Competing risks

Patrick Breheny

December 10

Introduction

- Throughout this class, we've looked at data and methods for studies involving the time until a single event
- Even when multiple events were present (e.g., death and liver failure in the PBC data), we combined them into a single event (progression-free survival)
- In this lecture, we'll consider the problem of how to analyze multiple, distinct failure types, treating them as separate outcomes rather than combining them (this is a rather complex topic, so we'll just get an overview of the main issues today)

Endometrial cancer data

- As a motivating example, we'll consider data from a study of endometrial adenocarcinoma, the most common gynecologic cancer in the United States
- The study focused on medically inoperable patients – i.e., women with serious comorbidities such as diabetes and cardiovascular disease that make surgery too risky as a treatment option
- These women were treated with radiation therapy only, with data coming from a consortium of five academic cancer centers

Endometrial cancer data (cont'd)

- Follow-up data on 74 women was available from the time of diagnosis until the time of either death or the end of the study
- The cause of death was also recorded, as either having been caused by the endometrial adenocarcinoma itself, or having been due to other causes
- In addition to the time until death, we also have data on the time until recurrence of the cancer following radiation therapy
- The three outcomes in this study provide a good illustration of the various kinds of relationships that can occur when considering multiple failure types

Recurrence vs. time to death from other causes

- For example, consider the relationship between recurrence and time to death from other causes
- It is possible that recurrence of the cancer increases the risk of death from other causes, but has no effect on the risk of say, dying due to cardiac problems
- This question could be reasonably handled through the use of time-dependent covariates, as discussed in the previous lecture

Recurrence vs. disease-specific death

- On the other hand, the relationship between recurrence and disease-specific death would not make much sense to model using time-dependent covariates, at least using a proportional hazards model, since it would be impossible to die of cancer without the cancer first recurring
- Instead, something like a multi-state model might be of interest:

$$\text{Remission} \xrightarrow{\lambda_1} \text{Recurrence} \xrightarrow{\lambda_2} \text{Death}$$

where λ_1 and λ_2 represent conditional hazards, also known in this context as transition rates

Disease-specific death vs. death from other causes

- The relationship between disease-specific death and death from other causes is different yet again, since these two outcomes are mutually exclusive
- In a sense, one could treat death from other causes as a censoring event, but this isn't exactly right, as it implies that once an individual dies from other causes, we still don't know when they might die of cancer
- This situation, in which only one event out of a group of potential events can occur in any given subject, is known as the problem of *competing risks*

Type-specific hazard

- Let us begin by extending our mathematical definitions of hazard and related quantities to accommodate multiple failure types
- Let T denote the time until failure, and K indicate the type of failure
- The type-specific hazard is then defined as

$$\lambda_k(t) = \lim_{h \to 0} \frac{\mathbb{P}\{t \leq T < t + h, K = k | T \geq t\}}{h}$$

Overall hazard and survival

- In the case where the failure types are mutually exclusive (i.e., the case of competing risks), the overall hazard is

$$\lambda(t) = \sum_k \lambda_k(t)$$

- Likewise, the overall survival is

$$S(t) = \exp\left\{-\int_0^t \lambda(s)ds\right\}$$

- Note that it makes sense to discuss type-specific hazards, but typically doesn't make sense to describe type-specific survival – if a subject survives, they survive all of the risks involved

Subdistributions

- Thus, while we have largely focused on estimating and examining survival functions throughout the course, we will have to use something else in the case of competing risks
- One possibility would be the cumulative hazard, but that tends to be unpopular due to the difficulty of interpreting it
- A widely used alternative is to extend the density and distribution functions to accommodate type-specific failures
- Because the time to a type-specific event no longer has a proper distribution, these extensions are known as the "subdensity" and "subdistribution" functions

The cumulative incidence function

- The *subdensity* function for type k is defined as

$$f_k(t) = \lambda_k(t)S(t)$$

- Similarly, the *subdistribution* function is defined as

$$F_k(t) = \int_0^t f_k(s)ds$$

- The subdistribution function is also known as the *cumulative incidence function*, which is the common name for this quantity in applied work

Nonparametric likelihood

- Nonparametric estimation of cumulative incidence functions is very similar to the nonparametric maximum likelihood estimation of survival functions in the Kaplan-Meier case
- Letting $t_1 < t_2 < \cdots$ denote the unique failure times, d_{jk} denote the number of failures of type k at time t_j, and n_j denote the number at risk at time t_j, the nonparametric MLEs for $F_1, \ldots, F_k$ can be found by maximizing the likelihood

$$\prod_j \prod_k \lambda_{jk}^{d_{jk}} (1 - \lambda_j)^{n_j - d_j},$$

where $\lambda_{jk} = \lambda_k(t_j)$, $\lambda_j = \sum_k \lambda_{jk}$, and $d_j = \sum_k d_{jk}$

Estimate of cumulative incidence

- Maximizing the likelihood on the previous slide yields

$$\hat{\lambda}_{jk} = \frac{d_{jk}}{n_j}$$

- These estimates, in turn, yield an estimate of the cumulative incidence function via

$$\hat{F}_k(t) = \sum_{t_j \leq t} \hat{\lambda}_{jk} \prod_{t_i < t_j} \{1 - \hat{\lambda}_i\};$$

note that the term involving the product is simply the Kaplan-Meier estimate $\hat{S}(t_j^-)$

R code

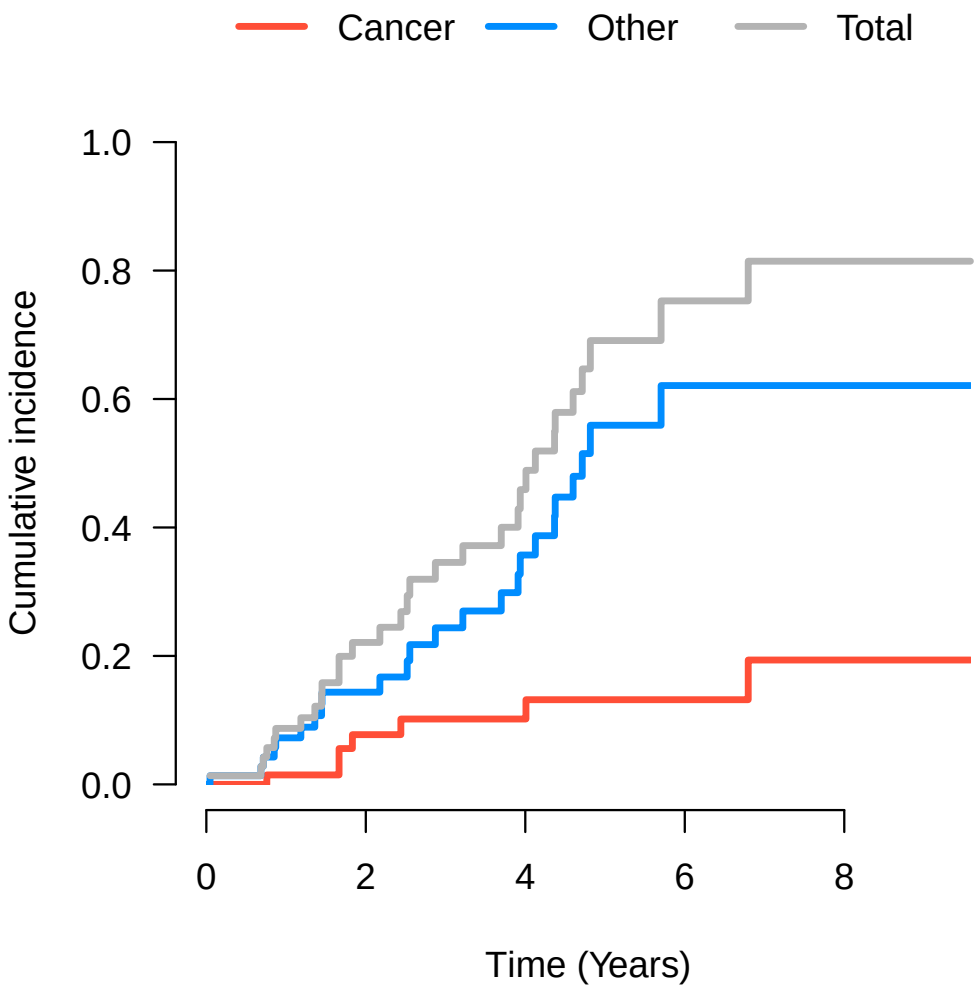
- The `survfit` function can also be used to fit incidence functions
- To use it, instead of passing 0/1 as the status indicator to `Surv`, one supplies a factor, with the first level taken to be the censoring indicator
- Otherwise, the code is the same:

```
fit <- survfit(Surv(tDeath, sDeath) ~ 1, Data)
```

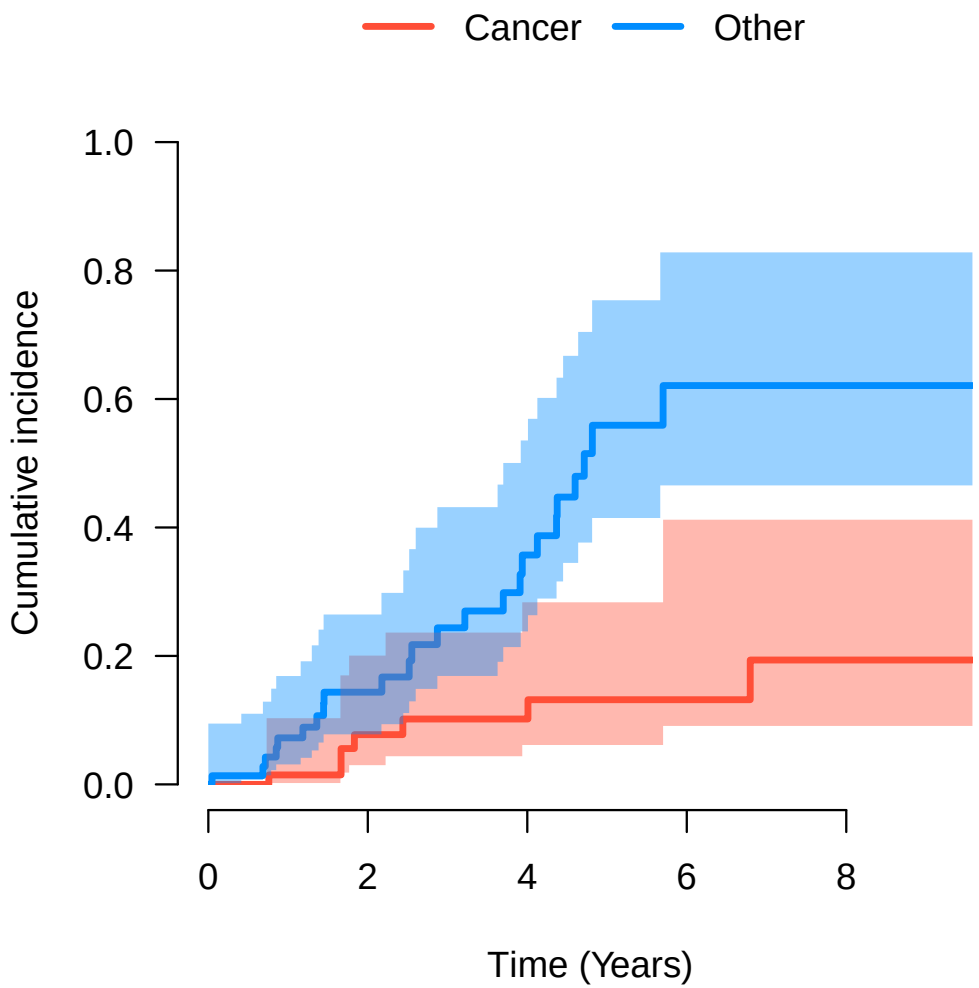
More on R

- The usual R functions for working with `survfit` objects also work with competing risks, such as `summary` for obtaining estimates at specific time points and `plot` for plotting the curves
- It is worth noting that the `survival` package refers to the cumulative incidence in each category as the "state probability"
- In particular, `fit$pstate` contains the cumulative incidence function estimates, with `fit$lower` and `fit$upper` containing the confidence interval endpoints; all three quantities are matrices now, with one column for each competing risk as well as one for the category of "alive"

Results


Cancer
Other
Total
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative incidence
0
2
4
6
8
Time (Years)

Results w/ confidence band



Comments

- Although the idea of a subdistribution may seem foreign at first, its ease of interpretation is one big reason they are widely estimated and analyzed
- For example, in the endometrial cancer data, they allow us to estimate that after 5 years, 13% of patients will have died due to cancer, 56% due to other causes, and the remaining 31% will still be alive
- This same goes for regression modeling approaches (we will discuss these later), which allow us to incorporate more specific information for a given subject in making those predictions

Conditional recurrence

- Cumulative incidence functions are not always the only quantities of interest when multiple time-to-event endpoints are present
- For example, one quantity of interest in the endometrial cancer study is the percent of surviving patients who have experienced recurrence by a given time
- This quantity was originally studied by Pepe (1991), "Inference for Events with Dependent Risks in Multiple Endpoint Studies", who referred to it as the *conditional prevalence*

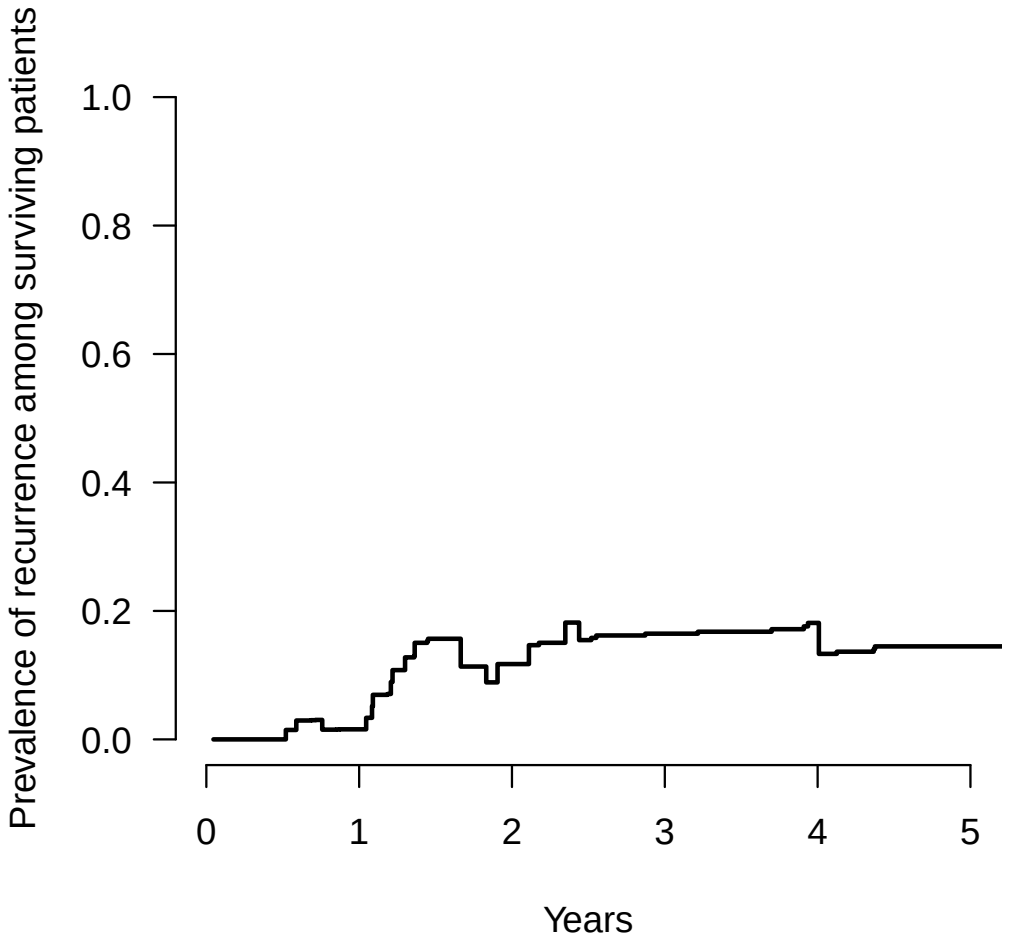
Estimate

- Interestingly, the conditional prevalence can be consistently estimated by a combination of simple Kaplan-Meier estimates
- Letting $\hat{S}_{\text{OS}}$ denote the estimated survival function with respect to overall survival (i.e., all-cause mortality) and $\hat{S}_{\text{PFS}}$ denote the estimated progression-free survival, the estimated conditional recurrence is given by

$$\text{RC}(t) = 1 - \frac{\hat{S}_{\text{PFS}}(t)}{\hat{S}_{\text{OS}}(t)}$$

- It is worth mentioning that unlike survival functions and cumulative incidence functions, the conditional prevalence is not necessarily monotone

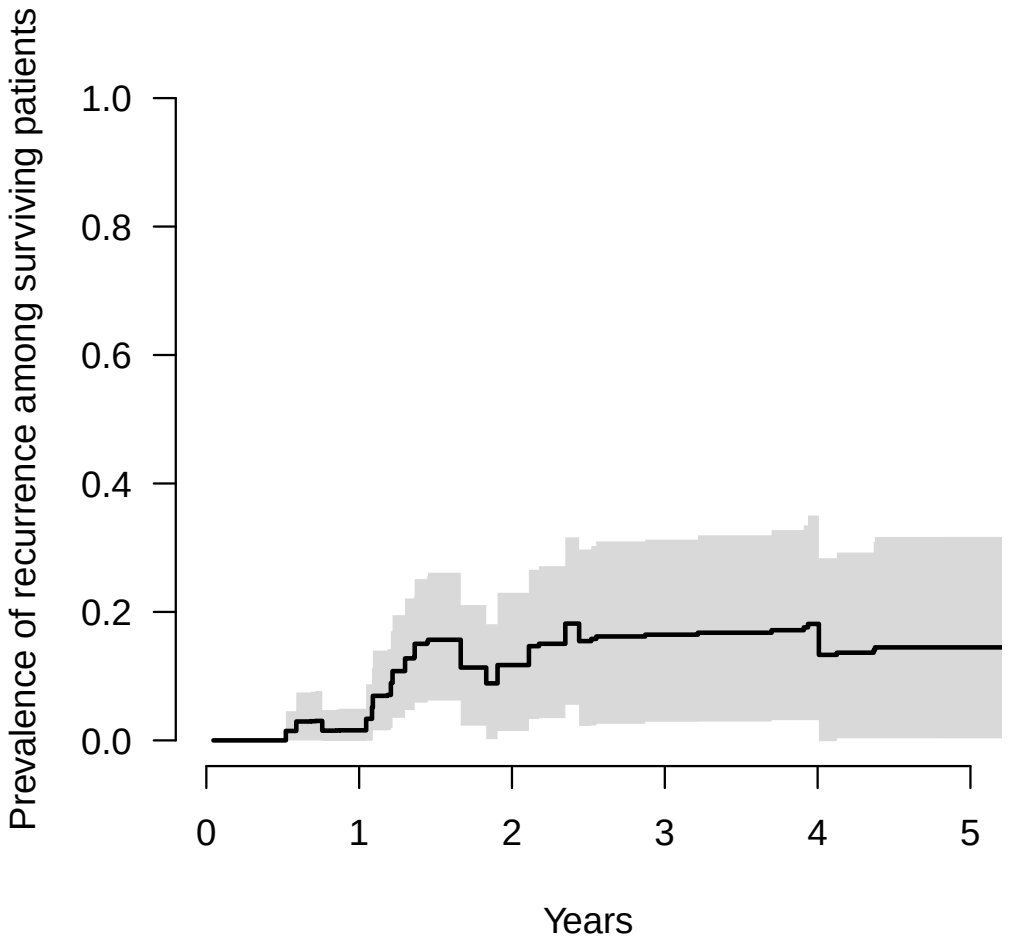
Conditional recurrence in endometrial cancer study



Confidence intervals

- Pepe (1991) provides some asymptotic derivations for the standard error of the conditional prevalence
- Another possibility is to use the bootstrap
- Here, we simply resample from the original group of subjects, calculate $\hat{S}_{\mathrm{OS}}(t)$, $\hat{S}_{\mathrm{PFS}}(t)$, and $\mathrm{RC}(t)$ based on the resampled subjects
- This is then repeated a large number of times (e.g., 1,000) and the 2.5th and 97.5th quantiles of the bootstrapped estimates form a confidence interval for $\mathrm{RC}(t)$
- This is known as the bootstrap percentile method; there are other ways of forming bootstrap confidence intervals as well

Results w/ confidence intervals



Remarks

- The cumulative incidence functions and the conditional recurrence plots both indicate that radiation therapy provides adequate management of cancer risk in these patients, with recurrence only occurring in approximately 16% of patients
- Furthermore, radiation therapy is likely preferable to more aggressive interventions, as the risk of death from other causes is the greater medical concern here, roughly 3 times higher than the risk of death due to cancer recurrence

Cause-specific Cox models

- A natural next question, which we will only briefly touch upon, is how to incorporate covariates into the explanation and prediction of competing risks
- Cause-specific hazards, $\lambda_j(t)$, can be estimated with separate, ordinary Cox regression models by redefining the failure indicator to be 1 if and only if the patient fails due specifically to cause j
- This analysis approach is known as *cause-specific Cox modeling*

Fine-Gray competing risk regression

- An alternative approach, proposed in a landmark paper by Fine and Gray (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk", focuses on directly modeling the cumulative incidence, while retaining the semiparametric nature of Cox regression
- This approach is not available in the `survival` package, but is implemented in the package `cmprsk` through the function `crr`, for competing risks regression
- It is worth noting that `crr` does not offer a formula interface (you have to specify the design matrix $\mathbf{X}$ directly); a wrapper to `crr` with a formula interface, `FGR()` is available in the `riskRegression` package, which also offers a wrapper, `CSC()`, for cause-specific Cox models

Comparing the two frameworks

- The key conceptual difference between the two is that for CSC models, an individual is removed from the risk set when a failure due to other causes occurs, whereas in the Fine-Gray model, that individual remains in the risk set
- In the Fine-Gray model, it is important to be aware that if a predictor increases the risk of failure A and has no effect on failure B, the predictor will have a negative coefficient for outcome B, not zero
- On the other hand, if a predictor increases the risk of all types of failure equally, the coefficient will be positive in the CSC model and may be zero in the Fine-Gray model, since this predictor doesn't affect the balance of risk across causes
- Both models have their uses – just be careful that you use the one that best reflects the scientific goals of the study