

Due: September 27, 2022 at 1:30 pm section
Possible points: 29

Note: For the questions marked “FOR DISCUSSION IN SECTION ONLY:”, please familiarize yourself with these items before section such that we can have a rewarding discussion. Also, please be sure to complete all of the required problem set questions. The first “FOR DISCUSSION” question (question 9d) does NOT mark the end of the required problem set.

1. PRACTICE QUESTIONS (please attempt to answer these practice questions 1a-e, but it is not required that you type your answers into this document for practice questions 1a-e):
State the measure of disease occurrence (e.g., point prevalence, cumulative incidence, etc.) that is given or can be calculated from the information provided for each study.

- (a) Among all persons being tested for SARS-CoV-2 (the causative viral agent of COVID-19) RNA through nasal swabs or provision of saliva, public health officials are closely following the fraction who are found to have SARS-CoV-2 RNA (the fraction who test positive).

[Practice question; click to see answer](#)

- (b) Investigators performed a one-time survey of 500 residents of California and asked whether respondents had experienced a “common cold” anytime during the past year. Two hundred residents responded “yes”. The survey was conducted from Jan 1, 2019 to Dec. 31, 2019.

[Practice question; click to see answer](#)

- (c) Among 600 initially HIV-negative (uninfected) men, 85 men acquired HIV infection during follow-up. There were no competing events. Because testing was anonymous, it is not known exactly which men acquired HIV. The mean follow-up time (defined as time from entry until either HIV acquisition, drop out, or administrative censoring) for the 600 men was 6.5 years.

[Practice question; click to see answer](#)

- (d) The probability that an elderly participant in the Framingham Heart Disease study, which followed residents of Framingham, Mass. for many years for heart disease outcomes, experienced death after 10 years of follow-up was estimated at 0.12.

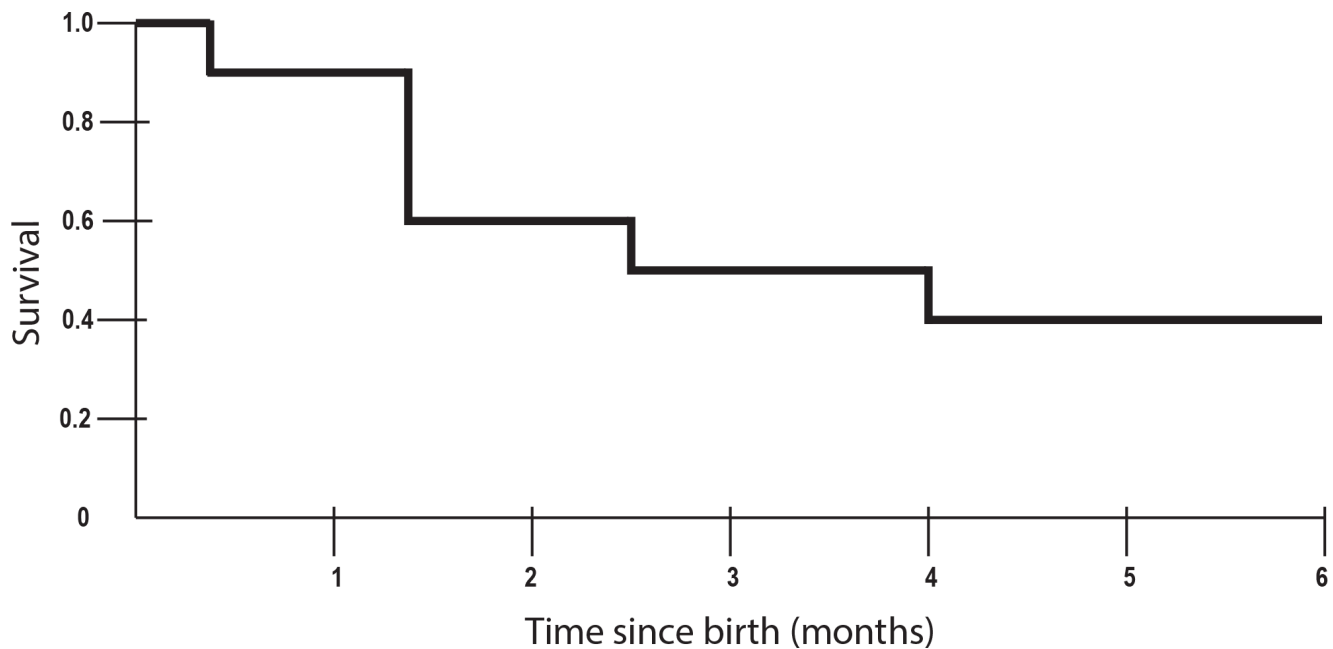
[Practice question; click to see answer](#)

- (e) Among 125 healthy attendees of a church picnic, all were contacted 5 days after the event and asked whether they had at least one diarrhea episode since the picnic. Each individual provided a yes or no answer, and a total of 34 reported having had diarrhea since the picnic.

[Practice question; click to see answer](#)

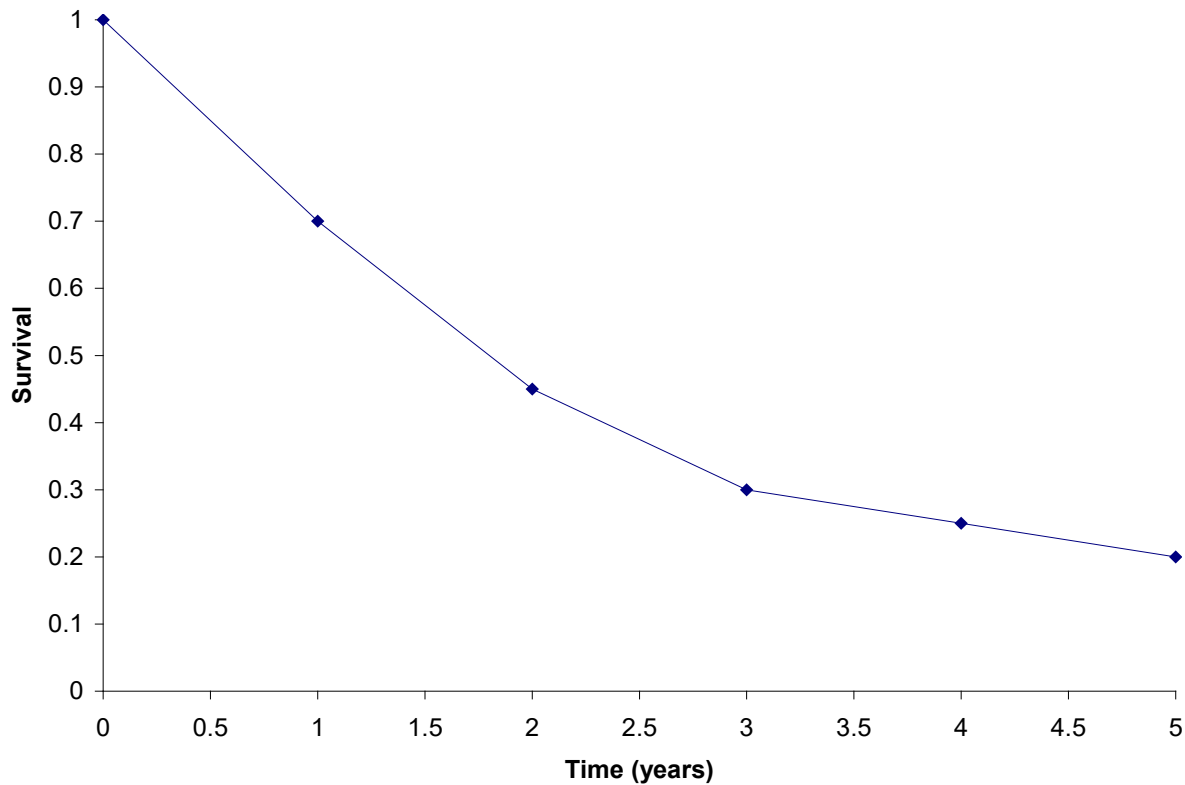
***NOTE: If you are unable to view the graphs in questions 2 and 3 (in the MS Word version of this problem set), please see the PDF version of Problem Set 2 (available on the EPI 203 CLE website).**

2. Results of Kaplan-Meier estimation are shown below from a cohort study of infants with very low birth weight. Infants were followed for death for 6 months following birth.



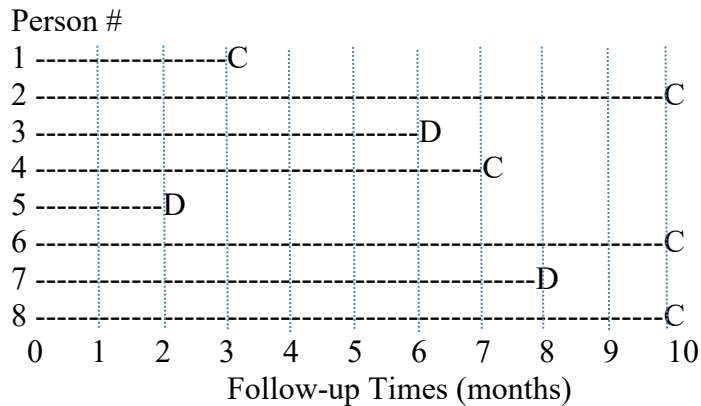
- (a) What is the cumulative survival and cumulative incidence of death at 2 months? How about at the end of follow-up? (2 pts)
- (b) What is the cumulative incidence of death at 8 months following birth? (1 pt)
- (c) At what timepoint does the cumulative incidence of death in this study sample first surpass 50%? (1 pt)
- (d) Why does the Kaplan-Meier plot look like a step function? In other words, why does it look like it does, with a series of steps? (1 pt)

3. Results from a well-ascertained fixed cohort study of 1000 patients with a form of lung cancer are shown below. The formatting of the plot was deemed technically correct by a group of methodologic experts. The 5-year cumulative incidence of death was 80%. Does this analysis use the life table or Kaplan-Meier approach to estimation of incidence? Explain your answer. (1 pt)



4. Using the Kaplan-Meier method of estimating cumulative incidence:

- (a) Display in a table (like that shown in the lecture slides (in the section regarding calculation of the Kaplan-Meier estimator) or Table 2-3 in Szklo and Nieto (p.59)) the data and the probabilities concerning death and survival in sequential time intervals for the information below on follow-up of 8 persons diagnosed with lung cancer. Please construct the table by hand, but it is fine to use a calculator for the calculations. In the figure below, C indicates censored and D indicates death. (2 pts)



- (b) Calculate the cumulative incidence of death at 8 months. (1 pt)
- (c) Draw a graph of the survival function (i.e., survival over time) - by hand (or use the manual drawing tools in your word processor). Please write neatly such that we can accurately evaluate your plot. Be sure that axes are labeled. (1 pt)

5. Consider the following fixed cohort study investigating the occurrence of death following diagnosis of pancreatic cancer in 9 patients.

Patient	Outcome
1	Died at 1 month
2	Still alive at 10 months
3	Died at 6 months
4	Moved away at 7 months
5	Died at 5 months
6	Last known to be alive at 9 months
7	Died at 4 months
8	Stopped coming back to clinic visits after 2 months
9	Died at 10 months

- (a) In preparation for a survival analysis estimating the cumulative incidence of death, enter the data into Stata and show a listing of the entire dataset. Show ALL your commands/output. (1 pt)
- (b) Using Stata, determine the 10-month cumulative incidence of death. Show your commands/output. (1 pt)

6. The San Francisco Men's Health Study enrolled 394 HIV-infected men who have sex with men in 1984 and designed the study to follow them for the next 10 years to determine the natural history of HIV infection (the research was all conducted prior to the current era where effective therapy is available). NOTE: this study was conducted when HIV/AIDS research was in its infancy and this study was one of the first to have long term follow-up. Follow-up occurred by means of examination of participants every 3 months at the central research clinic. The final research study visit for each participant, when available, was conducted after 9.75 years of follow-up and could be done on any day until 10 years of follow-up. The outcome was diagnosis of AIDS. Using Stata and the Stata data file that you can download from the course website, called "Epi203survival.dta":

(a) List the Kaplan-Meier estimates of the survivor function (i.e., probability of not experiencing AIDS) at different time points over the 10-year period. Show your commands/output. (Assume no competing events.) (2 pts) (Note: if you wish to save space, you may abbreviate the output after annotating what you have done.)

(b) Determine the cumulative proportion of HIV-infected persons developing AIDS (i.e., cumulative incidence of AIDS) at 8 years of follow-up. Give your answer to 4 significant digits. (1 pt)

(c) Plot the Kaplan-Meier survival curve for all participants using Stata. Include (1) a table under the plot showing the participants still at risk; and (2) the 95% confidence interval around each estimate on the curve. Show ALL of your commands/output and make sure to label your axes appropriately. Why is it useful to include such a table of subjects still at risk? (1 pt)

(Hint: Use the drop down menu for the graph command, found at: Statistics/Survival Analysis/Graphs/Kaplan-Meier Survival Function, to locate the options to show participants still at risk, the confidence intervals, as well as a variety of other options for the Kaplan-Meier graph. Or, search the online user's manual for the options associated with the "sts graph" command.)

(d) Plot the Kaplan-Meier failure curve (i.e., incidence of AIDS), and in this plot show:
-- the 99% confidence interval around the points
-- only the first five years of observation, and
-- to emphasize the high incidence of AIDS, limit the y axis to between 0 and 0.5. (1 pt)

7. Stata will show the numbers and pattern of censoring in a Kaplan-Meier analysis if you use the command `sts graph, lost`
- (a) Using the same Stata dataset (Epi203survival.dta), execute this command and display the graph. The number censored at different times of follow-up are displayed on the graph. As noted earlier, there were no competing events. Assuming that all the subjects started follow-up on the same day (January 5, 1984), is the censoring shown on the plot due to drop out, administrative censoring, or both? If both, which censoring events are drop out and which are administrative censoring? Explain your answer. (1 pt)
Hint: The numbers at the far right of the graph will benefit from magnification in order to be read. One approach, in addition to the main graph, is to magnify the far right by using the command:
`sts graph if obstime >9.5 & obstime <=10, noorigin yscale(off) ylabel(.9(.01)1)
xlabel(9.5(.05)10) lost`
- (b) Inspect the pattern of censoring in the plot from part (a). Assuming again that all the subjects started on the same day (January 5, 1984), are you concerned the nominal estimate of cumulative incidence of AIDS at 10 years might be wrong (i.e., biased)? If so, does the censoring pattern suggest anything about the possible direction of bias in the estimated cumulative incidence of AIDS at 10 years? (1 pt)

8. Let's assume that formal funding for the San Francisco Men's Health Study ended at 10 years such that the participants were no longer brought in for serial research-level evaluations. The investigators, however, continued to occasionally hear from the research participants and their families, usually (but not always) upon the occurrence of death of one of the participants from AIDS. They jotted down these communications and the AIDS diagnoses whenever they happened. At 15 years following start of the study, they decided to calculate another cumulative incidence of AIDS (extending their prior work by 5 more years) based upon the data they had collected by these various communications since the formal study ended. Do you think this updated 15-year estimate is valid? Why or why not? (1 pt)

9. Consider the following abstract from a study of the risk of type 1 diabetes in siblings of type 1 diabetic patients:

The aims of our analysis were to obtain the risk estimates for type 1 diabetes in the siblings of a Finnish population-based group of childhood-onset diabetic patients. We determined the diabetes status of all siblings of all children for whom type 1 diabetes was diagnosed before age 18 years between 1970 and 1979. Information of siblings' dates of birth was available from a national database. Siblings' diabetes status from their birth onwards was ascertained by a database and medical record search through 2001. The total number of person-years during the follow-up of the siblings was 405,685. Of the 10,168 siblings at risk, 647 (6.4%) had been diagnosed with type 1 diabetes by 2001. The cumulative incidence of type 1 diabetes by ages 10, 20, 30, 40, and 50 years in the siblings was 1.5, 4.1, 5.5, 6.4, and 6.9%, respectively.

Notes: Cumulative incidence was calculated using the Kaplan-Meier technique. Although a few deaths from non-diabetes causes occurred among siblings, consider these negligible in number and, therefore, for Questions 9a-c, assume no competing events. Assume that there were deaths after a diabetes diagnosis.

- (a) The abstract states that 647 (6.4%) of the siblings at risk developed type 1 diabetes. Interpret what this 6.4% means in terms of measure of disease occurrence? Is it a useful estimate of incidence? (2 pts)
- (b) Based on the information provided in the abstract, would a sibling born in 1991 of one of the childhood-onset diabetic patients contribute data to the estimation of cumulative incidence of diabetes by age 10 years? Would a sibling born in 1991 contribute to the estimation of cumulative incidence of diabetes by age 50 years? (1 pt)
- (c) Thinking about all of the siblings who would be included in this overall analysis leading to the results stated in the last line of the abstract, what do you believe is their earliest year of birth? (1 pt)

(d) FOR DISCUSSION IN SECTION ONLY:

The cumulative incidence of type I diabetes by age 20, among siblings of individuals with type 1 diabetes, is stated to be 4.1% in this article. However, in a survey taken in 2001 among 20-year-old siblings of individuals with type 1 diabetes in this region, the prevalence of type 1 diabetes was 12%. What could explain the discrepancy in these reports (4.1% vs. 12%), other than chance? Assume that the method of ascertainment of type 1 diabetes in each report was the same and that it was both highly sensitive and specific. Also assume that once develops type I diabetes, it cannot be cured. (Clue: we are not looking for a specific biological explanation but rather a general methodologic phenomenon).

10. The following abstract and results are from a study of nursing home placements:

Objective: To assess cumulative incidence of nursing home placement and non-cognitive factors predicting nursing home placement in a defined older population.

Design and setting: Six-year follow-up of a population-based cohort living in New South Wales.

Participants: 3654 non-institutionalized residents aged 49 years or older (82.4% of those eligible) participated in baseline examinations during 2012 to 2014.

Main outcome measure: Permanent nursing home admission for long-term institutionalized aged care.

Results: After excluding 384 participants who moved from the area or were lost to follow-up, 162 participants (5.0%) had been admitted to nursing homes on a permanent basis by October 2019. Six-year cumulative incidences for nursing home placement were 0.7%, 1.1%, 2.4%, 3.9%, 9.0%, 18.3% and 34.9% for people aged at baseline 55-59, 60-64, 65-69, 70-74, 75-79, 80-84 and 85 years or older, respectively.

Conclusions: Incidence of institutional aged care doubled for each five-year interval from the age of 60 years.

- (a) The authors excluded from the analysis the 384 participants who moved from the area or were lost to follow-up. This means they were taken out of the analysis dataset entirely. Do you agree with this exclusion? Explain your answer. (1 pt)
- (b) Instead of excluding these 384 participants, let's assume that the authors censored them at their last known time of observation and calculated cumulative incidence by the Kaplan-Meier technique. Describe the possible effect of this censoring on the reported cumulative incidence estimates. First, describe one scenario regarding the future fate of these 384 after censoring which would cause the reported (nominal) cumulative incidence to be an underestimate of truth. Second, describe one scenario which would cause the reported cumulative incidence to be an overestimate of truth. (1 pt)

(c) FOR DISCUSSION IN SECTION ONLY:

If the cumulative incidence estimates were calculated by the Kaplan-Meier technique, is there anything else – other than the moves/losses to follow-up mentioned in (a) – that makes you concerned about the validity of the estimates? Explain your answer.

11. For the following questions, please refer to the article by Ondrusova and Ondrus to be discussed in Journal Club. (Ondrusova and Ondrus. Epidemiology and treatment delay in testicular cancer: a retrospective study. *Int Urol Nephrol.* 40: 143-148, 2008)

- (a) What is the incidence of testicular cancer in 30 to 34-year old men in the Slovak Republic between 1993 and 2002? Is the estimate provided an example of cumulative incidence or an incidence rate? (1 pt)
- (b) Explain how you think the incidence of testicular cancer in 30 to 34-year-old men was derived? In other words, how do you think the authors got their estimate? (2 pts)

(c) FOR DISCUSSION IN SECTION ONLY:

The authors contend that the incidence of testicular cancer has increased from 1968 to 2002 in the Slovak Republic. We take this to mean that the biologic occurrence of testicular cancer in men has increased in a relatively short period of time in this country. From a public health perspective, this is a very worrisome development — a very big concern. Can you describe any alternative explanations for their data?

ANSWERS TO PRACTICE QUESTIONS 1a-e:

- (a) Answer: Point prevalence of SARS-CoV-2 RNA-positivity. Despite this clearly being a prevalence, both scientists and the lay press typically refer to this as the COVID-19 “positivity rate”. This is another example of terminology confusion. Regardless of what it is called, it is receiving much attention regarding how much SARS-CoV-2 is in a community and whether a community is performing enough testing to find all or most of its cases. Other than extremely high or level prevalence, we find the measure difficult to interpret or follow because it is highly influenced by the nature of what types of persons are tested.
- (b) Answer: Period prevalence of the common cold. The key element here is that the “common cold” could have occurred anytime during the participant’s prior 12 months of life. It may have been present at the beginning of the prior one-year period or it may have occurred sometime after the beginning of the period. Period prevalence is not often used but is most useful for measuring burden of disease for conditions with short-term duration. This study design should not be confused with a cohort study that would measure incidence. This is because, first, the design is counting the presence of prevalent common colds at the beginning of the period of questioning. Cohort designs usually study initially disease-free individuals for their subsequent occurrence of disease. Second, cohorts are defined by individuals at the beginning of a period of observation who are then followed forward and not by finding a group of persons who are asked about an outcome during some prior period of time. In this prevalence study, the residents who are questioned about their past year may be different from those who were in the study base at the beginning of the year period of observation. Specifically, some persons present at the beginning of the observation period may now be dead. One cannot identify a group of individuals in current time (i.e., today) and declare that they are a fixed “cohort” of persons who have experienced the last, say, 1 year together. Cohort studies are not constructed by working backwards. Instead, cohort studies are always defined by time 0 characteristics of its participants.
- (c) Answer: Incidence rate of HIV infection. (*Not required for credit:* this is technically an average incidence rate.) Using just the information that is provided, one can only calculate incidence using a person-time denominator (i.e., an incidence rate). The numerator will be 85 infections and the denominator 600 persons x 6.5 years to give 3900 person-years (the rate would be 2.2 per 100 person-years). This illustrates how one can estimate an incidence rate with scant information. We don’t know whether the rate is constant throughout the period of observation, but this does not preclude the calculation of the rate. Although cumulative incidence with the Kaplan-Meier method could theoretically be calculated in the group, you would need the time of each HIV infection for that calculation, as well as the time of observation for all those who did not develop HIV infection. We are told that these data do not exist (or at least they are not available to us). Therefore, it would be incorrect to say that we can estimate cumulative incidence by the Kaplan-Meier technique. It might be possible to construct some semblance of a life table from the information given, but only by making some strong assumptions. For example, you could assume an equal number of infections each year and an equal number of drop outs each year and then construct a collective population at risk over a series of time intervals that would add up to the 3900 person-years. This, however, would be a lot of work to create and a dubious application of the life table method particularly because it seems unlikely that a fixed cohort of men (as represented in a life table) would have a constant rate of HIV infection (instead, one would expect a decline in rate as the susceptibles are exhausted). Furthermore, the answer obtained in this approach, which takes a lot of work, would be no more accurate than simply using the assumed constant rate in the exponential formula to calculate cumulative incidence at any specific time point (will be explained in later lectures).

- (d) Answer: This probability (0.12) is the 10-year cumulative incidence of mortality. It is sometimes given the special term “cumulative mortality”. Because the estimate is a probability (a proportion), it is based on persons at risk. It is not a person-time rate. The presence of the heart disease outcomes presents no complications in the estimation of the cumulative incidence of death.
- (e) Answer: This study features equal duration of follow-up for all individuals. Thus, a simple proportion can be used to calculate cumulative incidence. In this case, the cumulative incidence of diarrhea is 34 out of 125 (27%) at 5 days. Although the metric in this scenario is commonly referred to as an “attack rate”, it is not a rate. Moreover, it would be a considerable stretch to attempt to derive an incidence rate with the available data, and this should not be one’s initial instinct. To derive an accurate rate, one would need to make some assumption about when the cases occurred in the five-day interval; we can see no practical reason why one would want to construct a rate. Finally, this could technically be construed as a period prevalence if prevalent diarrhea at the time of the picnic was counted, but this would be very unconventional and non-forward thinking. Speaking in terms of a cumulative incidence for a cohort at the time of its initial exposure is far more coherent than describing the prior health conditions of a group of people sometime following exposure.