tmerge

A Tool to Facilitate Creation of Multiple Time-Based Intervals per Subject

Cynthia Crowson, Terry Therneau and Elizabeth Atkinson

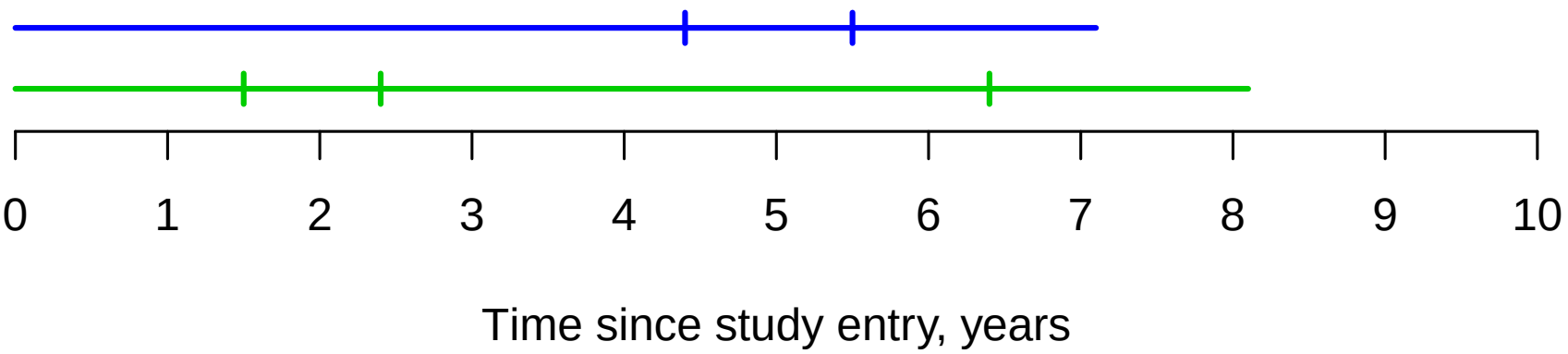
SDSS

May 18, 2018

MAYO CLINIC

Time-to-event analysis

- Often requires creation of multiple start/stop intervals per subject
  - time-dependent covariates
  - multiple outcomes per subject
  - multi-state models
- Deceptively simple task, easy to do incorrectly



MAYO CLINIC

Simple Example

- starting dataset has 1 observation per subject
- surgery is a time-dependent covariate

	id	age	tm_surg	futime	event
1	1	40	5	10	0
2	2	20	8	20	1
3	3	50	NA	30	1

- need to separate time periods before and after surgery

	id	age	tstart	tstop	death	surgery
1	1	40	0	5	0	0
2	1	40	5	10	0	1
3	2	20	0	8	0	0
4	2	20	8	20	1	1
5	3	50	0	30	1	0

MAYO CLINIC

`tmerge` function

- `tmerge` function in `survival` package in R makes this task easier
- sequential insertion
  - build the dataset one covariate or endpoint at a time
  - each addition will be "slipped in" to the original data in the same way that one would slide a new card into an existing deck of cards

MAYO CLINIC

tmerge function

- The basic form of the function call is

```
newdata <- tmerge(data1, data2, id,
                  newvar=tdc(time, value), ...)
```

- primary arguments:
  - data1: baseline data to be retained in the analysis dataset
  - data2: source for new data including events and time-dependent covariates
  - id: subject identifier used to merge the data together
  - ...: additional arguments that add variables to the dataset
  - tstart, tstop: used to set the time range for each subject
  - options

MAYO CLINIC

tmerge function

- The key part of the call are the "..." arguments, which each can be one of 4 types:
  - tdc() and cumtdc() add a time-dependent covariate
  - event() and cumevent() add a new endpoint
- resulting dataset has 3 new variables (at least):
  - `id`: identifier indicating which rows belong to the same subject
  - `tstart`: start of the interval
  - `tstop`: end of the interval

MAYO CLINIC

Example

- dataset: d1

```
  id age
1  1  40
2  2  20
3  3  50
```

- dataset: d2

```
  id tm_surg futime event
1  1       5     10     0
2  2       8     20     1
3  3      NA     30     1
```

MAYO CLINIC

Example: step 1 - create start/stop time

```
step1 <- tmerge(data1=d1, data2=d2, id=id,
                death=event(futime, event))
step1
```

```
  id age tstart tstop death
1  1  40      0    10     0
2  2  20      0    20     1
3  3  50      0    30     1
```

MAYO CLINIC

Example: step 2 - create time-dependent covariate

```
step2 <- tmerge(data1=step1, data2=d2,id=id,
                surgery=tdc(tm_surg))
step2
```

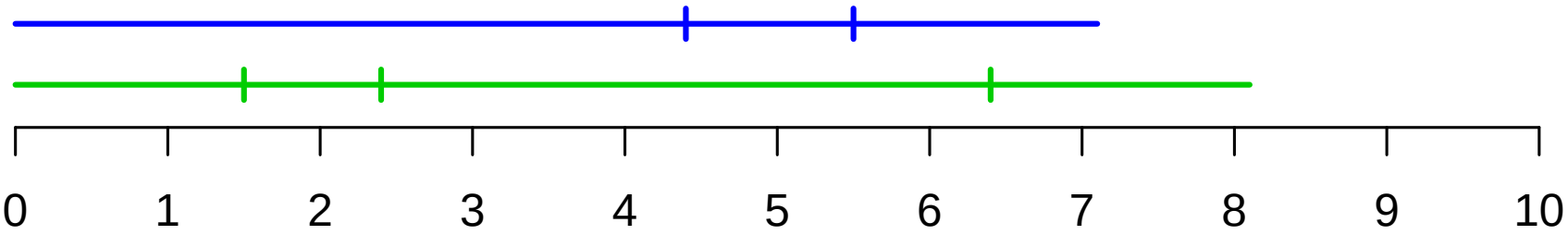
```
  id age tstart tstop death surgery
1  1  40      0     5     0       0
2  1  40      5    10     0       1
3  2  20      0     8     0       0
4  2  20      8    20     1       1
5  3  50      0    30     1       0
```

Note: this can also be done in just one step:

```
tmerge(data1=d1, data2=d2, id=id,
       death=event(futime, event),
       surgery=tdc(tm_surg))
```

How covariates differ from events

- time-dependent covariates
  - apply from the *start* of a new interval
  - persist for all remaining intervals unless subsequently changed
- events
  - occur at the *end* of an interval
  - only occur once
  - in time-to-event analyses, time intervals are open on the left and closed on the right, i.e., (tstart, tstop].



MAYO CLINIC

Example: Multiple Events

- Here is the data for the first 4 subjects

	id	treat	sex	age	futime	etime1	etime2	etime3
1	1	1	2	12	414	219	373	NA
2	2	0	1	15	439	8	26	152
3	3	1	1	19	382	NA	NA	NA
4	4	1	1	12	388	NA	NA	NA

MAYO CLINIC

Example: Multiple Events

```
dim(cgd0)
```

```
[1] 128  20
```

```
newcgd <- tmerge(data1=cgd0[, 1:13], data2=cgd0,
                 id=id, tstop=futime)
newcgd <- tmerge(newcgd, cgd0, id=id,
                 infect = event(etime1))
newcgd <- tmerge(newcgd, cgd0, id=id,
                 infect = event(etime2))
newcgd <- tmerge(newcgd, cgd0, id=id,
                 infect = event(etime3))
dim(newcgd)
```

```
[1] 196  16
```

MAYO CLINIC

Example: Multiple Events

	id	treat	sex	age	futime	tstart	tstop	infect
1	1	1	2	12	414	0	219	1
2	1	1	2	12	414	219	373	1
3	1	1	2	12	414	373	414	0
4	2	0	1	15	439	0	8	1
5	2	0	1	15	439	8	26	1
6	2	0	1	15	439	26	152	1
7	2	0	1	15	439	152	439	0

MAYO CLINIC

tcount attribute

```
attr(newcgd, "tcount")
```

```
        early late gap within boundary lead trail tied
infect      0    0   0     44        0    0     0    0
infect      0    0   0     16        0    0     1    0
infect      0    0   0      8        0    0     0    0
```

MAYO CLINIC

Example: Continuous values that change over time

- pbc data set contains baseline data and follow-up status for 312 subjects with primary biliary cirrhosis (one obs per person)
- pbcseq data set contains 1945 repeated laboratory values

```
pbc2 <- tmerge(pbc, pbc, id=id,
               death = event(time, status)) #set range
pbc2 <- tmerge(pbc2, pbcseq, id=id,
               ascites = tdc(day, ascites),
               bili = tdc(day, bili),
               albumin = tdc(day, albumin))
```

	id	tstart	tstop	death	ascites	bili	albumin
1	1	0	192	0	1	14.5	2.60
2	1	192	400	2	1	21.3	2.94
3	2	0	182	0	0	1.1	4.14
4	2	182	365	0	0	0.8	3.60
5	2	365	768	0	0	1.0	3.55

MAYO CLINIC

Example: Continuous values that change over time

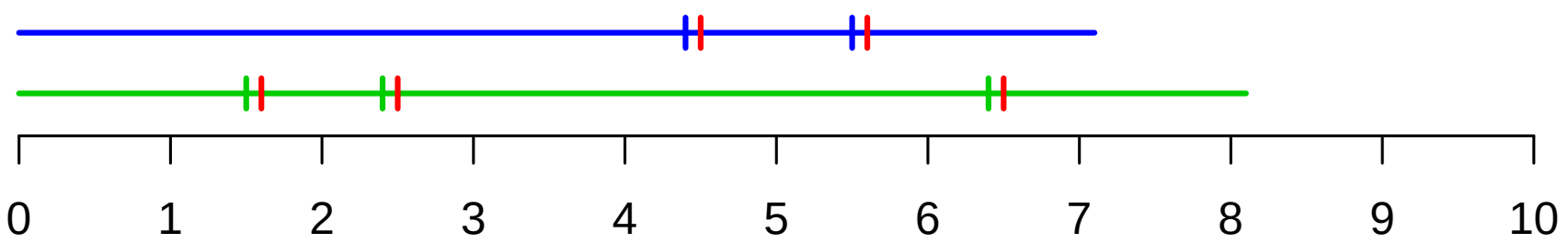
	early	late	gap	within	boundary	lead	trail	tied
death	0	0	0	0	0	0	312	0
ascites	0	131	0	1442	0	312	0	0
bili	0	138	0	53	1442	312	0	0
albumin	0	138	0	0	1495	312	0	0

- Missing values in time or value from `data2` are ignored
  - Consequence: "last value carried forward"
- Default can be changed by adding `options=list(na.rm=FALSE)` to the second call
  - Any `tdc` calls with a missing time are still ignored, independent of the na.rm value, since we would not know where to insert them.

MAYO CLINIC

Time delay

- For any data set containing constructed time-dependent covariates, it is a good idea to re-run the analyses after adding a 7-14 day lag to key variables.
  - One reason is to check for cases of reverse causality. A covariate measured soon before death may not be a predictor of death but rather is simply a marker for an event that is already in progress.
  - Even more subtle biases can occur via coding errors.
- When the results show a substantial change, understanding why this occurred is an critical step.

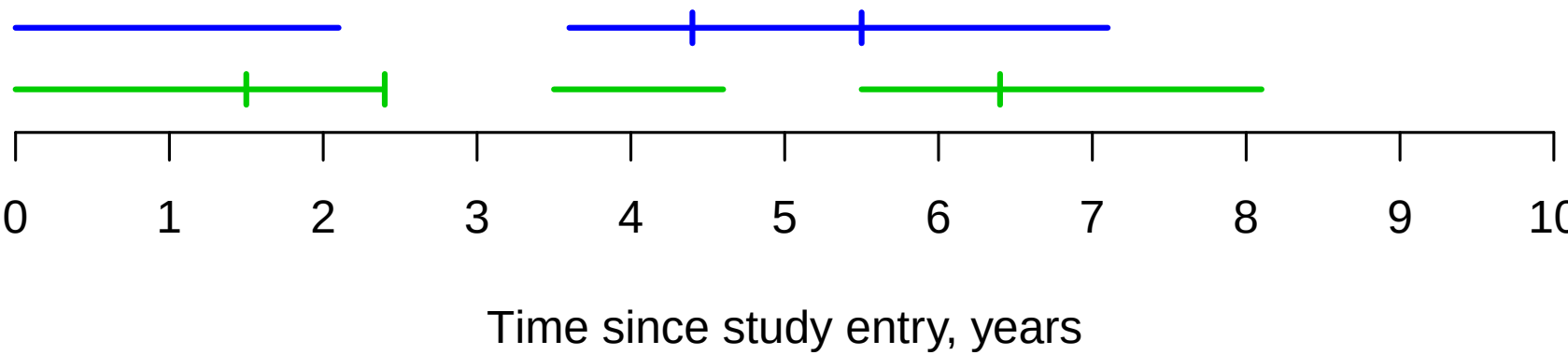


MAYO CLINIC

```
pbc2a <- tmerge(pbc2, pbcseq, id=id,
```

Gaps

- Time dependent covariates that occur before the start of a subject's follow-up interval or during a gap in time do not generate a new interval split, but they do set the value of that covariate for future times.
  - Rationale: during a subject's time within the county we would like the variable "prior diagnosis of diabetes" to be accurate, even if that diagnosis occurred during a non-resident period
- Events that occur in a gap are not counted.
  - For events outside of the timeline, we have no way to know who the appropriate comparison group is, and so must ignore those events.



MAYO CLINIC

Summary

- `tmerge` is a simple to use, flexible tool to create multiple start/stop intervals per subject
  - time-dependent covariates - both binary and continuous
  - multiple outcomes per subject
  - allows for gaps in time
- data checks can help avoid errors
  - `tcount` attribute
  - use of **delay** for time-dependent covariates

MAYO CLINIC