Joint Model of Longitudinal and Survival Data

Lei Liu
Department of Preventive Medicine
Robert H. Lurie Comprehensive Cancer Center
Northwestern University

Introduction

In many clinical studies, there are two types of data of interest: longitudinal measurements, and time to an event.

- Repeated measures: daily alcohol drinking level; Time to event: dropout
- Repeated measures: Prostate Specific Antigen (PSA); Time to event: prostate cancer recurrence.
- Repeated measures: monthly medical cost; Time to event: death

Research Goal I

- To account for potential informative dropout in the longitudinal outcomes
- Primary outcome: repeated measures (e.g., daily drinking level)
- Dropout needs to be tackled to alleviate potential bias

Research Goal II

- Use repeated measures of a biomarker (e.g., PSA) to predict time to event (prostate cancer recurrence)
- Primary outcome: time to event
- Repeated measures are a covariate process

Research Goal III

- Both longitudinal and survival outcomes are of primary interest
- Longitudinal (e.g., monthly) medical costs and survival are both needed to derive the cost effectiveness measures, e.g., Incremental Cost Effectiveness Ratio (ICER).

Goal I: Joint Model for Missing Data

- Missing at random is not testable (Little and Rubin 2002).
- Joint model provides a sensitivity analysis to assess the plausibility of the MAR assumption.

Alcohol Treatment Trial of Ondansetron

- Phase II randomized controlled trial of ondansetron in 283 alcohol-dependent individuals (Johnson et al. 2011)
- Two treatment groups: ondansetron and placebo
- Longitudinal outcome: daily drinking level was recorded in 11 weeks
- Covariates: baseline drinking level, age, gender, and 2 genotypes of interest: LL/LS/SS in the 5'-regulatory region of the 5-HTT gene, and a functional single-nucleotide polymorphism (TT/TG/GG), rs1042173, in the 3'-untranslated region
- Pharmacogenetic study: interaction between genotype combinations and treatment.

Dropout

- Only 64.3% in the ondansetron arm and 70.6% in the placebo arm completed the trial
- Question: is the dropout informative (depends on the drinking outcome)?
- Informative dropout could lead to selection bias: those who dropout early could have worse longitudinal outcomes, so the better values are over-represented in the sample.

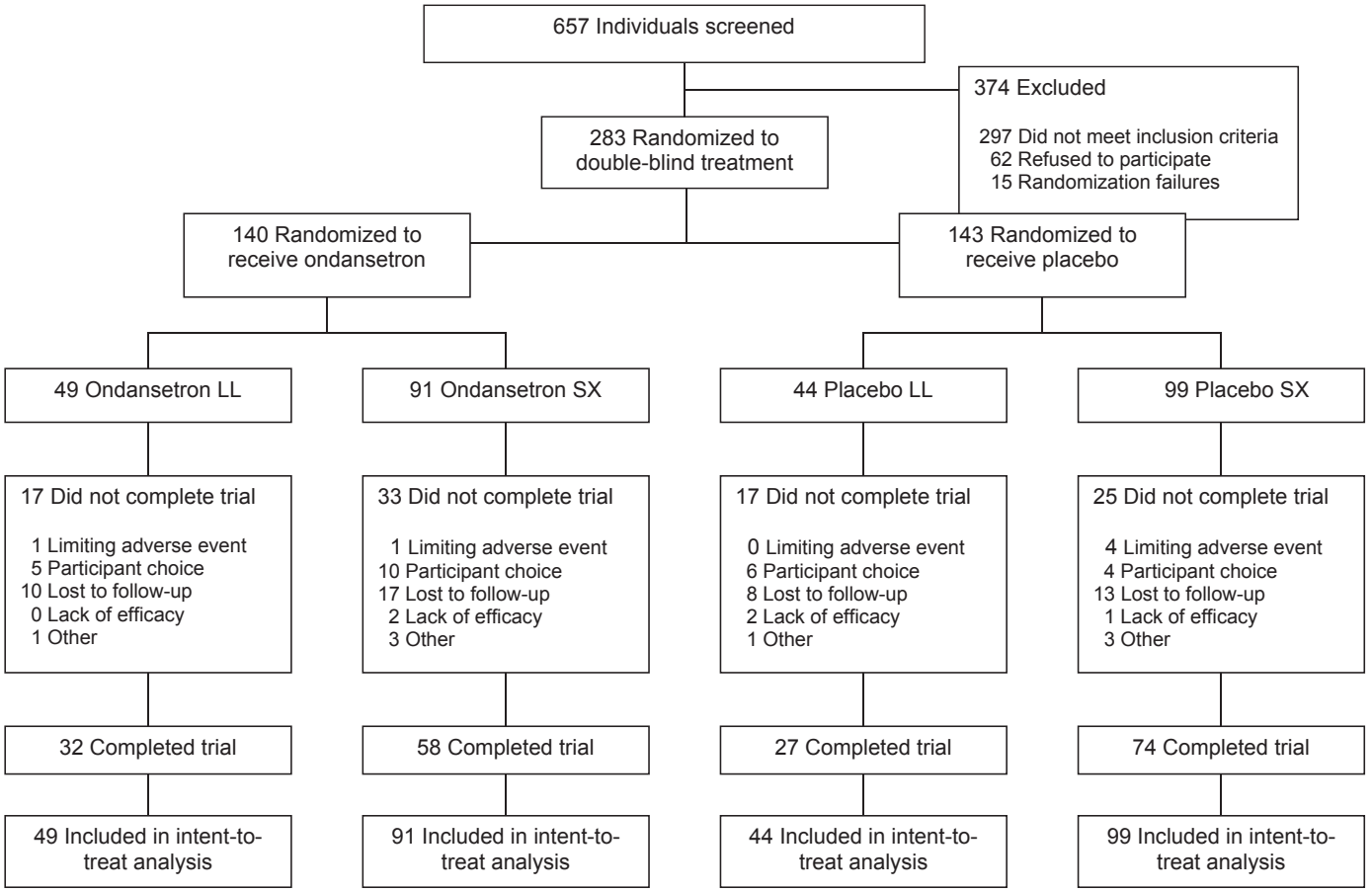

Figure 1: Flow chart of the ondansetron trial (Johnson et al. 2011).

Notations

- Y_{ij}: the drinking level of subject i measured at t_{ij}
- T_i: time to leave the study early
- C_i: independent censoring event, e.g., end of study
- $X_i = \min(C_i, T_i)$: the observed follow-up time
- $\Delta_i = I(T_i \leq C_i)$: the event indicator
- $h_i(t)$: hazard of dropout at time t

- Our Model

$$y_{ij} = Z_{ij}^T \boldsymbol{\beta} + a_i + b_i t_{ij} + e_{ij} \quad (1)$$

$$h_i(t) = h_0(t) \exp(W_i^T \alpha + \gamma_1 a_i + \gamma_2 b_i) \quad (2)$$

- By including random effects in the dropout model, we account for the informative dropout which may depend on the unobserved heterogeneity (e.g., health status, propensity to drink) in the longitudinal measures

- The relation between the longitudinal drinking level and time to dropout is denoted by γ_1 and γ_2
  - If $\gamma_1 > 0$, a higher drinking level is associated with a higher dropout rate
  - If $\gamma_1 < 0$, a higher drinking level is associated with a lower dropout rate
  - If $\gamma_1 = 0$, no association between the mean drinking level and dropout
  - If $\gamma_2 > 0$, a slower decline in drinking level during the trial is associated with a higher dropout rate
  - If $\gamma_2 < 0$, a slower decline in drinking level is associated with a lower dropout rate
  - If $\gamma_2 = 0$, no association between the slope of decline and dropout

Features of the Joint Model

- The dropout model and the longitudinal outcome model are developed jointly - they are correlated.
- The selection bias can be accounted for by the random effects shared between the mixed model of longitudinal outcomes and survival model for the time to dropout.
- Estimation is carried out by maximizing the joint likelihood of the two correlated processes.

Likelihood

- The likelihood for subject i is (with only random intercept a_i)

$$\begin{aligned} L_i &= \int \prod_{j=1}^{n_i} f(y_{ij}) \left[h_0(x_i) \exp(W_i^T \alpha + \gamma a_i)\right]^{\Delta_i} \\ &\quad \exp\left[-\int_0^{x_i} \exp(W_i^T \alpha + \gamma a_i) h_0(t) dt\right] \phi(a_i) da_i \end{aligned} \tag{3}$$

with $f(y_{ij}) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp(-\frac{(y_{ij} - Z_{ij}^T \boldsymbol{\beta} - a_i)^2}{2\sigma_e^2})$, and $\phi(a_i) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp(-\frac{a_i^2}{2\sigma_a^2})$

Estimation Method

- EM algorithm: Wulfsohn and Tsiatis (1997), Liu et al. (2007).
- Gaussian quadrature
  - SAS Proc NLMIXED (Guo and Carlin 2004, Vonesh et al. 2006, and Liu et al. 2008)
  - aML software (Lillard and Panis 2003, freely available at www.applied-ml.com): a multi-level multi-process (outcome) modeling tool
  - JM package in R (Rizopoulos 2012)

Gaussian Quadrature

- Approximate likelihood by a weighted average of the integrand assessed at quadrature points over the random effects (Pinheiro and Bates 1995, Liu and Pierce 1994)
- Adaptive (quadrature points are determined by the empirical Bayes estimate of random effect) and non-adaptive (quadrature points are fixed).
- Adaptive Gaussian Quadrature is more accurate for the same number of quadrature points - AGQ with 5 quadrature points is very accurate.

Baseline Hazard

- Nonparametric baseline hazard is challenging in Cox models with random effects (frailty): partial likelihood involves individual (unknown) frailty term.
- The baseline hazard $h_0(t)$ can be
  - Parametric (e.g., exponential, Guo and Carlin 2004) in SAS Proc NLMIXED
  - Piecewise constant (Vonesh et al. 2006, and Liu et al. 2008) in SAS Proc NLMIXED, and JM package in R
  - Piecewise linear: aML software

Implementation in SAS Proc NLMIXED

```
/* Repeated measures data. */
data one;
input id y trt year;
aa=1; /* aa=1 indicates a longitudinal observation */
cards;
1 0.5 1 0.5
1 1.2 1 1
1 2.0 1 1.5
1 3.4 1 2
2 2.3 0 1
2 2.5 0 2
2 2.7 0 2.5
......
;
run;

/* Survival data. Status=1: Event; Status=0: independent censoring */
data two;
input id stoptime status trt;
aa=2; /* aa=2 indicates a survival outcome */
```

```
cards;
1 10.0 1 1 /* event at time 10.0 */
2 7.9 0 0 /* independent censoring at time 7.9 */
3 6.0 0 1
......
;
run;

* Get all event times;
data three;
set two;
if status=1;
run;

* Calculate the quantiles of the event time;
proc univariate data=three noprint;
var stoptime;
output out=quant pctlpts=0 10 20 30 40 50 60 70 80 90 100 pctlpre=q;
run;

data quant;
set quant;
aa=2;
```

```
run;

* Merge data with the quantiles;
data four;
merge two quant;
by aa;
run;

* Calculate the duration and the indicator of event in each quantile interval;
data five;
set four;
array quant {11} q0 q10 q20 q30 q40 q50 q60 q70 q80 q90 q100;
array dur {10} dur1-dur10;
array event {10} event1-event10;

do i=1 to 10;
    dur{i}=0;
    event{i}=0;
end;

do i=2 to 11;
    if stoptime<=quant{i} then do;
        dur{i-1}=stoptime-quant{i-1};
```

```
        event{i-1}=status;
        i=11;
    end;
    else do;
        dur{i-1}=quant{i}-quant{i-1};
    end;
end;
run;

data six;
set one five;
run;

proc sort data=six;
by id aa;
run;

* Get the initial values for the longitudinal model;
proc mixed data=one;
model y = trt year /s;
    random Int year / type=un sub=id;
run;
```

```
* Get the initial values for the survival model;
proc lifereg data=two;
    model stoptime* status(0)=trt /d=EXPONENTIAL;
run;

proc nlmixed data=six qpoints=5;
parms h1=0.03 h2=0.03 h3=0.03 h4=0.03 h5=0.03 h6=0.03 h7=0.03 h8=0.03 h9=0.03 h10=0.03
alpha1=0 beta0=1 beta1=0 beta2=0 gamma1=-.1 gamma2=0 vara=1 varb=.5 covab=0 vare=.3;
bounds h1 h2 h3 h4 h5 h6 h7 h8 h9 h10 vara varb >=0;

base_haz=h1 * event1 + h2 * event2 + h3 * event3 + h4 * event4 + h5 * event5 + h6 * event
+ h7 * event7 + h8* event8 +h9 * event9 + h10 * event10;  /* baseline hazard */
cum_base_haz=h1 * dur1 + h2 * dur2 + h3 * dur3 + h4 * dur4 + h5 * dur5 + h6 * dur6
+ h7 * dur7 + h8* dur8 +h9 * dur9 + h10 * dur10;  /* cumulative baseline hazard */

if aa=1 then do; /* log likelihood for repeated measures */
    mu1= beta0 + beta1 * trt + beta2 * year  + a + year * b;
    loglik=-.5*(y-mu1)**2/vare-.5*log(vare);
end;

if aa=2 then do; /* log likelihood for survival */
    mu2= alpha1 * trt  + gamma1 * a + gamma2 * b;
    loglik2=-exp(mu2) * cum_base_haz;
```

```
    if status=0 then loglik=loglik2; /*log likelihood for censoring */
    if status=1 then loglik= log(base_haz) + mu2 + loglik2;  /*log likelihood for event *
end;
model id ~ general(loglik);
* "general" indicates that the likelihood is given by SAS statements;

random a b ~ normal([0, 0], [vara, covab, varb]) subject=id;

run;
```

Advantages of the estimation method

- Easy implementation
- Biases are small
- Standard error estimates are obtained directly
- Reasonable computational time

Results of the Ondansetron Alcohol Trial

- For the weekly average of drinks per drinking day, the estimates are

Table 1: Informative Dropout of the Ondansetron Alcohol Trial

Parameter	Estimates	SE	P-Value
γ_1	-0.02	0.05	0.60
γ_2	0.17	0.61	0.78

Informative drop-out

- After adjusting for risk factors, the time to dropout did not depend on the random effects from the longitudinal model of drinking outcomes, suggesting that dropout was not informative.
- Furthermore, the parameter estimates for the longitudinal outcome from this joint model were similar to those in the mixed model for longitudinal data only, which assumed that data were MAR.
- Our analysis provided evidence of the validity of the MAR assumption.
- It was well received, commented by a reviewer: "the analysis of the missing data, and the sophisticated analysis of whether these missing data were likely to be informative rather than random, was very instructive, and convincing".
- Of note, this paper was reported in the first page of *the Wall Street Journal* and highlighted in *JAMA*.

- Another paper using the similar method to tackle the informative dropout issue has been accepted by *JAMA Psychiatry*.
- May be a useful tool for a sensitivity analysis of missing data, especially in response to reviewers' concern on missing data mechanism.

Goal II

- Study the association between longitudinal biomarkers and event of interest
- Use longitudinal biomarkers to predict time to event

CPCRA study

- Terry Beirn community program for clinical research on AIDS study (CPCRA) (Abrams et al. 1994).
- Two arms: ddI and ddC
- CD4 was measured every two months from baseline to month 20.
- The median follow-up time is 13.2 months with censoring rate around 60%
- 100 patients died in the ddI arm, and 88 patients died in the ddC arm
- Goal: study the association of CD4 trajectory and mortality

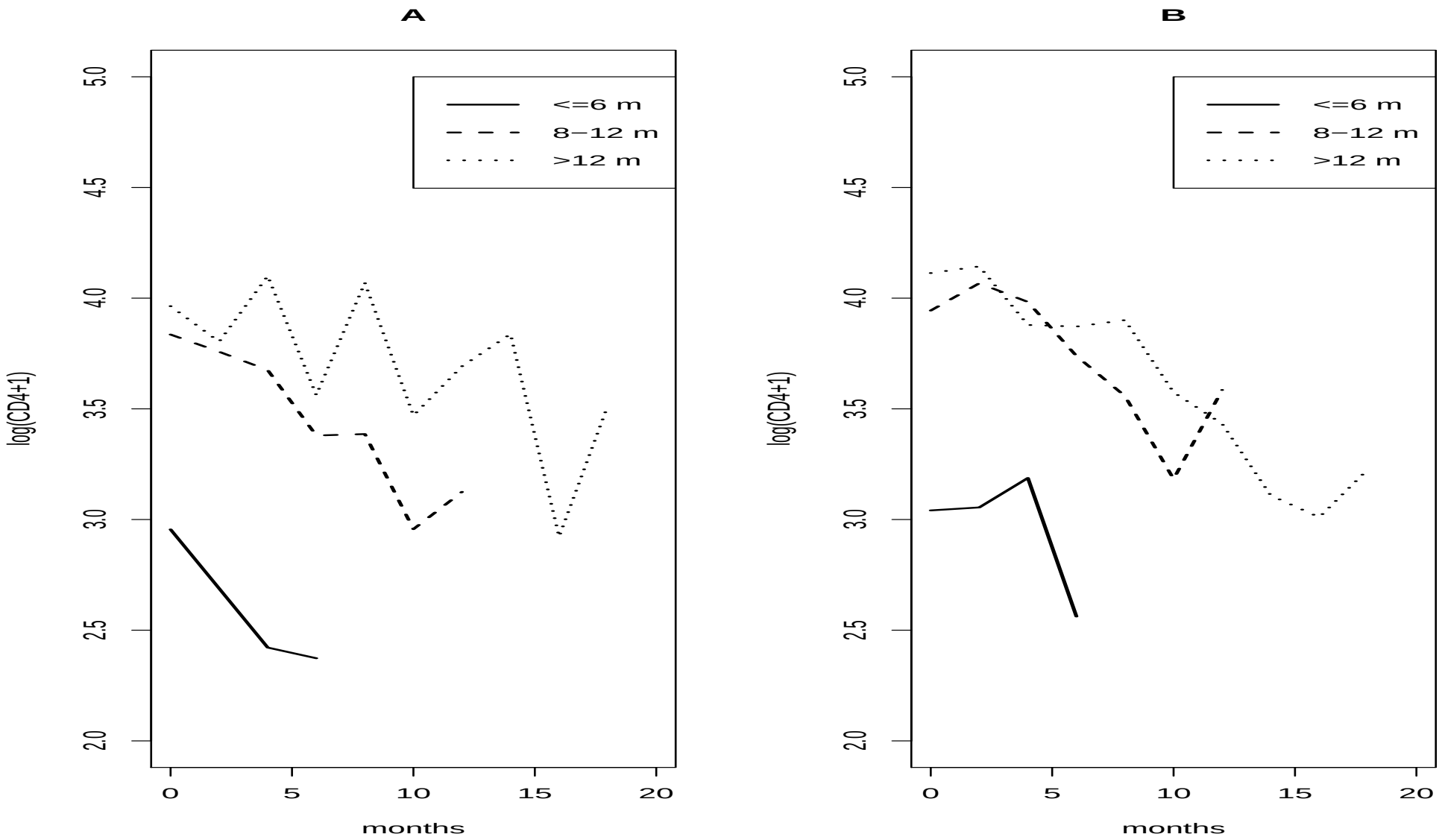

Figure 2: Exploratory plots of CD4 repeated measures with different followup times. (A) Mean CD4 count over time in the ddC group; (B) Mean CD4 count over time in the ddI group. Liu and Huang (2009)

Cox model with a time dependent covariate

- $Y_i(t)$: log(CD4+1) measure of subject i at time t
- The model is

$$h_i(t) = h_0(t) \exp(\alpha_1 Y_i(t) + W_i^T \alpha_2)$$

- $h_i(t)$ is the hazard of subject i
- $h_0(t)$ is the baseline hazard
- W_i includes other covariates measured at baseline: trt, gender, AIDS diagnosis, Hgb
- Can be fitted in SAS Proc PHREG

Drawbacks

- Requires a complete set of repeated measures in a time-continuous process.
  - In reality, the biomarker is measured only at discrete time points - may not be observable at the time of event occurrence.
  - Imputation methods (e.g., last observation carried forward) could be crude and lead to inappropriate inferences, especially when the time interval is long
- Patient survival might depend on the "underlying true" (or expected) values of biomarkers, instead of the observed values with measurement errors.
  - The estimated parameters from such a model are biased toward the null (Prentice, 1982)

- The hazard rate may depend largely on the change and variability in the repeated measures.
  - For example, a decrease in hemoglobin over time was associated with a higher death risk, independent of the baseline hemoglobin value (Regidor et al. 2006)
- However, fitting a Cox model with a time-varying covariate cannot capture such features.

Two stage model

- Proposed by Dafni and Tsiatis (1998)
- Stage 1
  - Obtain an estimated trajectory of the biomarker for each subject during the follow-up, by a random effects mixed model
  - The marker value at each event time can be calculated by the empirical Bayes estimate, i.e., the posterior expected value of random effects conditional on the observed data
- Stage 2
  - The empirical Bayes estimates are plugged in the survival model as predictors for the time to event

Two stage model

- Addresses the afore-mentioned concerns.
- Patients with poorer health have worse biomarker values (e.g., lower CD4), as well as a higher mortality rate or dropout rate (thus shorter follow-up) $\rightarrow$ selection bias

Joint model

- Can alleviate selection bias
- Model A

$$y_{ij} = Z_{ij}^T \boldsymbol{\beta} + a_i + b_i t_{ij} + e_{ij} \quad (4)$$

$$h_i(t) = h_0(t) \exp(W_i^T \alpha + \gamma_1 a_i + \gamma_2 b_i) \quad (5)$$

Table 2: Joint analysis of CD4 count and survival

	Model A			Model B		
	Est	SE	P-value	Est	SE	P-value
			CD4 value			
Trt	-0.105	0.107	0.33	-0.099	0.107	0.36
Time	-0.762	0.068	< 0.0001	-0.773	0.064	< 0.0001
			Survival			
Trt	-0.317	0.158	0.05	-0.294	0.153	0.05
γ_1	-0.506	0.077	< 0.0001			
γ_2	-0.344	0.189	0.07			

Interpretation

- Both γ_1 and γ_2 are significantly (or marginally significantly) less than 0, implying higher initial values and a slower drop in CD4 counts is associated with a better survival.
- After adjusting for other risk factors, a 10% decrease in CD4 value at baseline is associated with 5% increase in the death hazard, while a 10% drop of CD4 value per year is associated with 3% increase in the death hazard.

Another formulation

- Model B:

$$y_{ij} = Z_{ij}^T \boldsymbol{\beta} + a_i + b_i t_{ij} + e_{ij} \quad (6)$$

$$h_i(t) = h_0(t) \exp(W_i^T \alpha + \gamma(a_i + b_i t)) \quad (7)$$

- Hazard depends on the "underlying value" of repeated measures $a_i + b_i t$ through the random effects.
- The two formulations are not nested within each other when there exists random slope: equivalent if only random intercept is present.

- In Model B, the estimate for γ is -0.058 (p = .007), suggesting the negative association between the hypothetical "underlying value" of CD4 and death hazard, i.e., a higher underlying value of CD4 is associated with a lower mortality.
- The AICs for model A vs. model B are 2688.9 vs. is 2725.3 (smaller is better), suggesting the hazard of death is more likely to depend on the initial level and slope of CD4 count in different magnitudes, rather than dependent on the hypothetical "underlying value" of CD4.

Real Time Prediction

- Using PSA to predict prostate cancer recurrence (Taylor et al. 2013).
  - Individual real time prediction of residual time distribution $P(T_i > t|Z, a_i, \theta, T_i > s)$
  - Web calculator at http://psacalc.sph.umich.edu.
- Using longitudinal CD4 to predict time to death (Rizopoulos 2011): implemented in JM package
- Will it predict better than other approaches, e.g., Cox model with time varying covariates or two-stage method???

Goal III

- Joint modeling of longitudinal medical costs and survival: both outcomes are needed in the cost-effectiveness study
- Monthly costs could be 0: joint model of semi-continuous (zero-inflated continuous) data and survival (Liu 2009)
  - Part I: Logistic model for monthly costs being positive

$$\text{logit}P(Y_{ij} > 0) \quad = \quad X_{ij}^T\alpha + a_i \tag{8}$$

  - Part II: Amount of positive monthly costs

$$\log Y_{ij}|(Y_{ij} > 0) \quad = \quad X_{ij}^T\beta + \delta_1 a_i + b_i + e_{ij} \tag{9}$$

  - Joint model with survival

$$\lambda_i(t) \quad = \quad \lambda_0(t)\exp(X_i^T(t)\gamma + \delta_2 a_i + \delta_3 b_i) \tag{10}$$

- We can also consider cost at each (recurrent) hospital visit, so the recurrent hospital visit and medical cost at each visit are correlated (informative observational time, Liu et al. 2008).
  - Intensity of recurrent hospital visits (use of health services):

$$r_i(t) = r_0(t) \exp(X_i^T(t)\alpha + a_i) \tag{11}$$

  - Cost for each visit (linear mixed model):

$$\log Y_i(t) | (dN_i(t) = 1) = X_i^T(t)\beta + \gamma_1 a_i + b_i + e_i(t) \tag{12}$$

  - Joint model with survival

$$\lambda_i(t) = \lambda_0(t) \exp(X_i^T(t)\eta + \gamma_2 a_i + \gamma_3 b_i) \tag{13}$$

- Estimation of total medical costs from these joint models:
  - For the semi-continuous model: $E(\sum_{j=1}^{T_i} Y_{ij})$
  - For the recurrent marker model: $E(\int_0^{T_i} Y_i(t) dN_i(t))$, where $N_i(t)$ is the counting process for recurrent hospital visits
  - More work is needed to derive these quantities.

Informative Dropout vs. Death

- Informative drop-out event stops the follow-up, so that the subsequent measurements are not observable
  - However, it does not change the underlying course of the longitudinal process.
  - Repeated measures and dropout event are independent after conditioning on shared random effects

- A terminal event, such as death, changes the repeated measures process afterwards
  - For example, death precludes further accumulation of medical cost, so the repeated measures of medical cost (e.g., monthly medical cost) are zero after death.
  - For such an outcome, we observe the whole repeated measure process, even after the terminal event.
  - In this situation, repeated measures and survival are not independent, even after conditioning on random effects.
- Ribaudo et al. (2000), Pauler et al. (2003), and Kurland and Heagerty (2005) also distinguished drop-out and death explicitly.
- If both terminal event and informative dropout are present, we can use 3 equations to describe longitudinal measures, informative dropout and death with shared random effects.

Extensions of Joint Models

- Multi-level joint model (Liu et al. 2008): e.g., repeated measures of Hgb and death for subjects clustered within medical centers.
- Include possible nonlinear functional forms in longitudinal and/or survival models (Brown et al. 2005).
- Latent class joint model of longitudinal and survival data (Liu et al., submitted)

- Extension to multivariate processes is currently an active research area.
  - For example, Liu and Huang (2009) used a joint model to study the repeated measures CD4, recurrent opportunistic diseases, and a terminal event death in the CPCRA data set.
  - Prediction of the survival outcome by multivariate longitudinal biomarkers, e.g., BMI, SBP, CAC ...
  - A workshop on the statistical analysis of multi-outcome data in Paris, July 2012, available at http://www.lsta.upmc.fr/SAM2012/Program.html.
  - Second workshop on the statistical analysis of multi-outcome data will be held in Cambridge, UK, July 2014.

Related Topics

- Joint model of recurrent events and a terminal event (Liu et al. 2004, Liu and Huang 2008)
  - Intensity for recurrent events (e.g., recurrent hospitalizations):

$$r_i(t) = r_0(t) \exp(\beta^T Z_i + a_i) \tag{14}$$

  - Hazard for terminal event:

$$\lambda_i(t) = \lambda_0(t) \exp(\alpha^T Z_i + \gamma a_i) \tag{15}$$

- Both recurrent hospitalizations and death are of interest (Goal III).
- The FDA (CBER) has approved the endpoint of time to recurrent heart failure hospitalizations in the presence of terminal event as the primary endpoint for a Phase III study.

Acknowledgements

- R01 HS020263 (Joint PIs: Liu and Shih)
- Collaborators: Drs. Robert Wolfe, John Kalbfleisch, Zhangsheng Yu, Xuelin Huang, Bankole Johnson, Tina Shih, Robert Strawderman, John O'Quigley, Jennie Ma.

References

- Brown et al. (2005). *Biometrics*, **61**, 64-73.
- Dafni and Tsiatis (1998). *Biometrics*, **54**, 1445-1462.
- Guo and Carlin (2004). *American Statistician*, **58**, 16-24
- Henderson et al. (2000). *Biostatistics*, **1**, 465-480.
- Johnson et al. (2011). *American Journal of Psychiatry* **168**, 265-275.
- Lillard and Panis (2003). *aML multilevel multiprocess statistical software*, Version 2.0. Freely available at www.applied-ml.com.
- Liu (2009). *Statistics in Medicine*, **28**, 972-986.
- Liu and Huang. (2008). *Statistics in Medicine* **27**, 2665-2683.
- Liu and Huang. (2009). *Applied Statistics* **58**, 65-81.
- Liu and Liu (2013). Joint models for longitudinal and time-to-event occurrence. Book chapter in the Routledge International Handbook of Advanced Quantitative Methods in Nursing Research.

- Liu et al. (2007). *Statistics in Medicine* **26**, 139-155.
- Liu et al. (2008a). *Biometrics*, **64**, 950-958.
- Liu et al. (2008b). *Statistics in Medicine*, **27**, 5679-5691.
- Liu et al. (2013). *Joint latent class modeling in longitudinal and survival processes.* submitted.
- Prentice (1982). *Biometrika*, **69**, 331-342.
- Ratcliffe et al. (2004). *Biometrics*, **60**, 892-899.
- Regidor et al. (2006). *Journal of the American Society of Nephrology*, **17**, 1181-1191.
- Rizopoulos (2012). The JM package. Available at http://cran.rproject.org/web/packages/JM/JM.pdf.
- Rizopoulos (2011). *Biometrics*, **67**, 819829
- Taylor et al. (2013). *Biometrics*, **69**, 206213.
- Tsiatis and Davidian (2004). *Statistica Sinica*, **14**, 809-834.
- Vonesh et al (2006). *Statistics in Medicine*, **25**, 143-163.

- Wulfsohn and Tsiatis. (1997). *Biometrics*, **53**, 330-339.
- Xu and Zeger (2001). *Applied Statistics*, **50**, 375-387.
- Yu et al. (2004). *Statistica Sinica*, **14**, 835-862.

Thank you!