Flexible parametric inference for cause-specific hazard competing risks models under interval censoring

Michael J. Crowther

Associate Professor of Biostatistics
Biostatistics Research Group
Department of Health Sciences
University of Leicester
michael.crowther@le.ac.uk
@Crowther_MJ

ISCB Leuven - 15th July 2019

Background I

- Interval censoring is a common, yet often overlooked, problem in survival analysis
- Patients are observed at intermittent time points, and the event of interest occurs at some unknown timepoint between the observation points
- Occurs extensively in dentistry and cancer screening - both areas where conditions are only detected at a scheduled appointment
- Ignoring it can lead to bias - particularly in measures of absolute risk

Background II

- Jackson (2011) developed the widely used `msm` package in R, designed for longitudinal data, including interval censoring, where transition rates can be exponential or piecewise exponential
- Hudgens et al. (2014) took the Fine and Gray (1999) approach, using the Gompertz distribution, building on work by Jeong and Fine (2006)
- Li (2016) developed cause-specific proportional hazard modelling with interval censoring, using penalised splines to model the baseline hazards
  - assumed all events were interval censored
  - assumed proportionality
  - nothing on prediction
- Machado and van den Hout (2018) developed flexible approaches for multi-state models with interval censoring, based on splitting the timescale into intervals & approximating with splines

In this talk

Building on Li (2016), we provide a number of methodological developments:

- develop a full likelihood approach to handling competing risks with interval censoring, specified with cause-specific hazard models
  - Maintains ease of interpretation of hazard ratios
  - Can still easily calculate CIFs
- Provide a range of standard and flexible parametric models
- Extend to allow for non-linear effects, e.g. non-proportional hazards
- Derive clinically useful predictions
  - CIF, RMST, RMFT, contrasts, and associated confidence intervals
- Provide a user-friendly software package

Competing risks

Let T be the time to event of any of K competing causes, where $k = 1, \ldots, K$, and D denote the event type, where $D = 1, \ldots, K$. For exactly observed competing risks data, subject to right censoring, the ith likelihood contribution is,

$$L_i = \prod_{k=1}^{K} \left[h_k(t_i) S_k(t_i)\right]^{\delta_{ik}} S_k(t_i)^{1-\delta_{ik}}$$

where δ_{ik} is a cause-specific event indicator, taking the value of 1 if $D = k$ and 0 otherwise.

Competing risks

- Due to the factorisation of the likelihood, competing risks with exactly observed event data can be fitted in standard software
- Each cause-specific model can be fitted independently, censoring competing events.
- The same cannot be said when interval censoring is present, which raises a more technically challenging problem

Interval censored competing risks

- We know which particular event occurred, but we do not know exactly when
- Across the time domain that the event has occurred, we must still account for the fact that a patient is at risk of competing events, until said event has occurred
- For the ith patient, we define $\{t_{Li},\ t_{Ri}\}$, where $t_{Li} < t_{Ri}$, to be the limits of the interval during which the event occurred
- Therefore, for an interval censored observation, the ith contribution to the likelihood is,

$$L_i = \int_{t_{Li}}^{t_{Ri}} \prod_{k=1}^{K} f_k(u)^{\delta_{ik}} S_k(u)^{1-\delta_{ik}} \, \mathrm{d}u$$

Interval censored competing risks

$$L_i = \int_{t_{Li}}^{t_{Ri}} \prod_{k=1}^{K} f_k(u)^{\delta_{ik}} S_k(u)^{1-\delta_{ik}} \, \mathrm{d}u$$

- The challenge lies with the integral, as it means we can't fit separate models for each cause - we must estimate them jointly
- We could assume the event occurs at the beginning, end or midpoint
- We could assume competing events are censored at the beginning
- We could do things properly...

Interval censored competing risks

- To maintain generality, and flexibility in modelling frameworks, we propose to utilise numerical integration techniques to incorporate the full contribution to the log-likelihood
- Therefore we have,

$$L_i = \frac{t_{Ri} - t_{Li}}{2} \sum_{m=1}^{M} \left[w_m \prod_{k=1}^{K} f_k \left(\frac{t_{Ri} - t_{Li}}{2} x_m + \frac{t_{Li} + t_{Ri}}{2} \right)^{\delta_{ik}} \times S_k \left(\frac{t_{Ri} - t_{Li}}{2} x_m + \frac{t_{Li} + t_{Ri}}{2} \right)^{1-\delta_{ik}} \right]$$

- for example using Gauss-Legendre quadrature

Predictions

- The cause-specific CIF at time t, $F_k(t)$, is defined as,

$$F_k(t) = \int_0^t h_k(u) \left[\prod_{k=1}^{K} S_k(u) \right] \mathrm{d}u$$

- Restricted mean survival time from our competing risks model as,

$$RMST(t) = t - \int_0^t \sum_{k=1}^{K} F_k(u) \mathrm{d}u$$

- Restricted mean failure time, or total time lost due to any event, is defined as,

$$RMFT(t) = \int_0^t \sum_{k=1}^{K} F_k(u) \mathrm{d}u$$

- Confidence intervals obtained using the delta method

Application to prostate cancer

- We have 506 patients diagnosed with prostate cancer, who are randomly assigned a treatment of diethylstilbestrol or a placebo
- Time of death is measured to the nearest month, and whether patients died from cancer (n=155), cardiovascular disease (CVD) (n=141), or other causes (n=60), with 150 patients still alive at the end of follow-up
- I'll illustrate how to fit the models with the `merlin` package in `Stata` (Crowther, 2019)

Application to causes of death in prostate cancer

We begin by loading the prostate cancer dataset

```
. use "prostatecancer.dta",clear
```

List the data for the first 5 obs.

```
. list id time status trt if _n in 1/5, noobs
  +--------------------------+
  | id   time   status   trt |
  |--------------------------|
  |  1     72   Censor     0 |
  |  2      1   Cancer     0 |
  |  3     40      CVD     1 |
  |  4     20      CVD     0 |
  |  5     65   Censor     0 |
  +--------------------------+
```

Application to causes of death in prostate cancer

As `merlin` supports combinations of right censored (0), exactly observed events (1), and interval censoring (2), we must distinguish between such types

Dataset contains only interval censored events

```
. gen event = 0
. replace event = 2 if status > 0
(356 real changes made)
```

We need cause-specific model indicators

```
. gen cause1 = (status == 1 | status == 0)
. gen cause2 = (status == 2 | status == 0)
. gen cause3 = (status == 3 | status == 0)
```

Application to causes of death in prostate cancer

- Next we must define the limits of the intervals
- Event times were recorded to the nearest month, so if an event is recorded as 10 months, it may have occurred at some time between 9 and 10 months

```
. gen left = time - 1 if event == 2
(150 missing values generated)
```

Application to causes of death in prostate cancer

Our dataset now looks like this,

We need cause-specific model indicators

```
. list id left time event status trt if _n in 1/5, noobs
  +-------------------------------------------+
  | id   left   time   event   status   trt |
  |-------------------------------------------|
  |  1      .     72       0   Censor     0 |
  |  2      0      1       2   Cancer     0 |
  |  3     39     40       2      CVD     1 |
  |  4     19     20       2      CVD     0 |
  |  5      .     65       0   Censor     0 |
  +-------------------------------------------+
```

Application to causes of death in prostate cancer

We need a transition matrix

```
. matrix define tmat = (.,1,2,3\.,.,.,.\.,.,.,.\.,.,.,.)
. matrix rownames tmat = alive dead_cancer dead_cvd dead_other
. matrix colnames tmat = alive dead_cancer dead_cvd dead_other
. matrix list tmat
tmat[4,4]
                alive  dead_cancer     dead_cvd   dead_other
      alive         .             1            2            3
dead_cancer         .             .            .            .
   dead_cvd         .             .            .            .
 dead_other         .             .            .            .
```

Application to causes of death in prostate cancer

We can now specify our model

```
. merlin (time trt if cause1==1, family(rp, df(3) fail(event) linterval(left)))
>        (time trt if cause2==1, family(rp, df(3) fail(event) linterval(left)))
>        (time trt if cause3==1, family(rp, df(3) fail(event) linterval(left)))
>        , transmatrix(tmat)

Mixed effects regression model                  Number of obs     =        506
Log likelihood = -2117.5631
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
time:        |
         trt |  -.3972101   .1631877    -2.43   0.015    -.7170521   -.0773681
       _cons |   -1.08753   .1075837   -10.11   0.000     -1.29839   -.8766694
-------------+----------------------------------------------------------------
time:        |
         trt |   .1717434   .1694448     1.01   0.311    -.1603622     .503849
       _cons |  -1.441273   .1280155   -11.26   0.000    -1.692179   -1.190367
-------------+----------------------------------------------------------------
time:        |
         trt |  -.4559505   .2635969    -1.73   0.084     -.972591    .0606901
       _cons |  -1.707384   .1675541   -10.19   0.000    -2.035784   -1.378984
------------------------------------------------------------------------------
    Warning: Baseline spline coefficients not shown - use ml display
```

Application to causes of death in prostate cancer

Transition-specific distributions

```
. merlin (time trt if cause1==1, family(weibull,  fail(event) linterval(left)))
>        (time trt if cause2==1, family(rp, df(3) fail(event) linterval(left)))
>        (time trt if cause3==1, family(gompertz, fail(event) linterval(left)))
>        , transmatrix(tmat)
```

Time-dependent effects

```
. merlin (time trt trt#rcs(time, df(1)) if cause1==1,   ///
>                  family(rp, df(3) fail(event) linterval(left)))
>        (time trt trt#fp(time, pow(1 1)) if cause2==1, ///
>                  family(rp, df(3) fail(event) linterval(left)))
>        (time trt if cause3==1,                        ///
>                  family(rp, df(3) fail(event) linterval(left)))
>        , transmatrix(tmat)
```

The three models can be specified as simply or as complex as necessary

Application to causes of death in prostate cancer

We can predict the cause-specific cumulative incidence functions

```
range tvar 0 60 500
predict cif1, cif outcome(1) timevar(tvar) at(trt 1)
predict cif2, cif outcome(2) timevar(tvar) at(trt 1)
predict cif3, cif outcome(3) timevar(tvar) at(trt 1)
```

And quantify the impact of treatment

```
. predict cifdiff1, cifdifference outcome(1) timevar(tvar) ///
>                   at1(trt 1) at2(trt 0) ci
. predict cifdiff2, cifdifference outcome(2) timevar(tvar) ///
>                   at1(trt 1) at2(trt 0) ci
. predict cifdiff3, cifdifference outcome(3) timevar(tvar) ///
>                   at1(trt 1) at2(trt 0) ci
```

Application to causes of death in prostate cancer

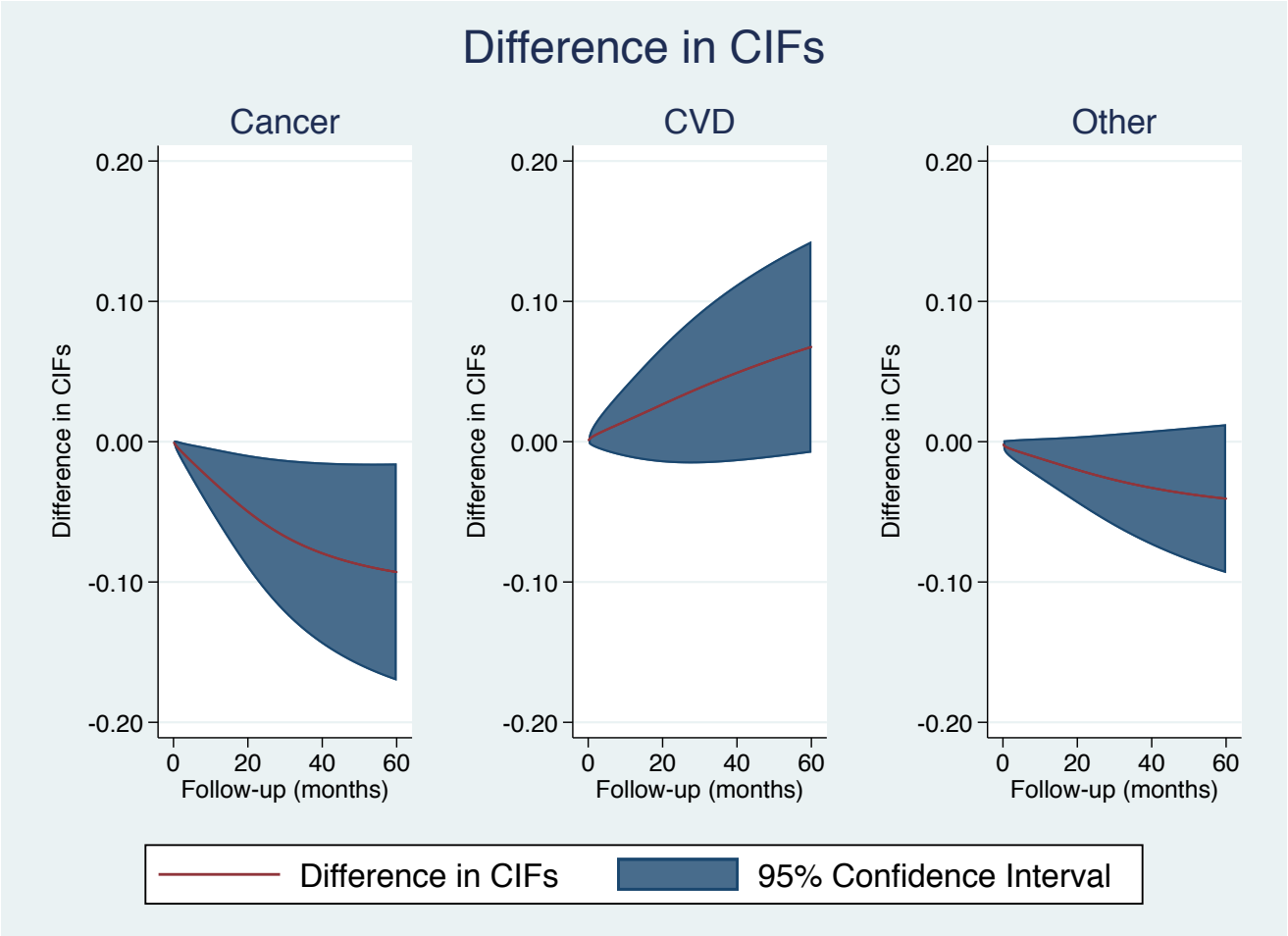

Figure 1: Difference in cumulative incidence functions, and associated 95% confidence interval, across treatment groups, for each cause of death.

Application to causes of death in prostate cancer

We can calculate the loss in (restricted) life expectancy for each cause

```
predict lile1, timelost outcome(1) timevar(tvar) at(trt 1) ci
predict lile2, timelost outcome(2) timevar(tvar) at(trt 1) ci
predict lile3, timelost outcome(3) timevar(tvar) at(trt 1) ci
```

Or directly calculate the total

```
predict lile0, totaltimelost outcome(1) timevar(tvar) at(trt 0) ci
predict lile1, totaltimelost outcome(1) timevar(tvar) at(trt 1) ci
```

Application to causes of death in prostate cancer

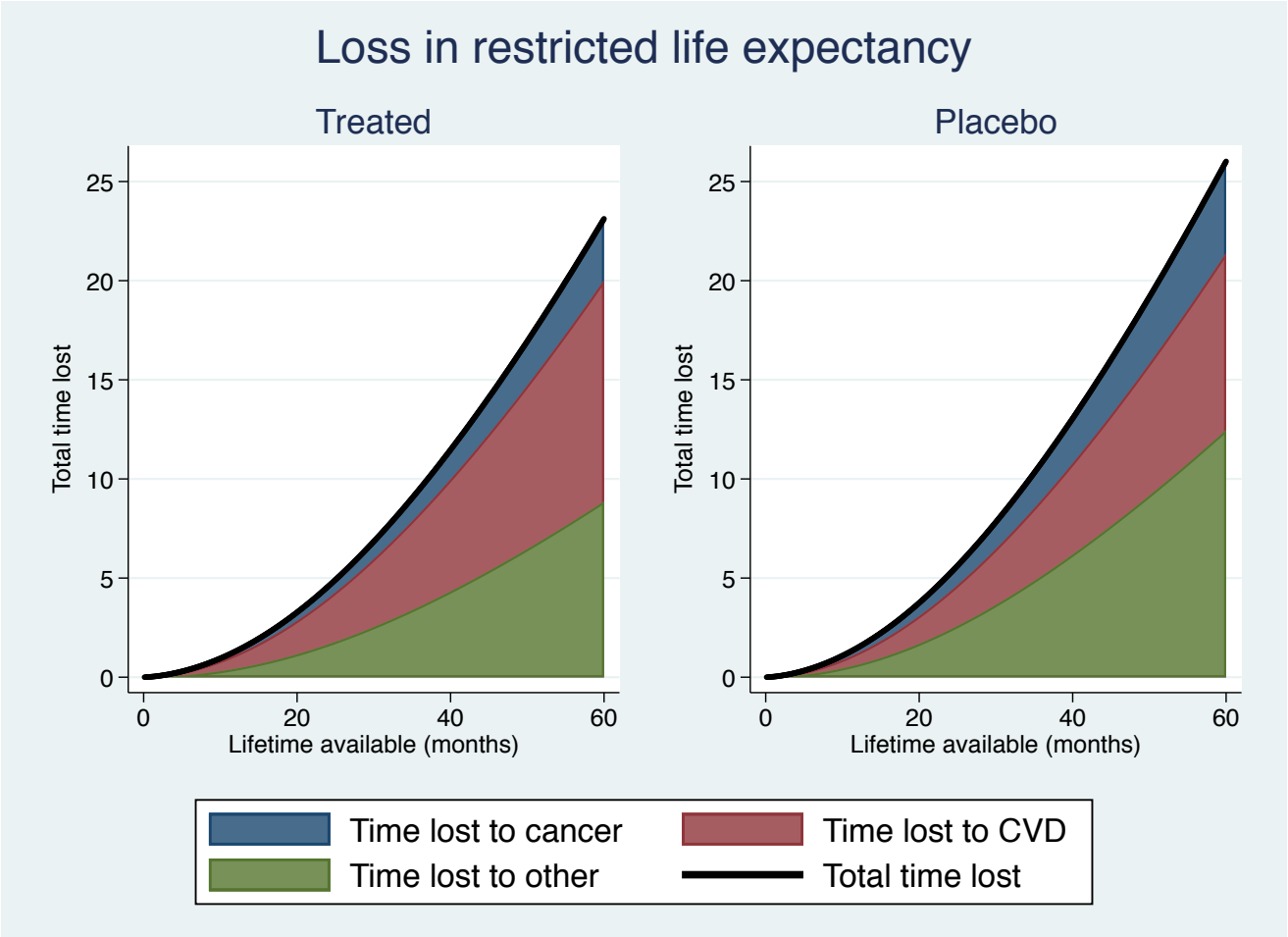

Figure 2: Loss in restricted life expectancy due to each cause of death.

Discussion and future work

- Interval censoring is often overlooked in competing risks
  - Technical challenges increase substantially
  - Limited software available
- Using standard numerical integration techniques, we develop a general and flexible, parametric framework to model cause-specific competing risks
  - Also allows for left truncation, exactly observed events, right censoring, which can vary across all cause-specific models
- We can still derive clinically useful predictions, regardless of the complexity of the underlying models
- Because it's been implemented in `merlin`, we get a lot of extensions for free...
- More examples on `mjcrowther.co.uk/software/merlin`

References

Crowther, M. J. 2019. merlin - a unified modelling framework for data analysis and methods development in Stata. *Stata Journal* . URL `https://arxiv.org/abs/1806.01615`.

Fine, J. P., and R. J. Gray. 1999. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94(446): 496–509.

Hudgens, M. G., C. Li, and J. P. Fine. 2014. Parametric likelihood inference for interval censored competing risks data. *Biometrics* 70: 1–9.

Jackson, C. H. 2011. Multi-state models for panel data: the msm package for R. *Journal of Statistical Software* 38(8): 1–29.

Jeong, J.-H., and J. Fine. 2006. Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55(2): 187–200. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2006.00532.x`.

Li, C. 2016. Cause-Specific Hazard Regression for Competing Risks Data Under Interval Censoring and Left Truncation. *Computational statistics & data analysis* 104: 197–208.

Machado, R. J. M., and A. van den Hout. 2018. Flexible multistate models for interval-censored data: Specification, estimation, and an application to ageing research. *Statistics in medicine* 37: 1636–1649.