ISSUES IN THE USE OF MULTI-STATE MODELS FOR EVENT HISTORY ANALYSIS

Jerry Lawless
University of Waterloo

Workshop on Current Issues in the Analysis of
Incomplete Longitudinal Data
October 13 - 15, 2005
Fields Institute

OUTLINE

- Multi-state models
- Incomplete data
- Some applications and illustrations
- Estimation and analysis
- Gaps in methodology

MULTI-STATE MODELS

- Individuals in some population may occupy states $1, 2, \ldots, k$ over some period of time

- Consider process $\{Y(t), t \geq 0\}$ where $Y(t) \epsilon \{1, 2, \ldots, k\}$ is the state occupied at time t.

- Transition probabilities (TP) are denoted

$$P_{ij}(t, t+s) = Pr\{Y(t+s) = j | Y(t) = i\}$$

- State prevalence or occupancy probabilities (if $Y(0) = 1$)

$$P_j(t) = Pr\{Y(t) = j | Y(0) = 1\}$$

TP's do not in general specify the process fully.

- Transition intensity functions: let $H(t)$ denote the process history $\{Y(u), 0 \leq u < t\}$ up to time t. Then for $i \neq j$

$$\lambda_{ij}\left(t|H(t)\right) = \lim_{s \downarrow 0} \frac{Pr\left\{Y(t+s) = j|Y(t-) = i, H(t)\right\}}{s}$$

Markov processes: $\lambda_{ij}\left(t|H(t)\right) = \lambda_{ij}(t)$

Semi-Markov processes: $\lambda_{ij}\left(t|H(t)\right) = \lambda_{ij}\left[B(t-)\right]$

where $B(t)$ = time since individual entered current state.

INCOMPLETE DATA

- Intermittent observation: subject i seen only at times $a_{ij} \;\; (j = 0, 1, \ldots, m_i)$, so that only $Y_i(a_{ij})$'s are known. Transitions between those times are unobserved.

- Initial conditions: information in $H(a_{i0})$, needed for intensity function modelling, may be missing.

- End of followup and loss to followup

- Missing covariate values

- Measurement error (transition times, covariates)

  - effects of intermittent observation

SOME APPLICATIONS AND ILLUSTRATIONS

- Disease processes
  e.g. simple illness - death process

  - onset of disease (e.g. diabetes, CD)

  - organ transplantation (1 - waiting list, 2 - transplanted, 3 - dead)

$$\begin{tikzcd} \boxed{1} \arrow[rr] \arrow[dr] & & \boxed{2} \arrow[dl] \\ & \boxed{3} & \end{tikzcd}$$

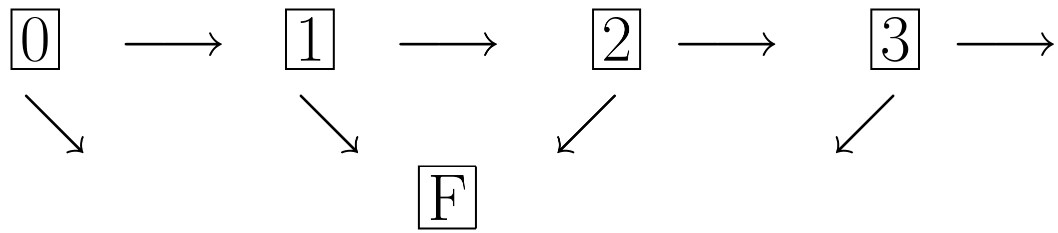
- Interactions between events

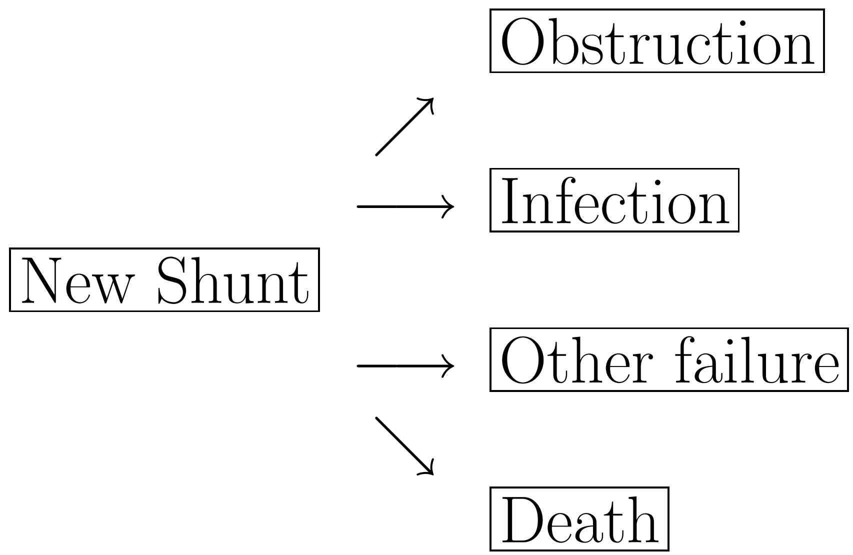
  - two events A and B (e.g. menopause, breast cancer)

$$\begin{tikzcd} \boxed{0} \arrow[r] \arrow[d] & \boxed{\mathrm{B}} \arrow[d] \\ \boxed{\mathrm{A}} \arrow[r] & \boxed{\mathrm{AB}} \end{tikzcd}$$

  - recurrent events and a failure time (e.g. strokes, death)

    e.g. Children with hydrocephalus and cerebrospinal fluid shunts.
    Shunt failures (due to infections, obstruction, other causes) that necessitate (partial) shunt replacement.
    Some patients die.


0
1
2
3
F


Obstruction
Infection
New Shunt
Other failure
Death

- Dependent loss to followup
  - Intermittent observation of subjects; subject has not been seen at recent observation times.
  - When to declare subject lost to followup (LTF) ?
  - Status re LTF may depend on process history.

$$
\begin{tikzcd}
\boxed{1} \arrow[rr] \arrow[dr] & & \boxed{2} \arrow[dl] \\
& \boxed{\text{LTF}} &
\end{tikzcd}
$$

- Cumulative cost models
  - Can associate a cost rate with different states
  - Useful in connection with medical costs etc.
  - Cumulative quality of life measures

ESTIMATION AND ANALYSIS

- Can write down likelihood functions with intensity-based models and complete observation (Andersen et al. 1993)
  - Allows maximum likelihood inference on intensities
  - For some models (Markov, Semi-Markov), survival analysis software can be used for estimation
    (e.g. Therneau and Grambsch 2000; Lawless 2003).
  - Survival models and software that handle time-varying covariates can deal with a wider range of multi-state models
  - Inference about transition probabilities or state duration distributions may be complicated

- Markov models (see Andersen et al. 1993)
  - Nonparametric estimation of transition probabilities (Aalen-Johansen estimate)
  - can fit proportional intensities models with Cox model methods $\lambda_{ij}(t|x) = \lambda_{ij0}(t) \exp(\beta' x)$
  - Parametric models can be fitted with survival or general optimization software
  - Key point: upon entry to a new state, consider the time T of exit from that state, and what other state is then entered. This is a competing risks failure time problem.

Markov models: T is left-truncated at time of entry to new state.

Semi-Markov models: "clock" starts at $T = 0$ at time of entry to new state.

- Intermittent Observation
  - Much more difficult to handle, aside from time- homogeneous Markov models (Gentleman et al. 1994, R function panel)
  - With equi-spaced observation times, models for longitudinal discrete (categorical) responses can be employed.
  - There is a severe shortage of methodology (and computational support) in this area.

- Missing covariates, measurement errors re events or covariates.
  - Almost nothing has been done

SOME GAPS IN METHODOLOGY

- Consider studies with intermittent observation of subjects

  - Statistics Canada Survey of Labour and Income Dynamics (SLID): persons seen once a year for 6 years

  - Followup of persons attending disease clinics

- Brief looks at dependent loss-to-followup; goodness of fit; missing covariates and response-selective observation; measurement error; modelling issues

Periodic Inspections and Non-Independent LTF

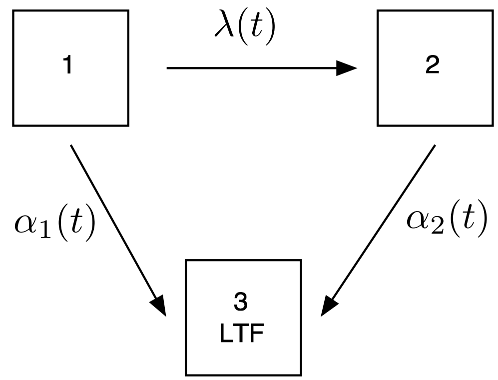
- Suppose individuals are inspected at times $a_0 < a_1 < a_2 < \cdots < a_k$ but that an individual may be found to be LTF at any time $a_j (j = 1, \ldots, k)$, and never seen henceforth.

- Independent inspections: next inspection time after a_{j-1} depends only on event history and covariates up to a_{j-1}.

- What if LTF at a_j is related to the event history over $(a_{j-1}, a_j]$, even after conditioning on covariates and event history up to a_{j-1}?

- Illustration of effects in event history setting: consider transitions from some state to another state, say state 1 to state 2.

  Consider effect of state-dependent LTF rates.



- Want to estimate $\lambda(t)$

- At inspection time a_j, the time of a $1 \longrightarrow 2$ transition during $(a_{j-1}, a_j]$ can be determined.

- When a person is found to be LTF (in state 3) at a_j, the time they became LTF cannot be determined.

- LTF is non-independent in this setting if $\alpha_1(t) \neq \alpha_2(t)$.

- Define for $s \leq t$

  $P_{ij}(s,t)$=P(in state j at time t| in state i at time s)

- For $a_{j-1} < t \leq a_j$, if we treated LTF as independent (non-differential, i.e. $\alpha_1(t) = \alpha_2(t)$), then non-parametrically we end up estimating not $\lambda(t)$ but

  P [entry to state 2 at t| in state 1 at $t-$, in states 1 or 2 at a_j]

$$= \frac{P_{11}(a_{j-1},t-)\lambda(t)P_{22}(t,a_j)}{P_{11}(a_{j-1},t-)[P_{11}(t-,a_j)+P_{12}(t-,a_j)]}$$

$$= \lambda(t)\left\{\frac{P_{22}(t,a_j)}{P_{11}(t,a_j)+P_{12}(t,a_j)}\right\}$$

$$= \lambda^*(t)$$

- If $\alpha_2(t) > \alpha_1(t)$, $\hat{\lambda}(t)$ is biased downward.

  e.g. $\alpha_1(t) = \alpha_1, \alpha_2(t) = \alpha_2, \lambda(t) = \lambda$. Then for $\alpha_2 - \alpha_1$ small,

$$\lambda^*(t) \simeq \left\{\frac{1}{1+(\alpha_2-\alpha_1)(a_j-t)}\right\}\lambda(t) \qquad\qquad a_{j-1} < t \leq a_j$$

- For **correct** estimation of $\lambda(t)$ we need (estimates of) $\alpha_1(t), \alpha_2(t)$ or at least their difference. (Can then use ML or weighted GEE's)
  - Can be estimated if there are data on transitions to LTF from both states 1 and 2 (e.g. unemployment studies)
  - If not, then look at sensitivity of inferences for $\lambda(t)$ to variations in $\alpha_2(t) - \alpha_1(t)$.

- Tracing studies: trace some persons LTF

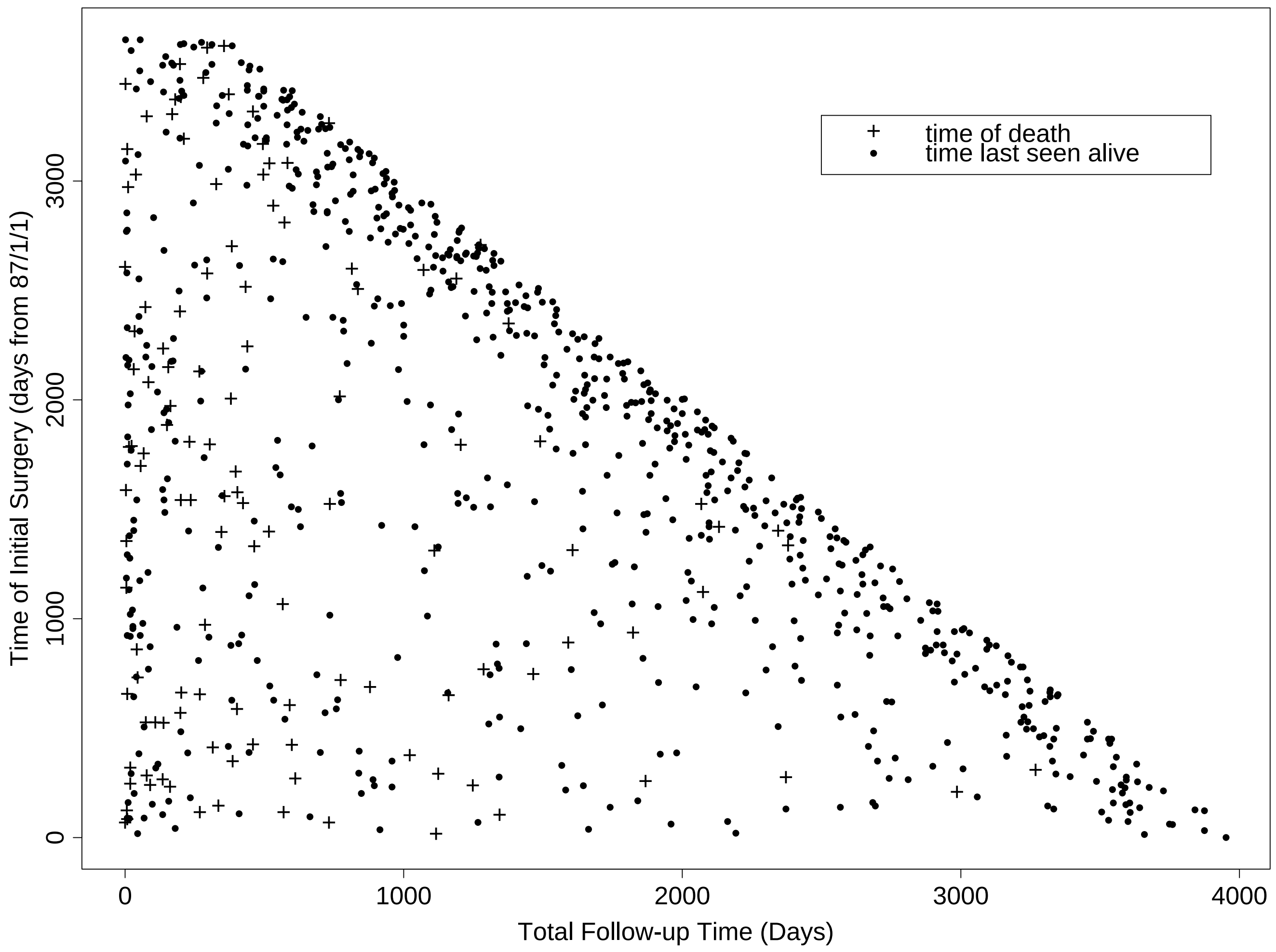
- Other issues in observational studies

  - Persons not seen for a long time
    (assignment of a LTF time? dependent LTF?)

  - Delayed reporting of terminal events
    e.g. Children with CSF shunts

  - See following plot of time of entry to study (time of initial shunt surgery) vs. length of followup as of December 1997, for children getting CSF shunts.

Rheumatic disease clinics: Farewell et al. (2003)



time of death
time last seen alive
Time of Initial Surgery (days from 87/1/1)
0
1000
2000
3000
Total Follow-up Time (Days)
0
1000
2000
3000
4000

Goodness of Fit

- Model expansion (tests model of interest vs a larger model)
  - effective methods ?

- Comparison of empirical and model-based estimates

  e.g. Aguirre-Hernandez and Farewell (2002) - Pearson test based on pseudo observed transition counts for Markov models

- Another idea: look at state prevalence probabilities

$$P_j(t) = Pr\left\{Y(t) = j|Y(0) = 1\right\}$$

  - Need an empirical (nonparametric) estimate of $P_j(t)$ that can be compared with the model-based one.

- One possibility: let T_j and W_j denote times of entry and exit from state j (assume can be occupied just once). Then

$$P_j(t) = Pr(T_j \leq t) - Pr(W_j \leq t)$$

  Estimate $Pr(T_j \leq t)$ and $Pr(W_j \leq t)$ nonparametrically (Turnbull estimates)

- This and alternatives when there is continuous observation of subjects: Cook and Lawless (2003).

- A possible alternative: develop nonparametric estimates of $P_j(t)$ for Markov models (robust in continuous observation case)

  - how to do when observation of individuals is intermittent ?

Longitudinal Multi-phase Observation

- Subjects seen at times $a_0 < a_1 < a_2 < \ldots$, at which the current states $Y(a_j)$ and covariates $x(a_j)$ are observed.

- A subset of subjects is selected at a_j , and harder-to-obtain covariates $z(a_j)$ are measured; the probability a subject is included in this subset depends on their current (and maybe past) values for Y and x.

- Objective is to model $Pr\{Y(t+s)|H(t), x(t), z(t)\}$.

Feasibility ?

Simple case: disease incidence studies

Measurement Error

- In many studies, the time of events or values of covariates in the time interval $(a_{j-1}, a_j]$ can be retrospectively ascertained at the observation time a_j.

- Same for initial conditions at time a_0

- Often subject to measurement errors.

How to deal with this ?

e.g. Survey of Labour and Income Dynamics (SLID)

- When person is "enrolled", suppose they are unemployed. Should data be collected on when they became unemployed?

Some Modelling Issues

- Hard to fit non-Markov models in many settings with intermittently observed states.
- Ability to check model assumptions depends on gaps between observation times.
- Robut methods related to longitudinal discrete response models (e.g. Carroll, Lin, many others)
  - categorical response $Y(t)$, covariates $x(t)$ with unequally spaced observation times
  - marginal methods that focus on $Pr\{Y(t)|(x(t)\}$ are quite well developed
  - conditional modelling that considers $Pr\{Y(t)|H(t),x(t)\}$ has not received so much attention
- Hidden Markov and other latent process models (e.g. Satten and Longini, 1996)

References

Aguirre-Hernandez, R and Farewell, VT (2002). *Statistics in Medicine*, 21, 1899-1911.

Anderson, PK, Borgan, O, Gill, RD and Keiding, N (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Cook, RJ, Lawless, JF and Lee,KA (2003). *Statistics and Operations Research Transactions (SORT)*, 27, 13-30.

Farewell, VT, Lawless, JF, Gladman, DD and Urowitz, MB (2003). *Applied Statistics*, 52, 445-456.

Gentleman, R, Lawless, JF, Lindsey, JC and Yan, P (1994). *Statistics in Medicine*, 13, 805-821.

Lawless, JF (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, Hoboken, NJ.

Lawless, JF, Wigg, MB, Tuli, S, Drake, J and Lamberti-Pasculli, M (2001). *Applied Statistics*, 50, 449-465.

Satten, G and Longini, I (1996). *Applied Statistics*, 45.

Therneau, TM and Grambsch, PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.