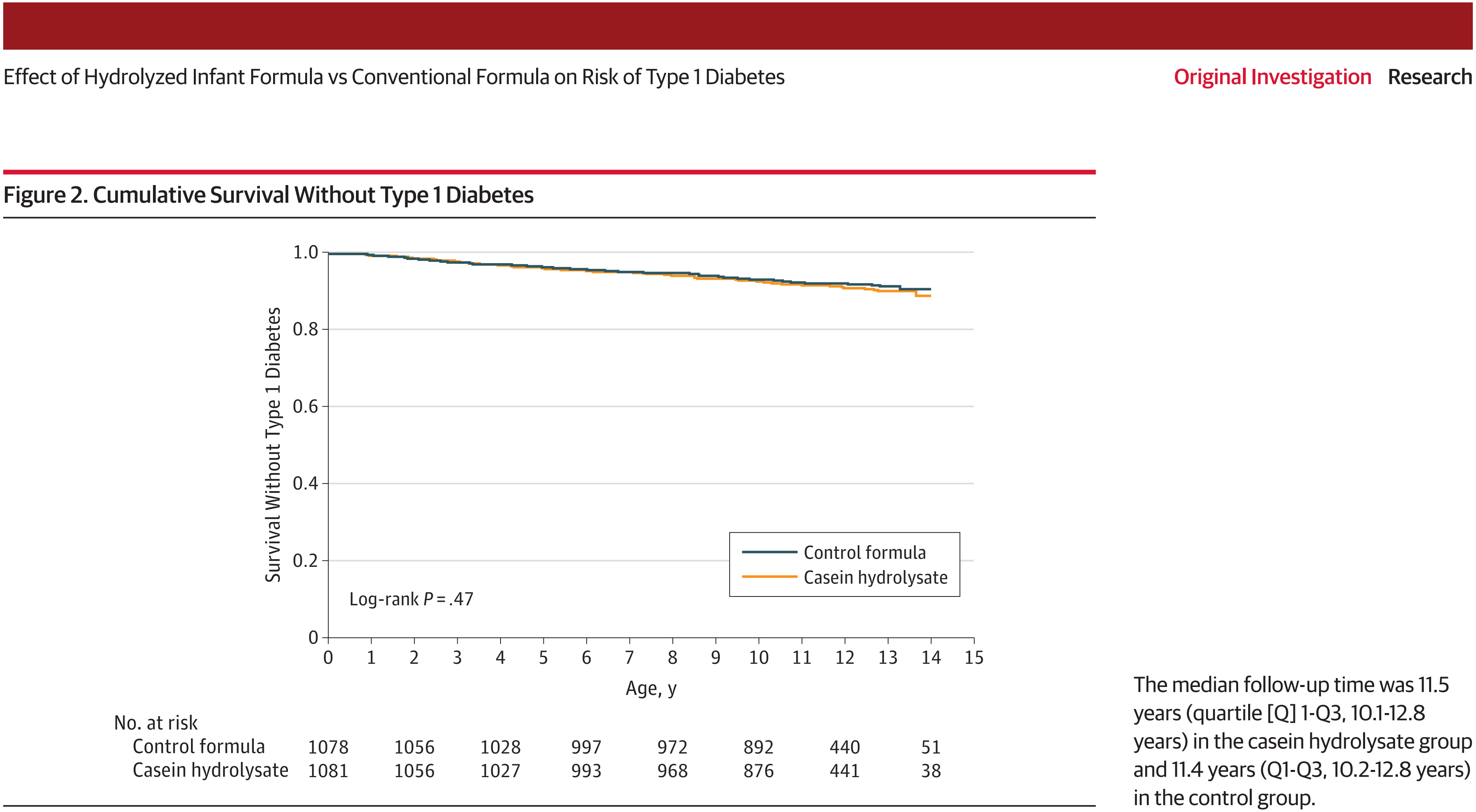
Combined therapy	206	111	21
Radiotherapy	210	94	19

Figure 2. Kaplan–Meier Estimates of Disease-free Survival.

Patients assigned to receive radiotherapy and concurrent chemotherapy had a higher rate of disease-free survival than patients assigned to receive radiotherapy alone (P=0.04 by the log-rank test). Tick marks indicate censored data.

Examples of Survivor Functions

Knip et al JAMA 2004: Diabetes in Infants



Effect of Hydrolyzed Infant Formula vs Conventional Formula on Risk of Type 1 Diabetes

Original Investigation Research

Figure 2. Cumulative Survival Without Type 1 Diabetes

No. at risk								
Control formula	1078	1056	1028	997	972	892	440	51
Casein hydrolysate	1081	1056	1027	993	968	876	441	38

The median follow-up time was 11.5 years (quartile [Q] 1-Q3, 10.1-12.8 years) in the casein hydrolysate group and 11.4 years (Q1-Q3, 10.2-12.8 years) in the control group.

Summary

Survival Analysis

- Time to event is the outcome of interest - can be thought of as extension of binary data
- Incomplete event times, or censoring, is the main distinguishing feature
- Key quantities in survival analysis: $\overset{\text{density}}{f(t)}$, $\overset{\text{hazard}}{h(t)}$, $\overset{\text{survival}}{S(t)}$ and $\overset{\text{cumulative hazard}}{H(t)}$.
- Next: Estimation and inference