Longitudinal Cohort data in epidemiology

Survival/time-to-event data

ERHS 732

Longitudinal Cohort data

- We have been using time-series data were day (or week) has been the unit of observation

```
> head(obs)
# A tibble: 6 x 14
  date        year month   day   doy dow     all
  <date>     <dbl> <dbl> <dbl> <dbl> <ord> <dbl>
1 1990-01-01  1990     1     1     1 Mon     220
2 1990-01-02  1990     1     2     2 Tue     257
3 1990-01-03  1990     1     3     3 Wed     245
4 1990-01-04  1990     1     4     4 Thu     226
5 1990-01-05  1990     1     5     5 Fri     236
6 1990-01-06  1990     1     6     6 Sat     235
>
```

Longitudinal Cohort data

- We now shift to the longitudinal cohort design were the person (or person-time) is the unit of observation

```
> head(fhs)
# A tibble: 6 x 39
  RANDID   SEX TOTCHOL   AGE SYSBP DIABP CURSMOKE CIGPDAY   BMI DIABETES
   <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl> <dbl>    <dbl>
1   2448     1     195    39  106   70          0       0  27.0        0
2   2448     1     209    52  121   66          0       0  NA          0
3   6238     2     250    46  121   81          0       0  28.7        0
4   6238     2     260    52  105   69.5        0       0  29.4        0
5   6238     2     237    58  108   66          0       0  28.5        0
6   9428     1     245    48  128.  80          1      20  25.3        0
# … with 23 more variables: PREVMI <dbl>, PREVSTRK <dbl>, PREVHYP <dbl>,
#   LDLC <dbl>, DEATH <dbl>, ANGINA <dbl>, HOSPMI <dbl>, MI_FCHD <dbl>, A
#   HYPERTEN <dbl>, TIMEAP <dbl>, TIMEMI <dbl>, TIMEMIFC <dbl>, TIMECHD <
#   TIMEDTH <dbl>, TIMEHYP <dbl>
> |
```

Individual level data

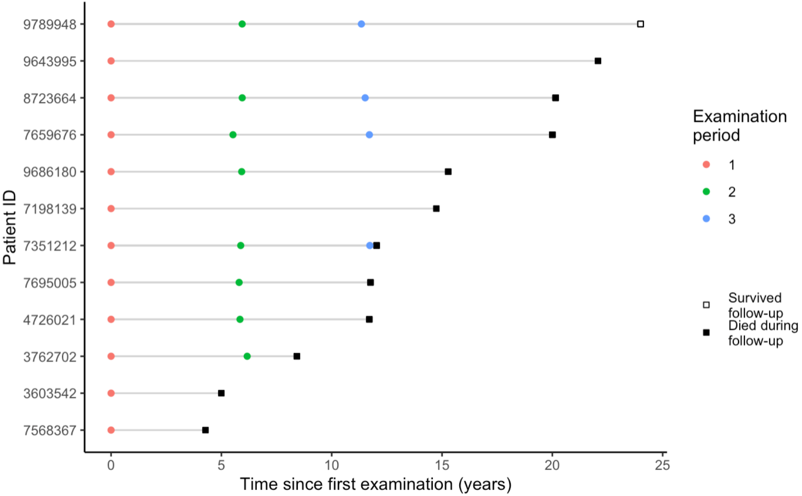
- This also shifts us from an ecological design, to an individual level with exposure, covariates (and outcomes) varying from participant to participant
- The longitudinal design also allows for change in these variables *within* individual over time
  - Repeated outcomes (mixed models (11/6)
  - Time-varying exposures (10/23 & 10/30)
  - Time-varying covariates (10/23 & 10/30)

Survival or time-to-event outcomes

- We will start with examining survival (time-to-event) outcomes
- Most simple examples are analysis of mortality outcomes where time-to-event is actually survival time, however the term is used when characterizing any time-to-event outcome (cancer incidence, CVD, birth etc)
- Measures of association in this framework are usually based on the risk and hazard (thought survival time ratios can also be estimated)

Survival or time-to-event outcomes

- The outcome is a combination of an event **and** the time the event happened



Survival or Right censored data

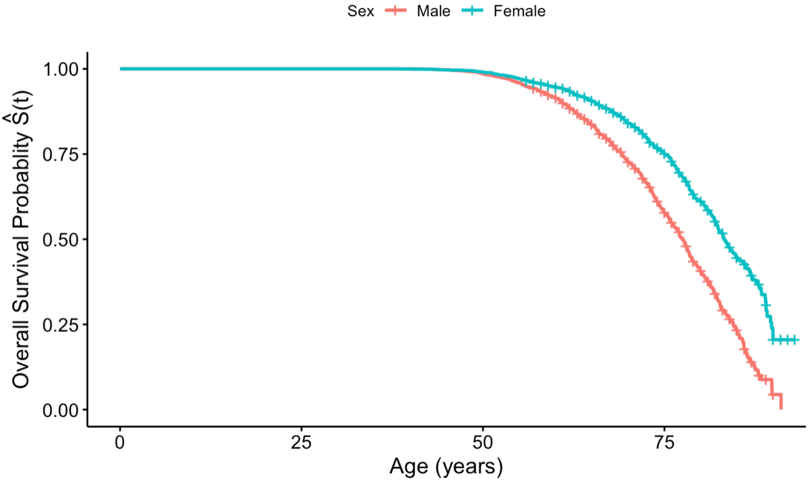
- Survival data are also called right censored data as participants (or person-time) is considered censored at some point
- A participant (or their person-time) will be censored
  - once they develop the outcome (for most survival outcomes the event can only occur once)
  - once we reach the end of follow-up (also called administrative end of follow-up) regardless of whether they have developed the outcome or not
  - if they experience an event that prevents us from assessing the outcome in the future (loss to follow-up, competing event)

Survival analysis

- We will examine survival (and hazards) over time as well as across groups
- A simple way to characterize survival are survival curves
- Modeling hazards typically relies on the Cox proportional hazards model (yielding Hazard Ratios comparing levels of exposure

Kaplan-Meier curves

- Survival curves (like the Kaplan-Meier) are very simple, but informative ways to portray survival data



Survival, Hazard and Risk

- Survival $S(t)$ is simply the proportion of people that have survived (or not had the outcome) by time t
- Risk $R(t)$ is the additive inverse of survival $(1 - S(t))$
  - $R(t)$ is the cumulative incidence at time t
- The (instantaneous) hazard is the probability of developing the outcome in a short interval of time (for example between $t - 1$ and t) among those still at risk (those that haven't developed the outcome by $t - 1$)

Cox Proportional Hazards Model

- Introduced in 1972 by statistician David Cox
- Advantageous in its simplicity
  - No distributional assumptions required
  - Main assumption is that of the proportional hazards (requires that covariates are multiplicatively related to the hazard and the hazards across levels of covariates remain proportional over time)
- Under a simple Cox model for a study with p covariates $X_i = (X_{i1}, \cdots, X_{ip})$ for each participant i then the hazard of an outcome of interest as a function of time is

$$\lambda(t|X_i) = \lambda_0(t) exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})$$

Framingham Heart Study (FHS)

- One of the most famous cohort studies
- Begun in 1948 and continues today following second generation (offspring of original cohort) and a third generation of participants (We will be using the original cohort data from the first generation of participants)
- Instrumental in identifying risk factors for cardiovascular disease ranging from smoking to blood pressure, cholesterol, BMI etc