The Devil Is In The Details... And The Data — Tutorial On Preparing Data for Multi-state Modelling

Enoch Yi-Tung Chen

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet

The 2022 Northern European Stata Conference
12 Oct 2022
https://github.com/enochytcchen/NordicStata2022
Illness-death model

Figure: Illness-death model
### Illness-death model

#### Example data from Crowther2017 (1)

```stata
. use http://fmwww.bc.edu/repec/bocode/m/multistate_example, clear
(Rotterdam breast cancer data, truncated at 10 years)
.
. // List one patient to see the variables
. list pid rf rfi os osi if pid == 1 | pid == 1371, sepby(pid) noobs
```

<table>
<thead>
<tr>
<th></th>
<th>pid</th>
<th>rf</th>
<th>rfi</th>
<th>os</th>
<th>osi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>59.1</td>
<td>0</td>
<td>59.1</td>
<td>alive</td>
</tr>
<tr>
<td></td>
<td>1371</td>
<td>16.6</td>
<td>1</td>
<td>24.3</td>
<td>deceased</td>
</tr>
</tbody>
</table>
Illness-death model

Transition matrix for Illness-death model

```stata
. matrix tmat = (.,1,2 
                .,,3 
. matrix colnames tmat = to:Health to:Relapse to:Death
. matrix rownames tmat = from:Health from:Relapse from:Death
. matrix list tmat

```

```
tmat[3,3]

<table>
<thead>
<tr>
<th></th>
<th>to:</th>
<th>to:</th>
<th>to:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Health</td>
<td>Relapse</td>
<td>Death</td>
</tr>
<tr>
<td>from:Health</td>
<td>.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>from:Relapse</td>
<td>.</td>
<td>.</td>
<td>3</td>
</tr>
<tr>
<td>from:Death</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
```

Figure: Illness-death model

Enoch Yi-Tung Chen (KI MEB) Preparing Data for Multi-state Modelling 12 Oct 2022
. mset, id(pid) states(rfi osi) times(rf os) transm(tmat)
. list pid rf rfi os osi _trans _start _stop if pid == 1 | pid ==1371, sepby(pid)

<table>
<thead>
<tr>
<th>pid</th>
<th>rf</th>
<th>rfi</th>
<th>os</th>
<th>osi</th>
<th>_trans</th>
<th>_start</th>
<th>_stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59.1</td>
<td>0</td>
<td>59.1</td>
<td>alive</td>
<td>1</td>
<td>0</td>
<td>59.104721</td>
</tr>
<tr>
<td>1</td>
<td>59.1</td>
<td>0</td>
<td>59.1</td>
<td>alive</td>
<td>2</td>
<td>0</td>
<td>59.104721</td>
</tr>
<tr>
<td>1371</td>
<td>16.6</td>
<td>1</td>
<td>24.3</td>
<td>deceased</td>
<td>1</td>
<td>0</td>
<td>16.558521</td>
</tr>
<tr>
<td>1371</td>
<td>16.6</td>
<td>1</td>
<td>24.3</td>
<td>deceased</td>
<td>2</td>
<td>0</td>
<td>16.558521</td>
</tr>
<tr>
<td>1371</td>
<td>16.6</td>
<td>1</td>
<td>24.3</td>
<td>deceased</td>
<td>3</td>
<td>16.558521</td>
<td>24.344969</td>
</tr>
</tbody>
</table>

![Illness-death model](image)

**Figure:** Illness-death model
Reversible illness-death model

Figure: Reversible illness-death model
Example data from Crowther2017

. use http://fmwww.bc.edu/repec/bocode/m/multistate_example, clear
(Rotterdam breast cancer data, truncated at 10 years)

. // Assume recovery indicator and recovery time
. set seed 12345

. // Recovery indicator
. gen rei = cond(runiform() < 0.5, 0, 1) if rfi == 1 & rf!== os
(1,464 missing values generated)

. // Recovery
. gen re = runiform(rf, os) if rei == 1
(2,243 missing values generated)

. save multistate_example_temp.dta, replace

. // List one patient to see the variables
. list pid rf rfi re rei os osi if pid == 2778, sepby(pid) noobs

+-------------------------------------+
<table>
<thead>
<tr>
<th>pid   rf   rfi    re   rei     os   osi</th>
</tr>
</thead>
<tbody>
<tr>
<td>2778   40.3    1  53.36185    1   114.0   alive</td>
</tr>
</tbody>
</table>
+-------------------------------------+
Transition matrix for reversible illness-death model

```
. matrix rtmat = (.,1,2\ 3,.,4\,.,.)
. matrix colnames rtmat = to:Health to:Relapse to:Death
. matrix rownames rtmat = from:Health from:Relapse from:Death
. matrix list rtmat
```

```
rtmat[3,3]

<table>
<thead>
<tr>
<th></th>
<th>to:</th>
<th>to:</th>
<th>to:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Health</td>
<td>Relapse</td>
<td>Death</td>
</tr>
<tr>
<td>from:Health</td>
<td>.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>from:Relapse</td>
<td>3</td>
<td>.</td>
<td>4</td>
</tr>
<tr>
<td>from:Death</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
```

**Figure:** Reversible illness-death model
Common mistake 1

. `msset, id(pid) states(rfi osi rei) times(rf os re) transm(rtmat)`

All elements of the lower triangle of `transmatrix()` must be coded `missing = .`.

. `matrix list rtmat`

```
rtmat[3,3]

  to:  to:  to:  
    Health  Relapse  Death
  from:Health    .  1   2
  from:Relapse   3    .  4
  from:Death     .    .  .
```

Common mistake 2

. `msset, id(pid) states(rfi osi rei) times(rf os re)`
. `list pid rf rfi re rei os osi _trans _start _stop _status if pid == 2778`

```
+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+
<table>
<thead>
<tr>
<th>pid</th>
<th>rf</th>
<th>rfi</th>
<th>re</th>
<th>rei</th>
<th>os</th>
<th>osi</th>
<th>_trans</th>
<th>_start</th>
<th>_stop</th>
<th>_status</th>
</tr>
</thead>
<tbody>
<tr>
<td>11725.</td>
<td>2778.</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>1</td>
<td>0</td>
<td>40.279263</td>
</tr>
<tr>
<td>11726.</td>
<td>2778.</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>2</td>
<td>0</td>
<td>40.279263</td>
</tr>
<tr>
<td>11727.</td>
<td>2778.</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>3</td>
<td>0</td>
<td>40.279263</td>
</tr>
<tr>
<td>11728.</td>
<td>2778.</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>4</td>
<td>40.279263</td>
<td>53.361855</td>
</tr>
<tr>
<td>11729.</td>
<td>2778.</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>5</td>
<td>40.279263</td>
<td>53.361855</td>
</tr>
</tbody>
</table>
+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+--------------------------+
Set up a multi-state structured data

msset created the following variables

/*
 _from    float  %9.0g  Starting state
 _to      float  %9.0g  Receiving state
 _status  byte   %8.0g  Event (transition) indicator
 _start   double %10.0g Starting time for each transition
 _stop    double %10.0g Stopping time for each transition
 _flag    byte   %8.0g  Data modified
 _trans   float  %9.0g  Transition number
 _trans1  byte   %8.0g  _trans== 1.0000
 _trans2  byte   %8.0g  _trans== 2.0000
 _trans3  byte   %8.0g  _trans== 3.0000
*/

// Generate variables
gen _from = .
gen _to = .
gen _start = .
gen _stop = .
gen _status = .

Enoch Yi-Tung Chen (KI MEB)  Preparing Data for Multi-state Modelling  12 Oct 2022  10 / 29
Set up a multi-state structured data

Target: make wide-format data into long-format

. // List one patient to see the variables
. list pid rf rfi re rei os osi if pid == 2778 , sepby(pid) noobs
+--------------------------------------------------+
| pid   rf   rfi   re   rei   os   osi |
+--------------------------------------------------+
| 2778   40.3   1   53.36185   1   114.0 alive |
+--------------------------------------------------+
. matrix list rtm
rtmat[3,3]

| to: to: to: |
| Health Relapse Death |
| from:Health . 1 2 |
| from:Relapse 3 . 4 |
| from:Death . . . |

Figure: Reversible illness-death model
Set up a multi-state structured data

// Make 4 duplicates for each patient to define transitions
expand 4

// Manually make msset format
bysort pid: gen _trans = _n
// Generate _episode for potential recurrent events after recovery
gen _episode = 1
expand 2 if (_tran == 1 | _tran == 2) & rei == 1, gen(du)
replace _episode = 2 if du == 1
drop du

Figure: Reversible illness-death model
Set up a multi-state structured data

// Check the duplicates were done correctly
list pid rf rfi re rei os osi _trans _episode if pid == 2778 , sepby(pid) noobs

<table>
<thead>
<tr>
<th>pid</th>
<th>rf</th>
<th>rfi</th>
<th>re</th>
<th>rei</th>
<th>os</th>
<th>osi</th>
<th>_trans</th>
<th>_episode</th>
</tr>
</thead>
<tbody>
<tr>
<td>2778</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2778</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2778</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2778</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2778</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2778</td>
<td>40.3</td>
<td>1</td>
<td>53.36185</td>
<td>1</td>
<td>114.0</td>
<td>alive</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure: Reversible illness-death model
Set up a multi-state structured data

Specify _from_ _to

```
. matrix list rtmat

rtmat[3,3]

              to:     to:     to:  
    from:Health .       1       2
     from:Relapse  3       .       4
   from:Death    .       .       .

replace _from = 1 if _trans == 1 | _trans == 2
replace _from = 2 if _trans == 3 | _trans == 4

replace _to = 1 if _trans == 3
replace _to = 2 if _trans == 1
replace _to = 3 if _trans == 2 | _trans == 4
```
Set up a multi-state structured data

Specify _start _stop

local condition "(_trans == 1 | _trans == 2) & _episode == 1"
replace _start = 0 if ‘condition’  // T0 is 0
replace _stop = min(rf,os) if ‘condition’  // Relapse, death/censoring,
  // whichever happens first
replace _stop = 120 if _stop == . & ‘condition’  // Censor everyone after 120 mos

Figure: Reversible illness-death model
Set up a multi-state structured data

**Specify _start _stop**

```plaintext
local condition "(_trans == 3 | _trans == 4)"
replace _start = rf if ‘condition’    // T0 is time since relapse
replace _stop = min(re,os) if ‘condition’   // Recovery, death/censoring,
                     // whichever happens
replace _stop = 120 if _stop == . & ‘condition’ // Censor everyone after 120 mos
```

![Figure: Reversible illness-death model](image)

Enoch Yi-Tung Chen (KI MEB) Preparing Data for Multi-state Modelling 12 Oct 2022
Set up a multi-state structured data

**Specify _start _stop**

```plaintext
local condition "(_trans == 1 | _trans == 2) & _episode == 2"
replace _start = re if 'condition'  // T0 is time since recovery
replace _stop = os if 'condition'   // Death/censoring,
                   // whichever happens
replace _stop = 120 if _stop == . & 'condition'  // Censor everyone after 120 mos
```

**Figure: Reversible illness-death model**
Set up a multi-state structured data

Specify _status

// _trans == 1 & _episode == 1
replace _status = 1 if _trans == 1 & _episode == 1 & ///<
rfi == 1 & min(rf,os) == rf //
// Relapse as an event and happens first
replace _status = 0 if _trans == 1 & _episode == 1 & _status != 1

// _trans == 2 & _episode == 1
replace _status = 1 if _trans == 2 & _episode == 1 & ///<
osi == 1 & rfi == 0 //
// Death/censoring as an event and relapse never happens
replace _status = 0 if _trans == 2 & _episode == 1 & _status != 1

Figure: Reversible illness-death model
Specify _status

// _trans == 3
replace _status = 1 if _trans == 3 & rfi == 1 & /// Relapse has happened
  rei == 1 // Recovery as an event
replace _status = 0 if _trans == 3 & rfi == 1 & /// Relapse has happened
  _status != 1

// _trans == 4
replace _status = 1 if _trans == 4 & rfi == 1 & /// Relapse has happened
  osi == 1 & rei == 0 //
  // Death/censoring as an event and happens first
replace _status = 0 if _trans == 4 & rfi == 1 & /// Relapse has happened
  _status != 1

Figure: Reversible illness-death model
Specify _status

// _trans == 1 & _episode == 2
replace _status = 1 if _trans == 1 & _episode == 2 & ///
  rfi == 1 & rei == 1 & // Relapse and recovery have happened
  min(rf,re) == re // Recovery happens first, impossible here.
replace _status = 0 if _trans == 1 & _episode == 2 & ///
  rfi == 1 & rei == 1 & // Relapse and recovery have happened
  _status != 1
// _trans == 2 & _episode == 2
replace _status = 1 if _trans == 2 & _episode == 2 & ///
  rfi == 1 & rei == 1 & // Relapse and recovery have happened
  osi == 1 & min(re,os) == re // Death/censoring as an event and
replace _status = 0 if _trans == 2 & _episode == 2 & ///
  rfi == 1 & rei == 1 & _status != 1

Figure: Reversible illness-death model
Set up a multi-state structured data

Check

// List those who are not at risk in each transition
// There shouldn’t be any missing tho
// If there is, it means there’s something wrong

list pid _start _stop _from _to _status _trans ///
  if _start == . | _stop == . | _status == .

Figure: Reversible illness-death model
Set up a multi-state structured data

Check

```
. list pid rf rfi re rei os osi _trans _episode _start _stop _status if pid == 2846

     +---------------------------------------------+------+
<table>
<thead>
<tr>
<th>pid</th>
<th>rf</th>
<th>rfi</th>
<th>re</th>
<th>rei</th>
<th>os</th>
<th>osi</th>
<th>_trans</th>
<th>_episode</th>
<th>_start</th>
<th>_stop</th>
<th>_status</th>
</tr>
</thead>
<tbody>
<tr>
<td>11281</td>
<td>2846</td>
<td>24.3</td>
<td>1</td>
<td>26.14</td>
<td>1</td>
<td>72.7</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11282</td>
<td>2846</td>
<td>24.3</td>
<td>1</td>
<td>26.14</td>
<td>1</td>
<td>72.7</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11283</td>
<td>2846</td>
<td>24.3</td>
<td>1</td>
<td>26.14</td>
<td>1</td>
<td>72.7</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11284</td>
<td>2846</td>
<td>24.3</td>
<td>1</td>
<td>26.14</td>
<td>1</td>
<td>72.7</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13397</td>
<td>2846</td>
<td>24.3</td>
<td>1</td>
<td>26.14</td>
<td>1</td>
<td>72.7</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13398</td>
<td>2846</td>
<td>24.3</td>
<td>1</td>
<td>26.14</td>
<td>1</td>
<td>72.7</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Figure: Reversible illness-death model
### Check

```
.tab _tr_epi _status

<table>
<thead>
<tr>
<th>_status</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>_tr_epi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_1</td>
<td>1,464</td>
<td>1,518</td>
<td>2,982</td>
</tr>
<tr>
<td>1_2</td>
<td>739</td>
<td>0</td>
<td>739</td>
</tr>
<tr>
<td>2_1</td>
<td>2,787</td>
<td>195</td>
<td>2,982</td>
</tr>
<tr>
<td>2_2</td>
<td>217</td>
<td>522</td>
<td>739</td>
</tr>
<tr>
<td>3_1</td>
<td>779</td>
<td>739</td>
<td>1,518</td>
</tr>
<tr>
<td>4_1</td>
<td>963</td>
<td>555</td>
<td>1,518</td>
</tr>
<tr>
<td>Total</td>
<td>6,949</td>
<td>3,529</td>
<td>10,478</td>
</tr>
</tbody>
</table>
```

```
preserve
.use multistate_example_temp, clear
(Rotterdam breast cancer data, truncated at 10 years)

tab rfi

<table>
<thead>
<tr>
<th>indicator</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1,464</td>
<td>49.09</td>
<td>49.09</td>
</tr>
<tr>
<td>1</td>
<td>1,518</td>
<td>50.91</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>2,982</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```

---

**Figure:** Reversible illness-death model

[Diagram showing reversible illness-death model with states Health, Relapse, and Death, connected by arrows indicating transitions.]
. tab _tr_epi _status

| _status
_tr_epi | 0  | 1  | Total
-----------+----------------------+----------
1_1       | 1,464 | 1,518 | 2,982
1_2       | 739 | 0 | 739
2_1       | 2,787 | 195 | 2,982
2_2       | 217 | 522 | 739
3_1       | 779 | 739 | 1,518
4_1       | 963 | 555 | 1,518
-----------+----------------------+----------
Total     | 6,949 | 3,529 | 10,478

. preserve
. use multistate_example_temp, clear
(Rotterdam breast cancer data, truncated at 10 years)

. tab rei

| rei | Freq. | Percent | Cum. |
-----------+----------------------+----------
0       | 779 | 51.32 | 51.32 |
1       | 739 | 48.68 | 100.00 |
-----------+----------------------+----------
Total    | 1,518 | 100.00 |

Figure: Reversible illness-death model
. \texttt{tab \_tr\_epi \_status}

<table>
<thead>
<tr>
<th>_tr_epi</th>
<th>_status</th>
<th>_status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>_tr_epi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_1</td>
<td>1,464</td>
<td>1,518</td>
<td>2,982</td>
</tr>
<tr>
<td>1_2</td>
<td>739</td>
<td>0</td>
<td>739</td>
</tr>
<tr>
<td>2_1</td>
<td>2,787</td>
<td>195</td>
<td>2,982</td>
</tr>
<tr>
<td>2_2</td>
<td>217</td>
<td>522</td>
<td>739</td>
</tr>
<tr>
<td>3_1</td>
<td>779</td>
<td>739</td>
<td>1,518</td>
</tr>
<tr>
<td>4_1</td>
<td>963</td>
<td>555</td>
<td>1,518</td>
</tr>
<tr>
<td>Total</td>
<td>6,949</td>
<td>3,529</td>
<td>10,478</td>
</tr>
</tbody>
</table>

\texttt{preserve}

. \texttt{use \textit{multistate\_example\_temp}, clear}
(Rotterdam breast cancer data, truncated at 10 years)

. \texttt{tab osi}

<table>
<thead>
<tr>
<th>survival</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>alive</td>
<td>1,710</td>
<td>57.34</td>
<td>57.34</td>
</tr>
<tr>
<td>deceased</td>
<td>1,272</td>
<td>42.66</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>2,982</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

. \texttt{restore}

\textbf{Figure:} Reversible illness-death model
Summary

1. Always put the diagram aside
2. Thinking about competing risk
   (What has happened? What may happen next? Risk set? )
3. No error ≠ correct data. The devil is in the details and the data!
Summary

1. Always put the diagram aside
2. Thinking about competing risk
   (What has happened? What may happen next? Risk set? )
3. No error $\neq$ correct data. The devil is in the details and the data!
4. Then...happy hour for multi-state modelling!
Acknowledgements

Nikolaos Skourlis (Karolinska Institutet)
Appendix

Slides and syntax of this presentation can be found at: https://github.com/enochytchen/NordicStata2022