Analysis of Longitudinal Data

- Patrick J. Heagerty PhD
- Department of Biostatistics
- University of Washington

LDA Progress!

- During the last couple of decades statistical methods have been developed (ie. LMM, GEE) that can analyze longitudinal data with:
  - ▷ Unequal number of observations per person (n_i)
  - ▷ Unequally spaced observations (t_{ij})
  - ▷ Time-varying covariates (x_{ij})

- Regression questions:

$$\mu_i(t) = E[Y_i(t) \mid X_i(t)]$$

- Q: When should we directly apply these now standard longitudinal methods to data with the features listed above?

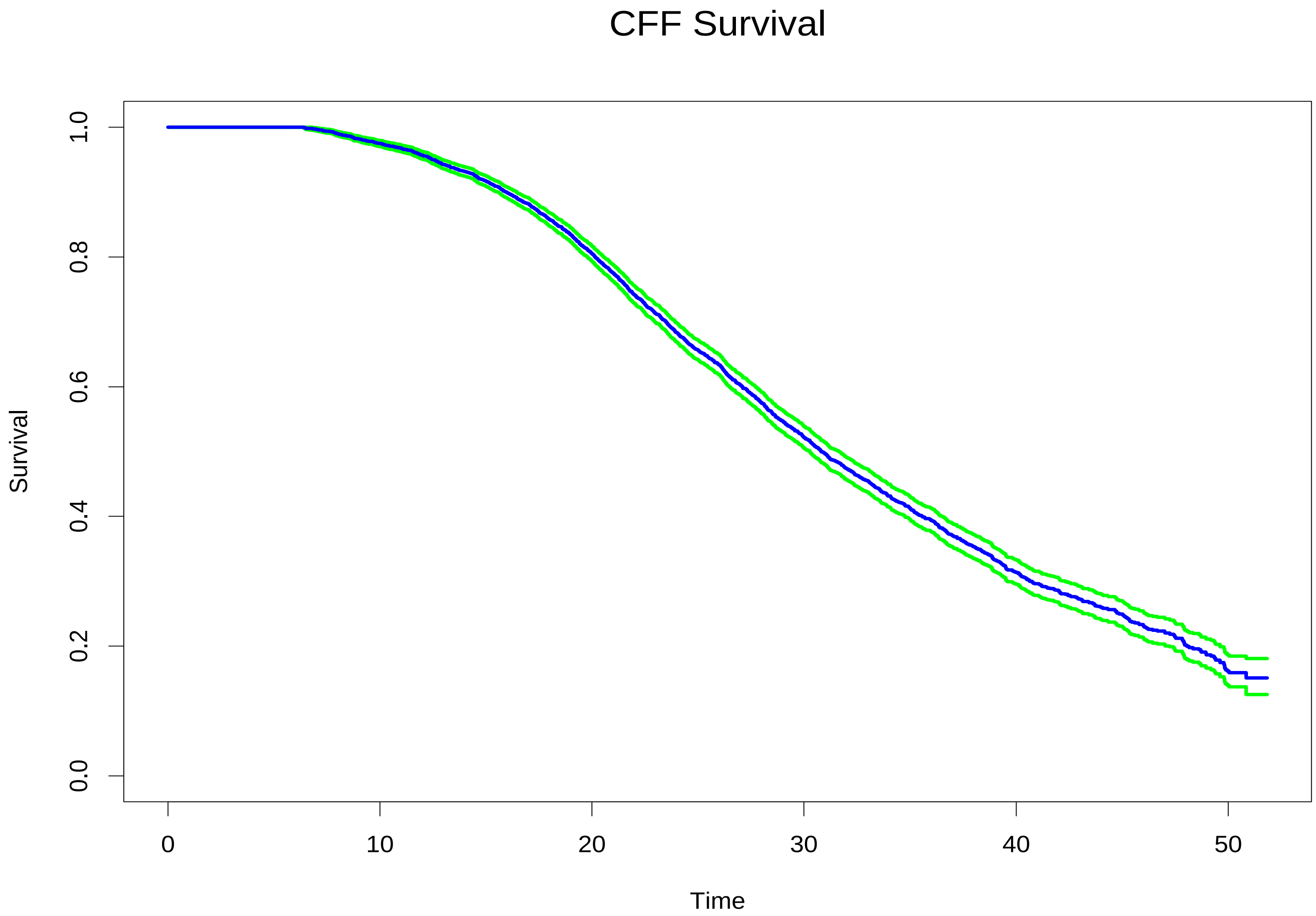
Session Eight Outline

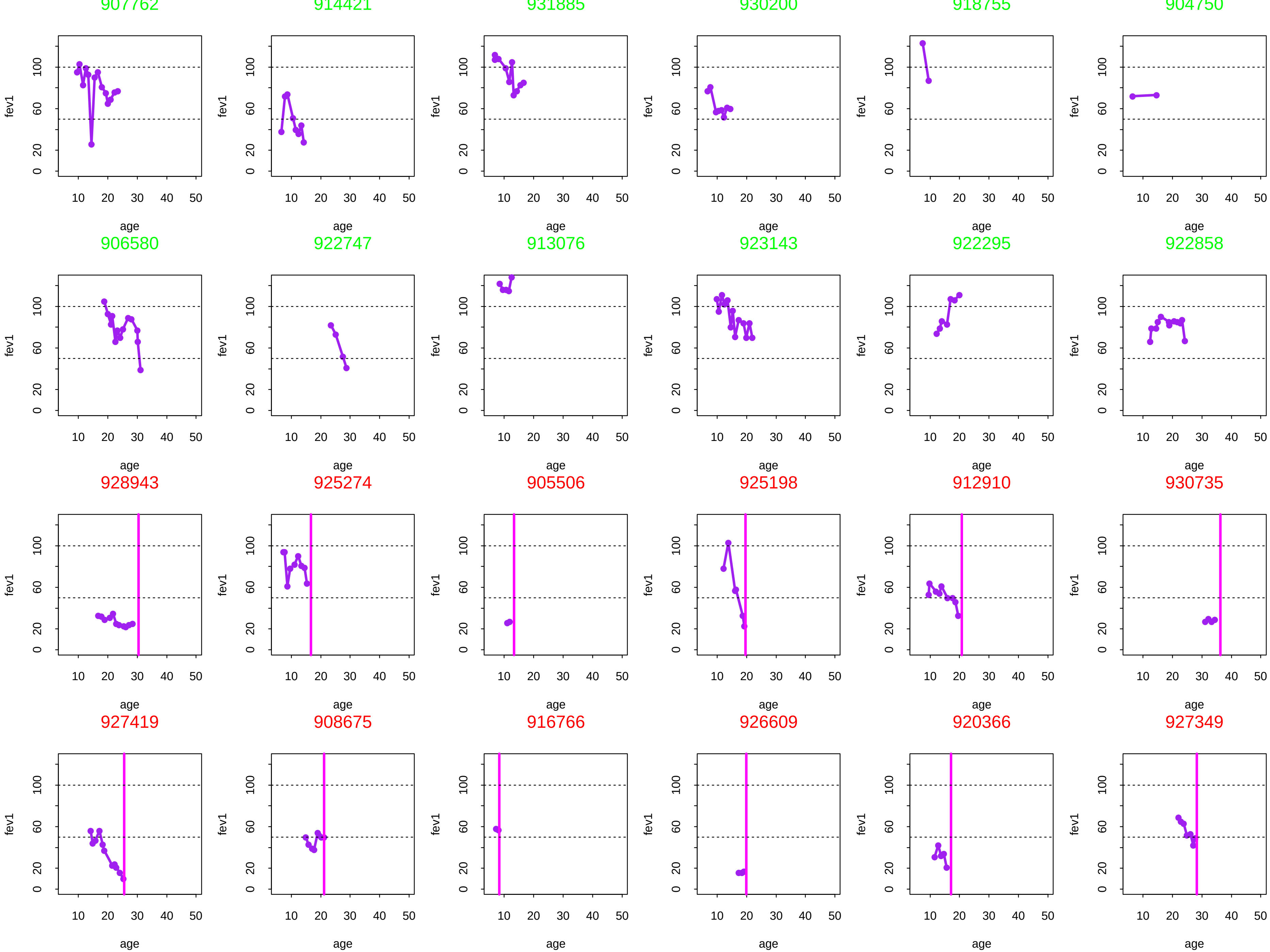
- Examples
  - ▷ Cystic Fibrosis Foundation (CFF)
  - ▷ Maternal Stress and Child Morbidity (MSCM)
  - ▷ United States Renal Data System (USRDS)
- Time-varying Covariate Processes
  - ▷ Exogenous
    - ∗ Lagged covariates
  - ▷ Endogenous
    - ∗ Fixed vs Dynamic exposure
- Analysis with Death
  - ▷ Specification of model
  - ▷ Inference

Repeated Measures Data

Cystic Fibrosis Data

- $N = 23,530$ subjects, $4,772$ deaths, 1986-2000
- $n = 160,005$ longitudinal observations
- Longitudinal measurements: FEV1, weight, height
- Goal: identify factors associated with decline in pulmonary function.
- (Another Goal: predict mortality; transplantation selection)

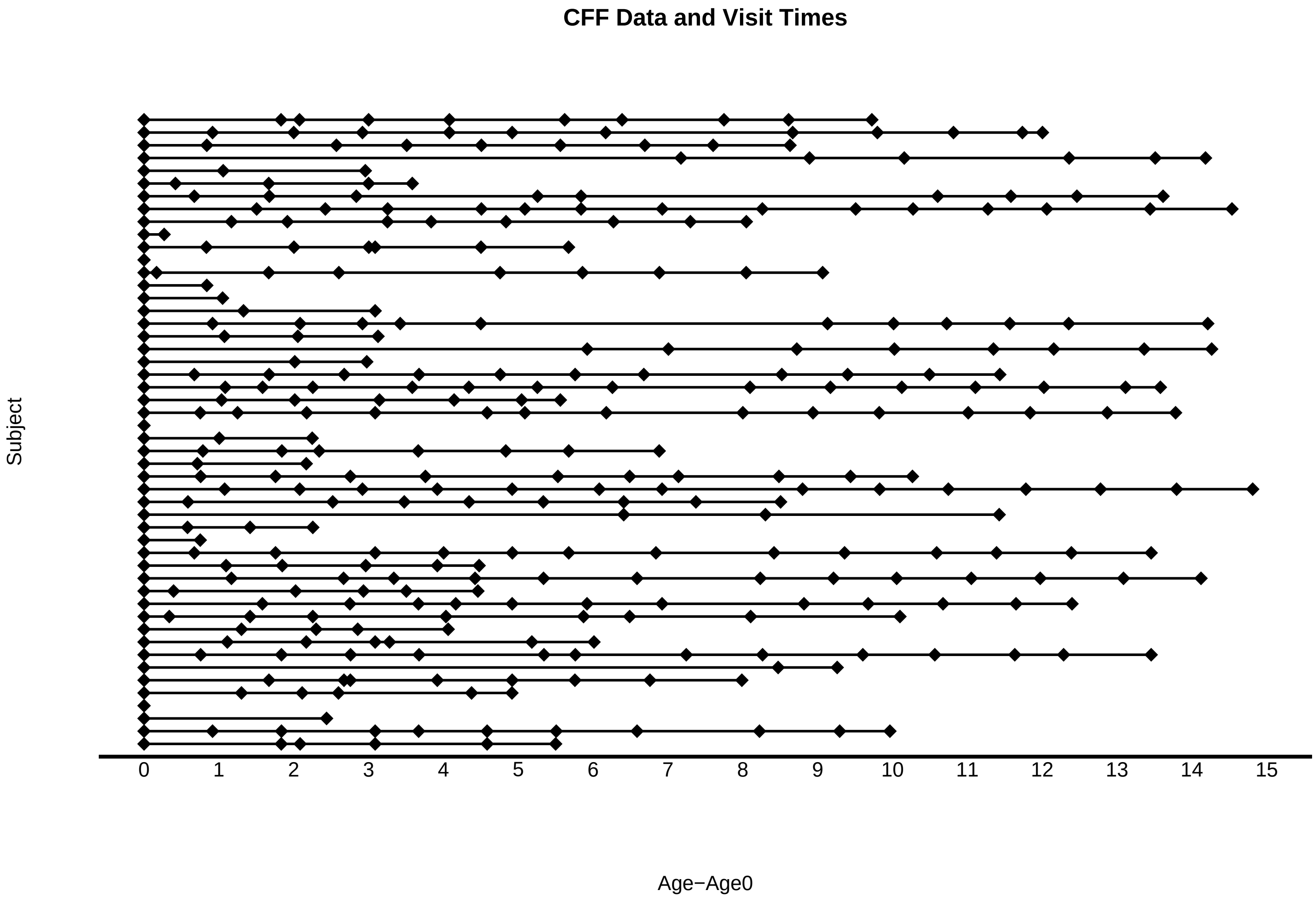

CFF Survival
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Time

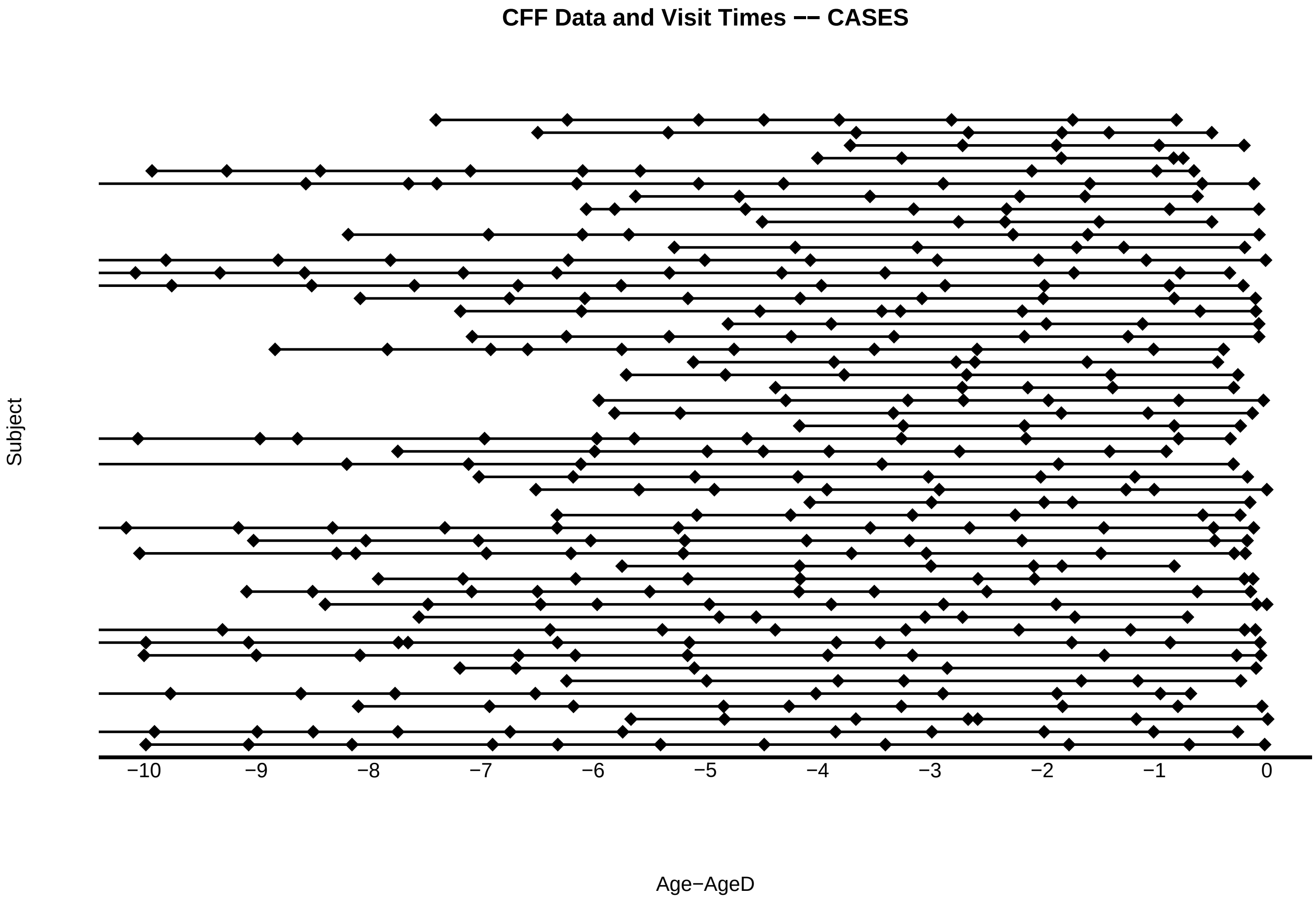

907762
914421
931885
930200
918755
904750
906580
922747
913076
923143
922295
922858
928943
925274
905506
925198
912910
930735
927419
908675
916766
926609
920366
927349
fev1
age

Example: Scientific Goals & CF

- Parad RB, Gerard CJ, Zurakowski D, Nichols DP, Pier GB "Pulmonary outcome in cystic fibrosis is influenced primarily by mucoid Pseudomonas aeruginosa infection and immune status and only modestly by genotype."
  Infect. Immun., **67**(9): 4744-50, 1999.

- Variables:
  - ▷ Measurement time: t_{ij}
  - ▷ Pulmonary function: $Y_i(t_{ij})$
  - ▷ Time-dependent covariate: $X_i(t_{ij})$ – infection status
  - ▷ Death: $D_i(t)$ counting process for T_i


CFF Data and Visit Times
Subject
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Age−Age0


CFF Data and Visit Times −− CASES
Subject
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
Age−AgeD

Maternal Stress and Child Morbidity

Example 2: Time-dependent covariates

- daily indicators of stress (maternal), and illness (child)
- primary outcome: illness, utilization
- covariates: employment, stress
- **Q**: association between employment, stress and morbidity?
- **Q**: Does stress cause morbidity?

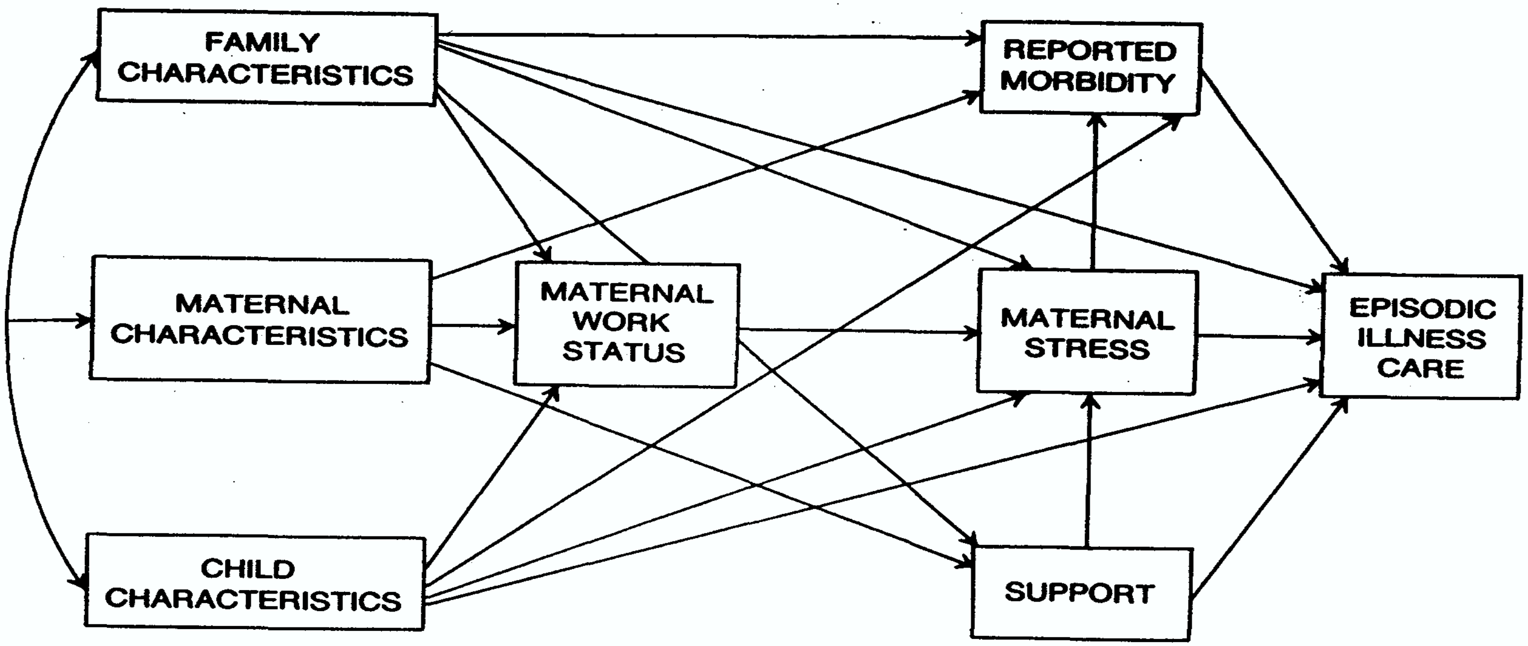
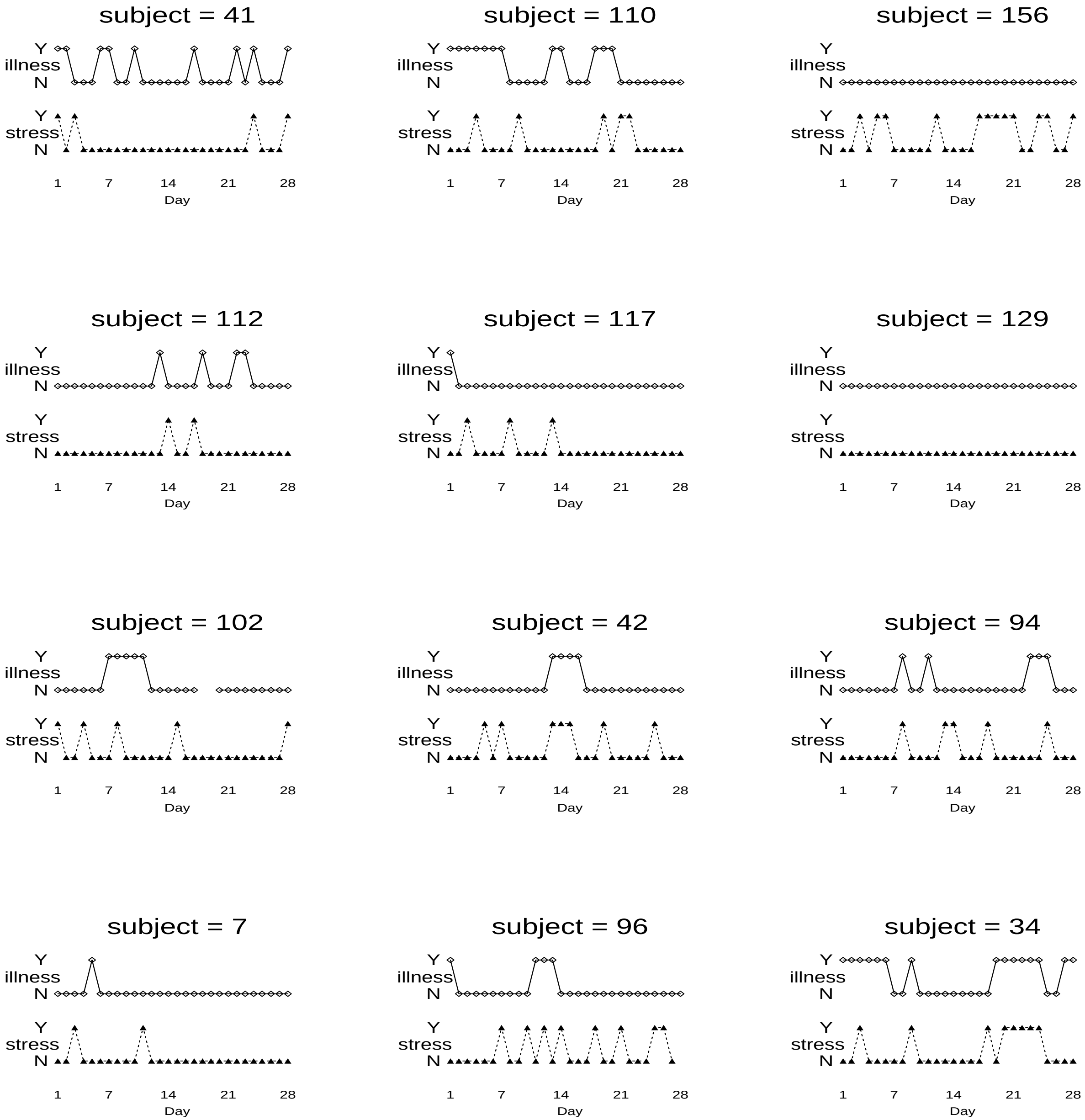


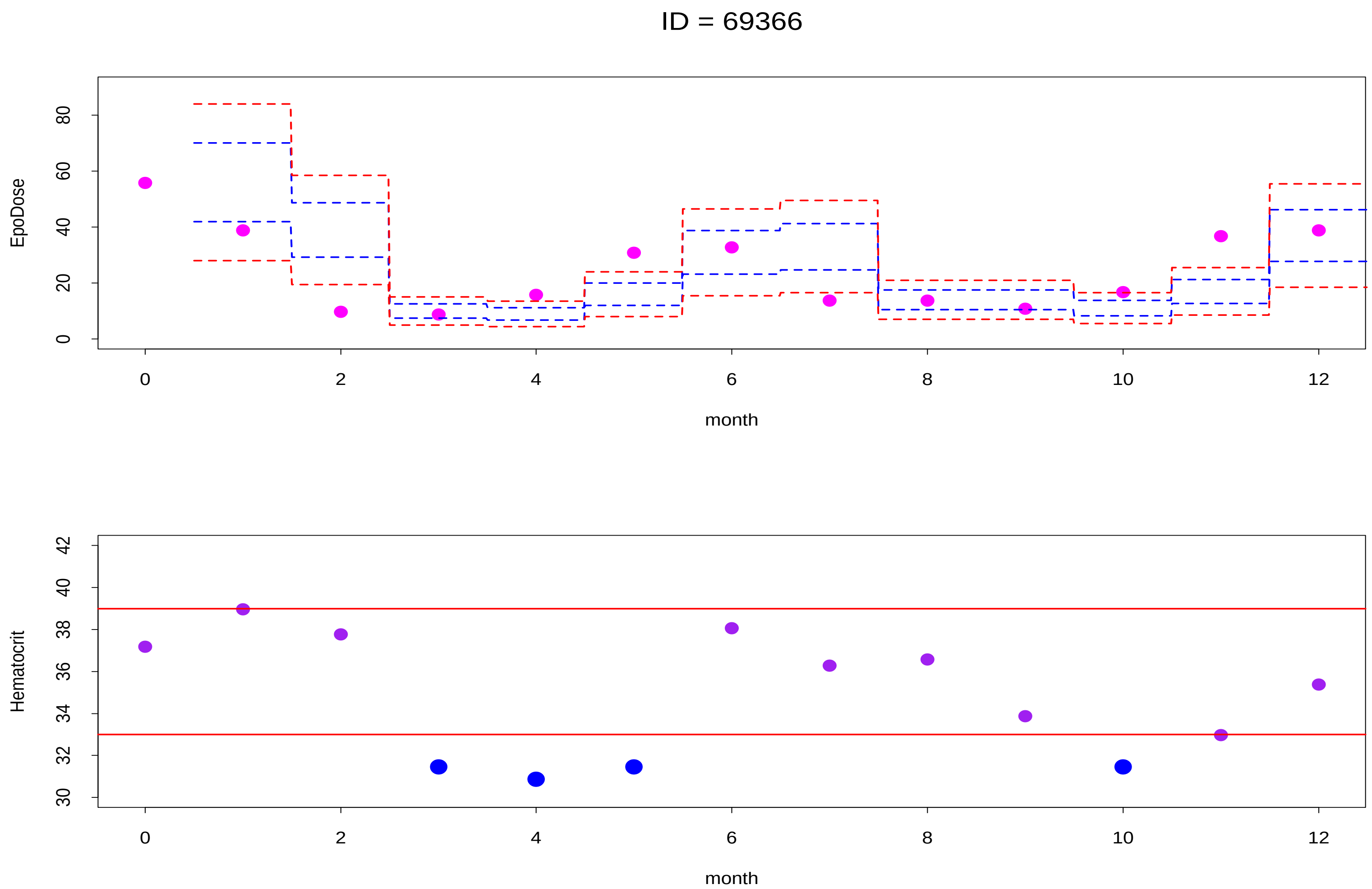
FIG. 1. Determinants of episodic illness care utilization.


subject = 41
subject = 110
subject = 156
subject = 112
subject = 117
subject = 129
subject = 102
subject = 42
subject = 94
subject = 7
subject = 96
subject = 34
Y
illness
N
Y
stress
N
1 7 14 21 28
Day

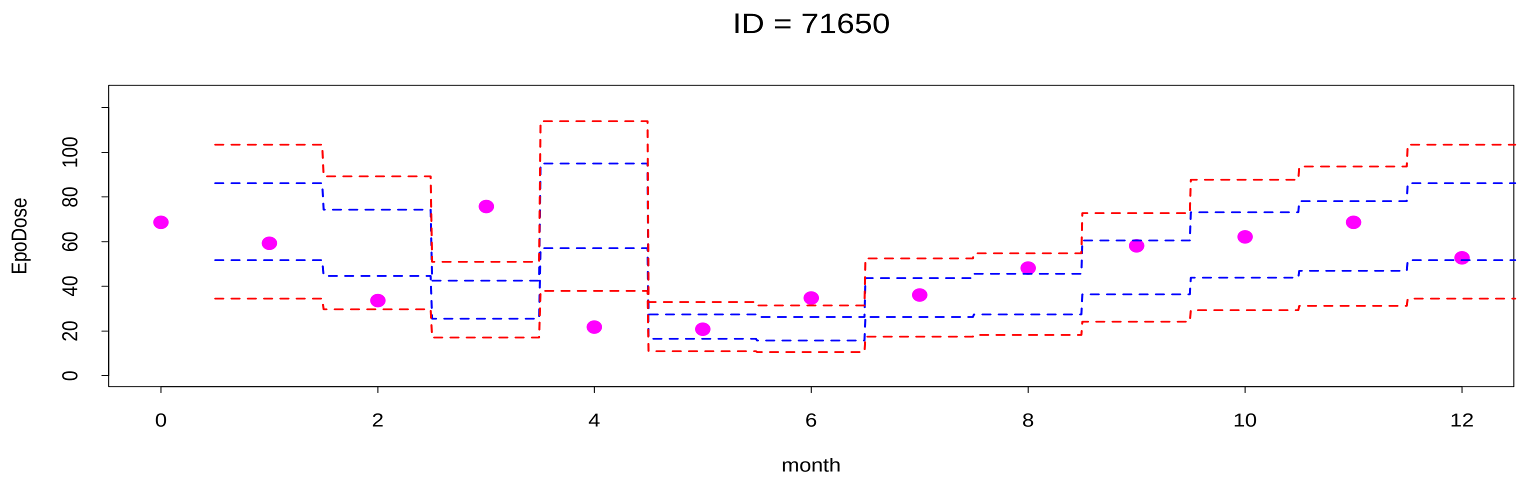
USRDS Data: Safety of ESAs?

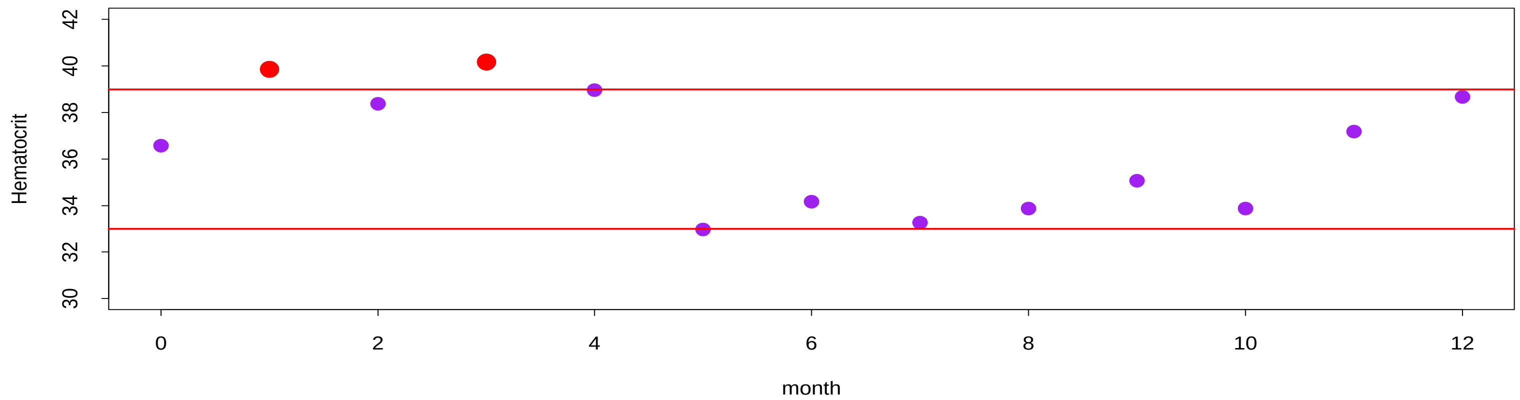
- End Stage Renal Disease (ESRD)
  - ▷ Poor kidney function
  - ▷ Dialysis
  - ▷ Fail to stimulate formation of red blood cells
- Epoetin
  - ▷ Anemia treatment
  - ▷ $3 billion Medicare / year
- Studies show an association between high dose and risk of death
  - ▷ Adverse outcomes?
  - ▷ Confounding by indication?

USRDS Dialysis Data



USRDS Dialysis Data





The Processes

	Primary Process
$Y_i(t)$	The response process
	Secondary Processes
$X_i(t)$	The covariate process
$S_i(t)$	The scheduling process (**not today**)
$R_i(t)$	The recording process
$D_i(t)$	The death process
$B_i(t)$	The birth process (**not today**)

LDA and Regression

- Most statistical representations focus on discussion of

$$\mu_i(t) = E[Y_i(t) \mid X_i(t)]$$

- But what about the other processes? Do we mean:

$$\text{CFF} \quad : \quad E[Y_i(t) \mid X_i(t), X_i(s), S_i(t) = 1, R_i(t) = 1, D_i(t) = 0]$$

$$\text{USRDS} \quad : \quad E[Y_i(t) \mid X_i(t), X_i(s), R_i(t) = 1, D_i(t) = 0]$$

$$\text{MSCM} \quad : \quad E[Y_i(t) \mid X_i(t), X_i(s), R_i(t) = 1]$$

Motivation: Hospitalization and EPO Dose?

- Background:
  - **NEJM – November 2006**
    - RCTs target high versus low hemoglobin
    - Higher target $\rightarrow$ higher Epo dose
    - Higher target associated with AEs
  - **FDA – March 2007**

    Issued a "black box warning" which indicated that aggressive use of erythropoiesis-stimulating agents to raise hemoglobin to a target of 12 g/dL or higher was associated with "serious and life-threatening side-effects and/or death."
- General Question:
  - **Q**: Are higher doses of EPO associated with greater rates of adverse events such as hopitalization?

Motivation: Full Data History

- Regression:

$$E[\text{Hosp}(t) \mid \text{Dose}(t-1),\ \text{Dose}(t-2), \ldots, X]$$

- Statistical Issues:
  - ▷ What aspects of **exposure history** are associated with current hosp?
  - ▷ What is the role of the **outcome history** $\text{Hosp}(t-1),\ \text{Hosp}(t-2), \ldots$?
  - ▷ What is the role of **intermediate history** $\text{Hem}(t-1),\ \text{Hem}(t-2), \ldots$?

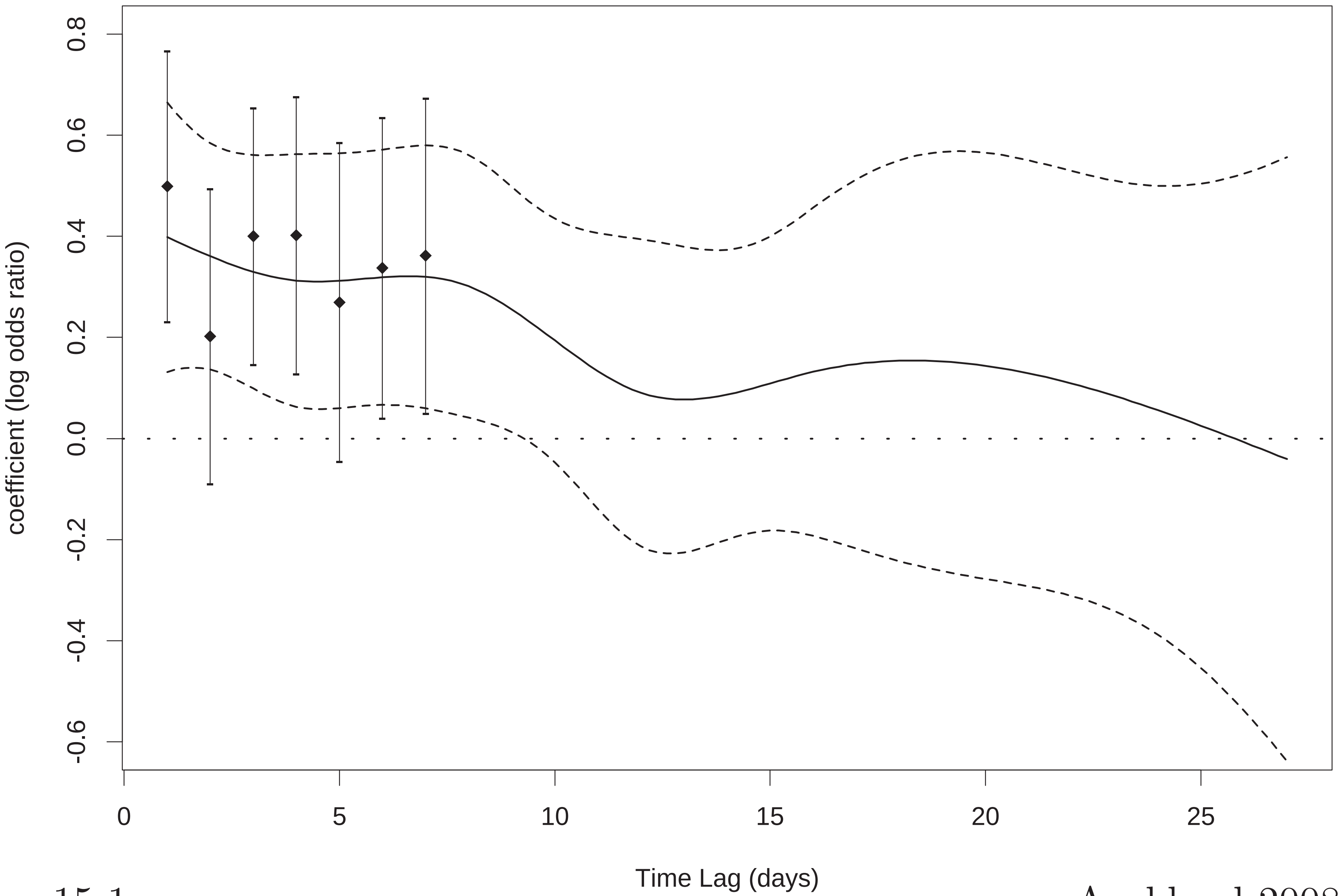
Time-dependent Covariates: Lagged Covariates

- **Exogenous** – future covariates are not influenced by current / past outcomes.

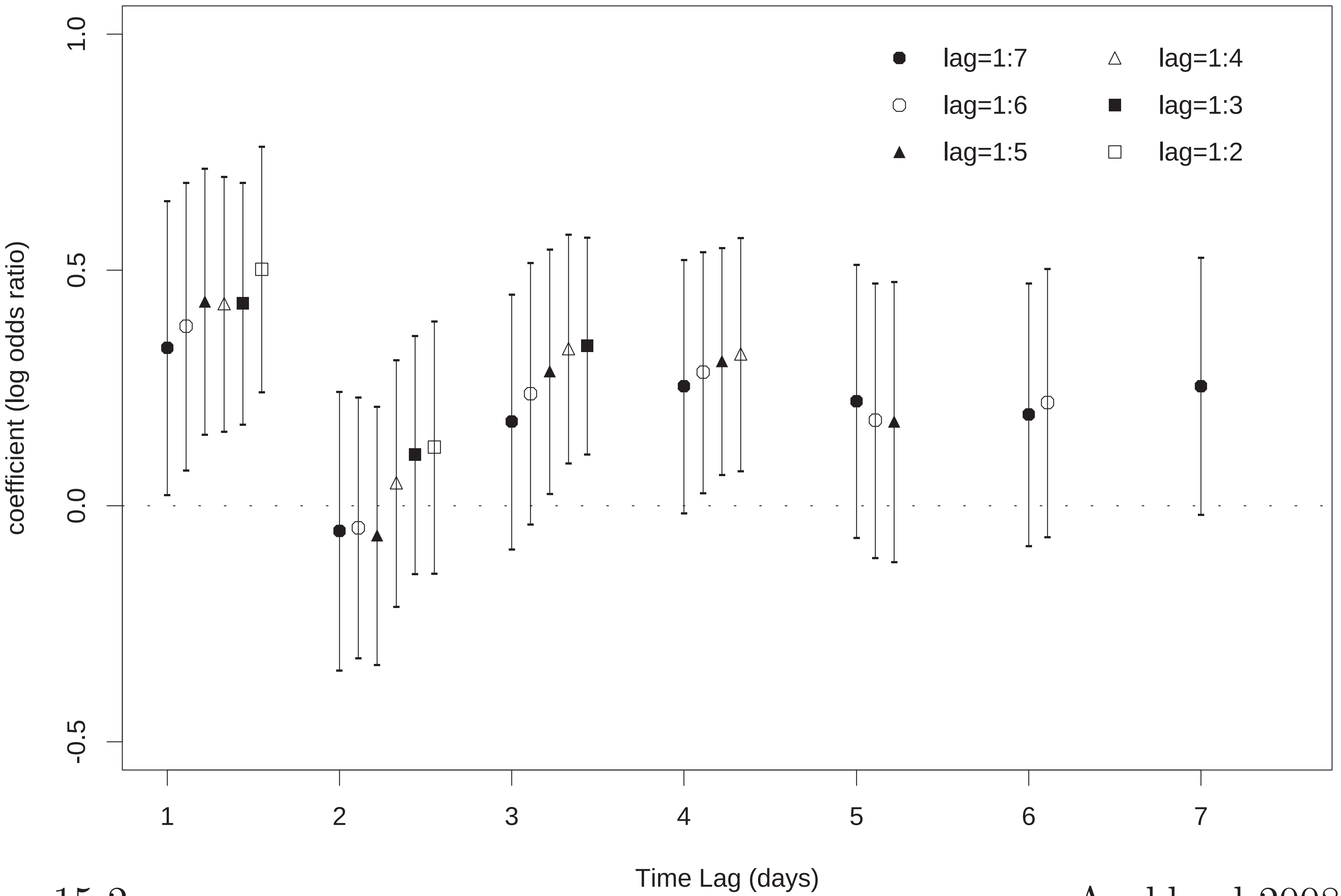
$$[X(t+1) \mid Y(t), X(t)] \sim [X(t+1) \mid X(t)]$$

- **Analysis Issues:**
  - ▷ Include single lagged covariates (current, cumulative)
    - ∗ **MSCM**: $E[\text{Sick}(t) \mid \text{Stress}(t-k)]$
    - ∗ **USRDS**: $E[\text{Hosp}(t) \mid \text{Dose}(t-k)]$
  - ▷ Include multiple lagged covariates
    - ∗ **MSCM**: $E[\text{Sick}(t) \mid \text{Stress}(t-1), \text{Stress}(t-2)]$
    - ∗ **USRDS**: $E[\text{Hosp}(t) \mid \text{Dose}(t-1), \text{Dose}(t-2)]$

Lag Coefficient Function



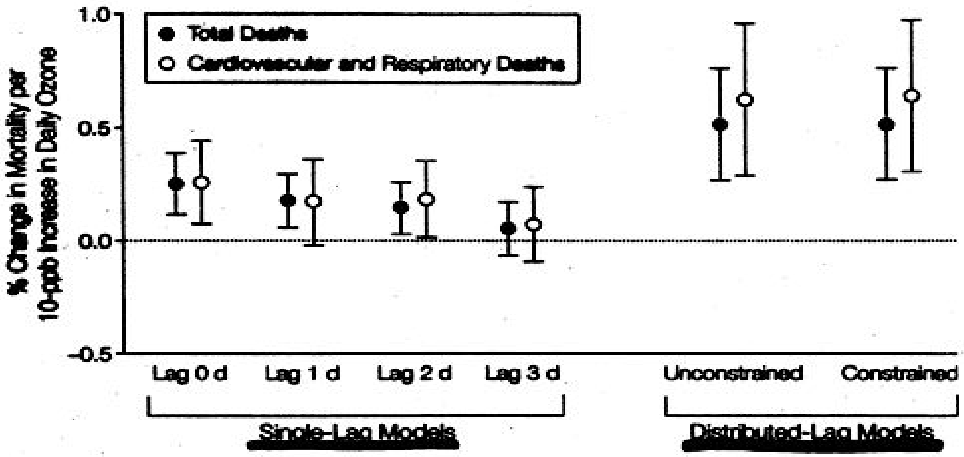
Multivariate models with different lags



Bell et al. *JAMA* (2004)

OZONE AND MORTALITY IN US URBAN COMMUNITIES

Figure 1. Percentage Change in Daily Mortality for a 10-ppb Increase in Ozone for Total and Cardiovascular Mortality, for Single-Lag and Distributed-Lag Models



The single-lag model reflects the percentage increase in mortality for a 10-ppb increase in ozone on a single day. The distributed-lag model reflects the percentage change in mortality for a 10-ppb increase in ozone during the previous week. Error bars indicate 95% posterior intervals.

Endogenous: Analysis

- **Definition**: – The covariate is influenced by past outcomes (or intermediate variables)

$$Y(t) \rightarrow X(t+1)$$

- **Implication**:

$$E[Y_i(t) \mid X_i(1), \ldots, X_i(n)]$$

depends on $X_i(s)$ for $s > t$ (future values of covariate).

- Role for causal inference concepts.
- See: DHLZ (2002) section 12.5 for introduction.

Causal Targets of Inference

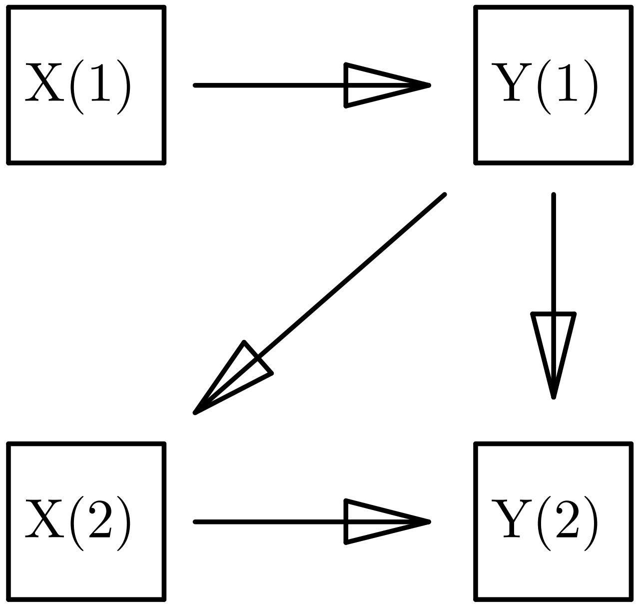
- Longitudinal Treatment

$$\text{vec}(X_0) \equiv [X(1) = \mathbf{0}, X(2) = \mathbf{0}, \ldots, X(n) = \mathbf{0}]$$
$$\text{vec}(X_1) \equiv [X(1) = \mathbf{1}, X(2) = \mathbf{1}, \ldots, X(n) = \mathbf{1}]$$

- Population Means
  - ▷ Mean of population if all subjects had $X = 1$ at all times, and similar population mean if $X = 0$ at all times.

$$\mu_0(n) \equiv E\left[Y(n) \mid \text{vec}(X_0)\right]$$
$$\mu_1(n) \equiv E\left[Y(n) \mid \text{vec}(X_1)\right]$$

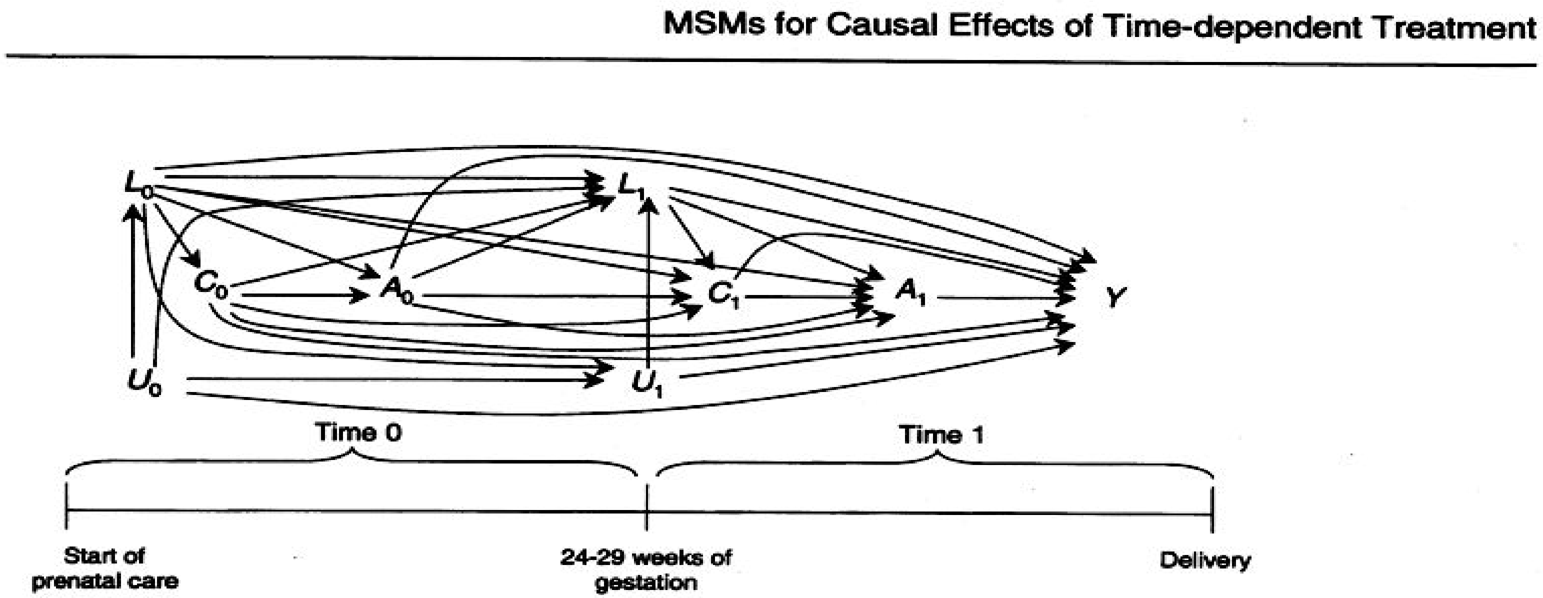
Endogenous Covariates



treatment / exposure

response

Bodnar et al. *AJE* (1997)



Model / Estimation

- **G-computation**

  ▷ **Model**: model **outcome** given past outcomes / exposure.

$$P[Y(t) \mid X(t), \{Y(s), X(s)\}\ s < t] \quad : \quad \textbf{outcome}$$

$$P[X(t) \mid \{Y(s), X(s)\}\ s < t] \quad : \quad \textbf{exposure}$$

  ▷ **Compute**: compute means of interest by allowing intermediate effects, $Y(s)$, to occur naturally, but controlling exposure.

$$\mu_1(t) = E_t \{E_s[Y(t) \mid \mathbf{X(t)=1}, \{Y(s), \mathbf{X(s)=1}\}\ s < t]\}$$

Model / Estimation

- **Marginal Structural Models**
  - **Model**: model **exposure** given past outcomes / exposure.

$$[X(t) \mid \{Y(s), X(s)\}\ s < t]$$

  - **Compute**: compute a regression of the outcome using inverse probability weights (IPW) to control for exposure selection bias.

Table 1: Regression of stress, S_{it}, on illness, I_{it-k} $k = 0, 1$, and previous stress, S_{it-k} $k = 1, 2, 3, 4+$ using GEE with working independence.

	est.	s.e.	Z
(Intercept)	-1.88	(0.36)	-5.28
I_{it}	0.50	(0.17)	2.96
I_{it-1}	0.08	(0.17)	0.46
S_{it-1}	0.92	(0.15)	6.26
S_{it-2}	0.31	(0.14)	2.15
S_{it-3}	0.34	(0.14)	2.42
mean(S_{it-k}, $k \geq 4$)	1.74	(0.24)	7.27
employed	-0.26	(0.13)	-2.01
married	0.16	(0.12)	1.34
maternal health	-0.19	(0.07)	-2.83
child health	-0.09	(0.07)	-1.24
race	0.03	(0.12)	0.21
education	0.42	(0.13)	3.21
house size	-0.16	(0.12)	-1.28

Table 2: MSM estimation of the effect of stress, S_{it-k} $k \geq 1$, on illness, I_{it}.

	est.	s.e.	Z
(Intercept)	-0.71	(0.40)	-1.77
S_{it-1}	0.15	(0.14)	1.03
S_{it-2}	-0.19	(0.18)	-1.05
S_{it-3}	0.18	(0.15)	1.23
mean(S_{it-k}, $k \geq 4$)	0.71	(0.43)	1.65
employed	-0.11	(0.21)	-0.54
married	0.55	(0.17)	3.16
maternal health	-0.13	(0.10)	-1.27
child health	-0.34	(0.09)	-3.80
race	0.72	(0.21)	3.46
education	0.34	(0.22)	1.57
house size	-0.80	(0.18)	-4.51

method	logOR
GEE cross-sectional association	0.66
GEE with seven days lagged	1.38
Transition model (direct effect)	0.50
G-computation	0.80
MSM	0.85

Summary of Endogenous

- Interest in exposure over time – more than simply the acute (most recent) exposure.
- A variable (perhaps outcome) is both a **consequence** of exposure at early times, and a **cause** of exposure at later times.
- Intermediate and confounder.
- **G-computation**
- **MSM**
- Interest in outcomes under a controlled and **static** treatment plan.

EPO: November 2006 NEJM

- **Drüeke** CREATE
  - ▷ Control **Hemoglobin** rather than fix the dose.
    - ∗ Low group (11.0-12.5)
    - ∗ Normal group (13.0-15.0)
- **Singh** CHOIR
  - ▷ Control **Hemoglobin** rather than fix the dose.
    - ∗ Low group (11.3)
    - ∗ Normal group (13.5)
- Research Question(s)

  Q: What target hemoglobin should be used? How to use observational data to compare different targets and/or compare mortality experience to RCT data?

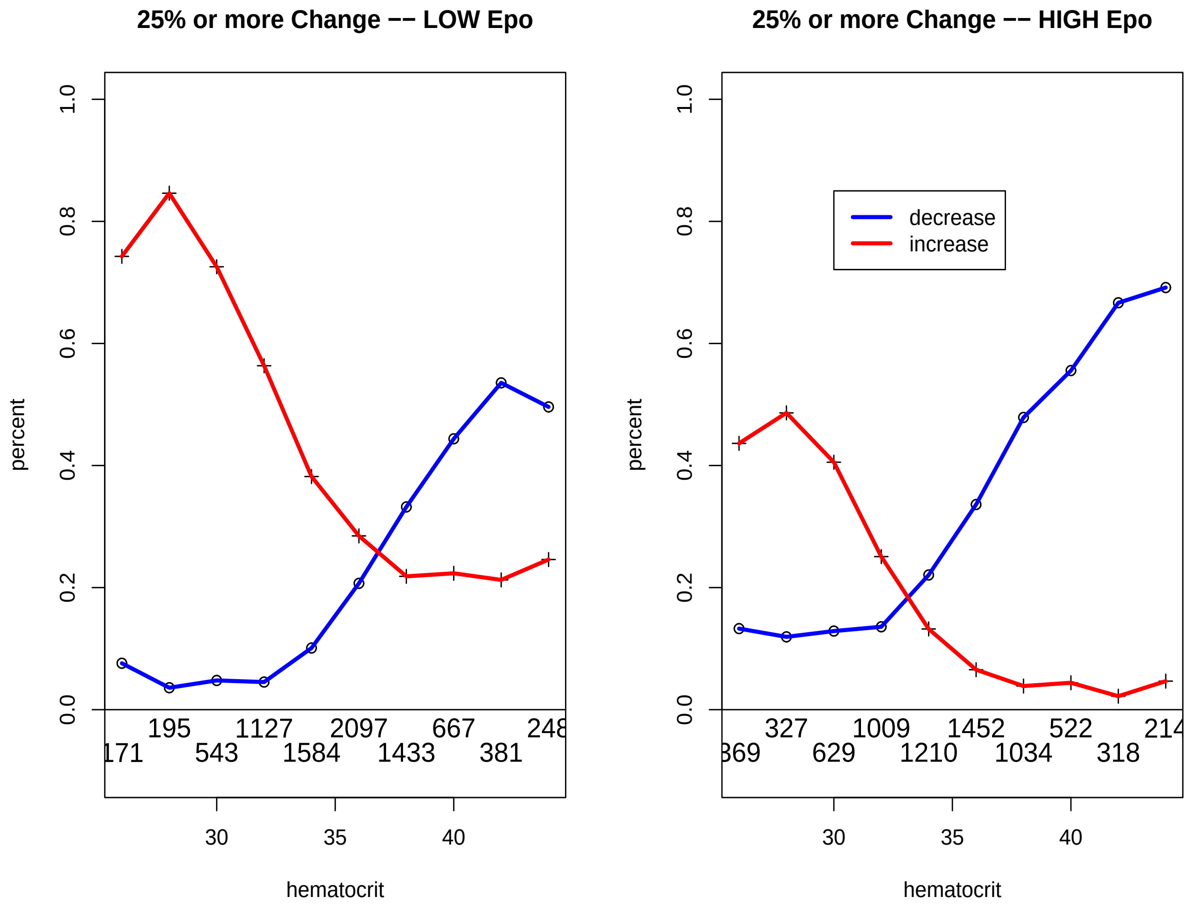
Analysis of Dynamic Treatment

- **Note**: The guidelines for Epo do not suggest a **static** dose be administered. Rather, dose is driven by the state of the intermediate (Hb):

$$X(t+1) = \begin{cases} 1.25 \times X(t) & \text{if} \quad Z(t) \leq 11 \\ X(t) & \text{if} \quad 11 < Z(t) \leq 13 \\ 0.75 \times X(t) & \text{if} \quad Z(t) > 13 \end{cases}$$

- This corresponds to a **dynamic treatment guideline**, $\mathcal{G}_1$.
- **Q**: How to formulate DOSE questions in this setting?
  - ▷ $\mathcal{G}_1$ corresponds to correction of ±25% at Hb=(11,13).
  - ▷ Compare to a $\mathcal{G}_2$ which uses alternative target Hb threshold(s).

USRDS Data (2003 sample)



LDA with Death

- Different than drop-out
- With Drop-out:

$$E[Y_i(t) \mid X_i] = E[Y_i(t) \mid X_i, R_i(t) = 1] \times P[R_i(t) = 1 \mid X_i] + E[Y_i(t) \mid X_i, R_i(t) = 0] \times P[R_i(t) = 0 \mid X_i]$$

- Linear Mixed Models (LMM) applied to the observed data where $R_i(t) = 1$ can validly estimate parameters in the mean $E[Y_i(t) \mid X_i]$ when data are MAR.
- With Death:

$$E[Y_i(t) \mid X_i] = E[Y_i(t) \mid X_i, D_i(t) = 0] \times P[D_i(t) = 0 \mid X_i] + E[Y_i(t) \mid X_i, D_i(t) = 1] \times P[D_i(t) = 1 \mid X_i]$$

LDA with Death: Analysis

- Analysis conditional on death information:
  - Full (future) stratification:

$$E[Y_i(t) \mid X_i(t), \mathbf{T_i = s}] \;\; s > t$$

    - See: Pauler, McCoy & Moinpour (2003)
  - Partial (current status) conditioning:

$$E[Y_i(t) \mid X_i(t), \mathbf{T_i > t}]$$

    - See: Kurland and Heagerty (2004)
  - Conditional on principal strata (potential status):

$$E[Y_i(t \mid 1) - Y_i(t \mid 0) \mid \{\mathbf{T_i(0) > t, T_i(1) > t}\}]$$

    - See Frangakis and Rubin (2002), Rubin (2007)

LDA with Death: Comments on Analysis

- Full stratification using $[\mathbf{T_i = s}]$ $s > t$
  - ▷ Compares groups defined by X_i comparable in terms of death.
  - ▷ Conditions on future (not yet observed) information.
- Partial (current status) conditioning: $[\mathbf{T_i > t}]$
  - ▷ Conditions on observed vital status.
  - ▷ Compares groups defined by X_i after selection by death.
- Principal stratification: $[\{\mathbf{T_i(0) > t, T_i(1) > t}\}]$
  - ▷ Compares subgroups defined by X_i comparable in terms of death.
  - ▷ Conditions on unobservable potential status.

Some recommendations

- In applications we should identify factors that influence the **secondary stochastic processes** and choose appropriate statistical techniques in order to validly answer the scientific question.

- In statistical research reports we should be explicit about the **assumptions** we are making regarding the secondary stochastic processes.

- For **time-dependent covariates** ask about associations with both past and future covariate values – consider the factors that drive the covariate.

Thanks!