Missing Data/ Time Dependent Covariates

22 October 2013
Dave Glidden
Biostatistics 210

SCUT Example

- Missing outcome data only
- Baseline + 3 follow-up timepoints
- Inverse weights can "fix up" GEE
- But have patterns of missingness
  not "all or nothing"
- For month 3: can use only month 3 data
  forces data into "all or nothing"

SCUT Example

complete data

```
GEE population-averaged model                   Number of obs      =      1568
Group variable:                        id       Number of groups   =       392
Link:                            identity       Obs per group: min =         4
Family:                          Gaussian                      avg =       4.0
Correlation:                  independent                      max =         4
                                                Wald chi2(3)       =    402.09
Scale parameter:                  .3325843      Prob > chi2        =    0.0000

Pearson chi2(1568):                521.49       Deviance           =    521.49
Dispersion (Pearson):           .3325843        Dispersion         =  .3325843

                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      logmar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vaspoint |
      Week 3 |  -.3191327   .0204863   -15.58   0.000    -.3592851   -.2789802
     Month 3 |   -.467602   .0240345   -19.46   0.000    -.5147088   -.4204953
    Month 12 |  -.5094388   .0260931   -19.52   0.000    -.5605803   -.4582973
             |
       _cons |   .9094388   .0321733    28.27   0.000     .8463803    .9724973
------------------------------------------------------------------------------
```

SCUT Example

available data

```
GEE population-averaged model                   Number of obs      =       662
Group variable:                        id       Number of groups   =       392
Link:                            identity       Obs per group: min =         1
Family:                          Gaussian                      avg =       1.7
Correlation:                  independent                      max =         4
                                                Wald chi2(3)       =    460.71
Scale parameter:                  .3234649      Prob > chi2        =    0.0000

Pearson chi2(662):                 214.13       Deviance           =    214.13
Dispersion (Pearson):           .3234649        Dispersion         =  .3234649

                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      logmar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vaspoint |
      Week 3 |  -.4927721    .039634   -12.43   0.000    -.5704533   -.4150909
     Month 3 |  -.6148736   .0512371   -12.00   0.000    -.7152965   -.5144506
    Month 12 |  -.7569388   .0411782   -18.38   0.000    -.8376465    -.676231
             |
       _cons |   .9094388   .0321733    28.27   0.000     .8463803    .9724973
------------------------------------------------------------------------------
```

Inverse Weighting

- MAR-based method
- Builds a model for missingness
  think propensity score
  predictors: MAR variables
- Calculate probability of non-missing data
  use this as an inverse probability weight
- Conduct a weighted analysis
  "upweights" to correct for non-response

Steps in the Analysis

- Develop logistic model for missingness
- Extract the predicted values
- Make weights inverse of predicted probability
- Weights fixed up to be "1" for baseline timepoint -- *no missingness to correct for*
- Stata must be tricked into GEE
  doesn't permit wts to change within person

The Stata Code

- gen obs = 1-miss
- logistic obs base if vaspoint==3
- predict out, p
- gen weight = 1/out
- replace weight=1 if vaspoint==1
- reg logmar i.vaspoint [pweight=weight] if miss==0, robust

Data Structure

patid	vaspoint	logmar	missing	weight
101C	Baseline	1.3	No	1
101C	Month 3		Yes	
102D	Baseline	1.6	No	1
102D	Month 3		Yes	
103E	Baseline	0	No	1
103E	Month 3	-0.1	No	2.512119
104F	Baseline	1.8	No	1
104F	Month 3		Yes	
105G	Baseline	0.5	No	1
105G	Month 3	0	No	3.432192
106H	Baseline	0.9	No	1
106H	Month 3		Yes	
107J	Baseline	1.2	No	1
107J	Month 3		Yes	
108K	Baseline	0.3	No	1
108K	Month 3		Yes	
109M	Baseline	0.3	No	1
109M	Month 3	0	No	3.011097
110A	Baseline	1.2	No	1
110A	Month 3		Yes	
111B	Baseline	0	No	1

Weighted Results

```
. reg logmar i.vaspoint [pweight=weight] if miss==0, robust
(sum of wgt is   7.8571e+02)

Linear regression                                      Number of obs =     484
                                                       F(  1,   482) =   26.23
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1103
                                                       Root MSE      =  .61456

------------------------------------------------------------------------------
             |               Robust
      logmar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vaspoint |
    Month 3  |  -.4319015   .0843343    -5.12   0.000    -.5976098   -.2661932
       _cons |   .9094388   .0321988    28.24   0.000     .8461714    .9727062
------------------------------------------------------------------------------
```

Weighting Method

- Non-parametric technique for missingness
- Limited applicability
- Works for "all or none missingness" or dropouts
- Better when there are few covariates
  avoids very "wild" weights
- Can be considerably inefficient
- Easy to mess up
- Not a recommended method

Maximum Likelihood

- Models which are based on maximum likelihood can be useful for MAR data
- Models based on ML include
  - logistic regression
  - linear regression
    if the data follows normal distn
  - mixed models
  - just not GEE

xtmixed Results

```
Mixed-effects ML regression                     Number of obs      =       484
Group variable: id                              Number of groups   =       392

                                                Obs per group: min =         1
                                                               avg =       1.2
                                                               max =         2


                                                Wald chi2(1)       =     96.94
Log likelihood = -414.02921                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      logmar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vaspoint |
    Month 3  |  -.4017007   .0407987    -9.85   0.000    -.4816648   -.3217366
       _cons |   .9094388   .0314948    28.88   0.000     .8477101    .9711674
------------------------------------------------------------------------------
```

Relative efficiency is nearly 4!
Weighting would be a massive cut in power

ML and MAR

- Discussed in Sect. 11.9 of VGSM
- ML robustness to MAR is not obvious
- ML is projects values of missing data based on model using EM algorithm
- ML is busy in the background

ML and MAR

- Parametric model must be correct
- Missingness is only in the outcome(s)
  non-response, dropout repeated measures data
- Variables for MAR assumption in the model as predictors or as the outcome
  - May include variables you don't want in model
  - Can use "margins" command

Recommendations

- Investigate reasons for missing data
  gain insights into modeling and assumptions
- Approach using MAR assumption
  MCAR is simply too restrictive
- MI is preferred approach to missingness
  chained equation gives most flexibility
- Missing only data in outcome use ML
  unless you'd need to adjust for mediators then use MI

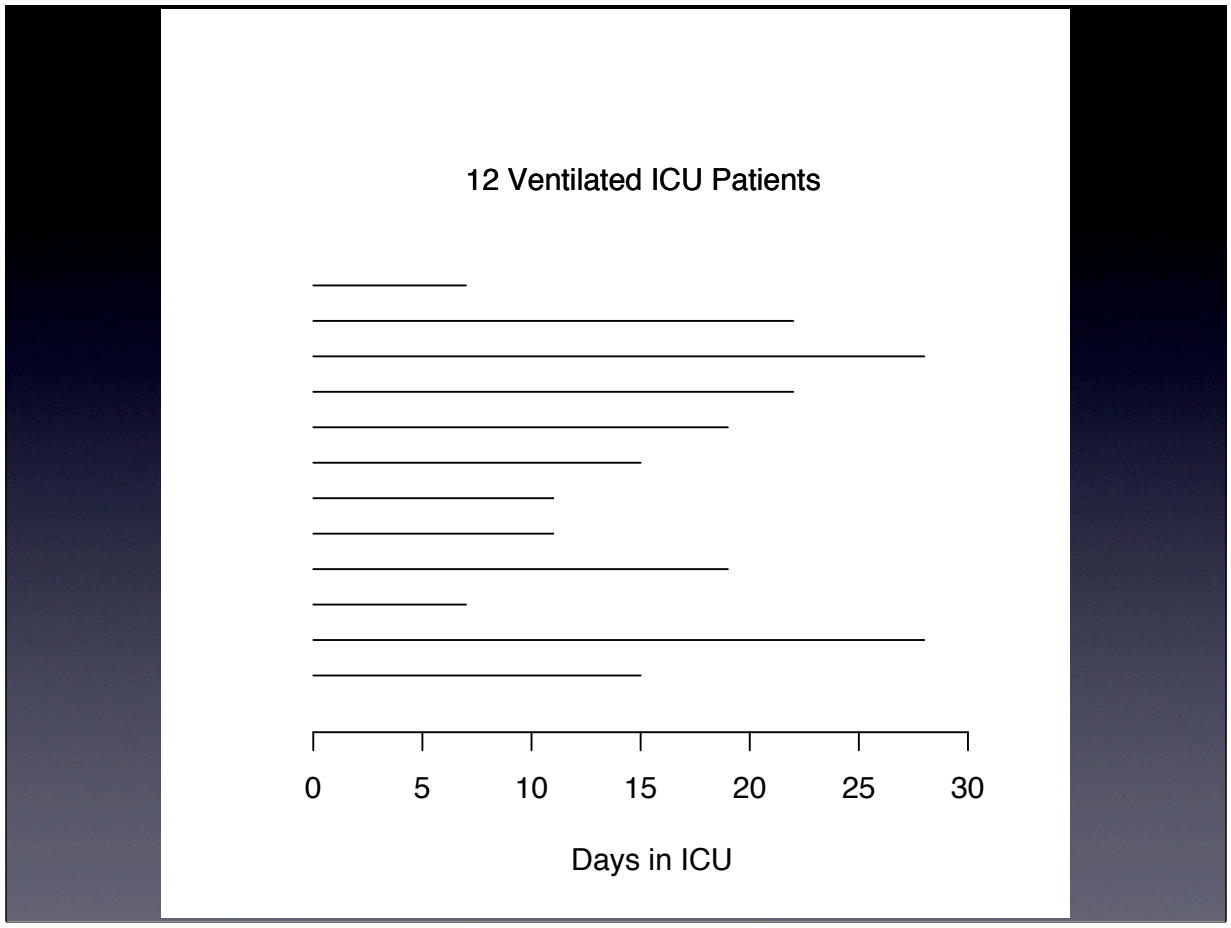
Recommendations

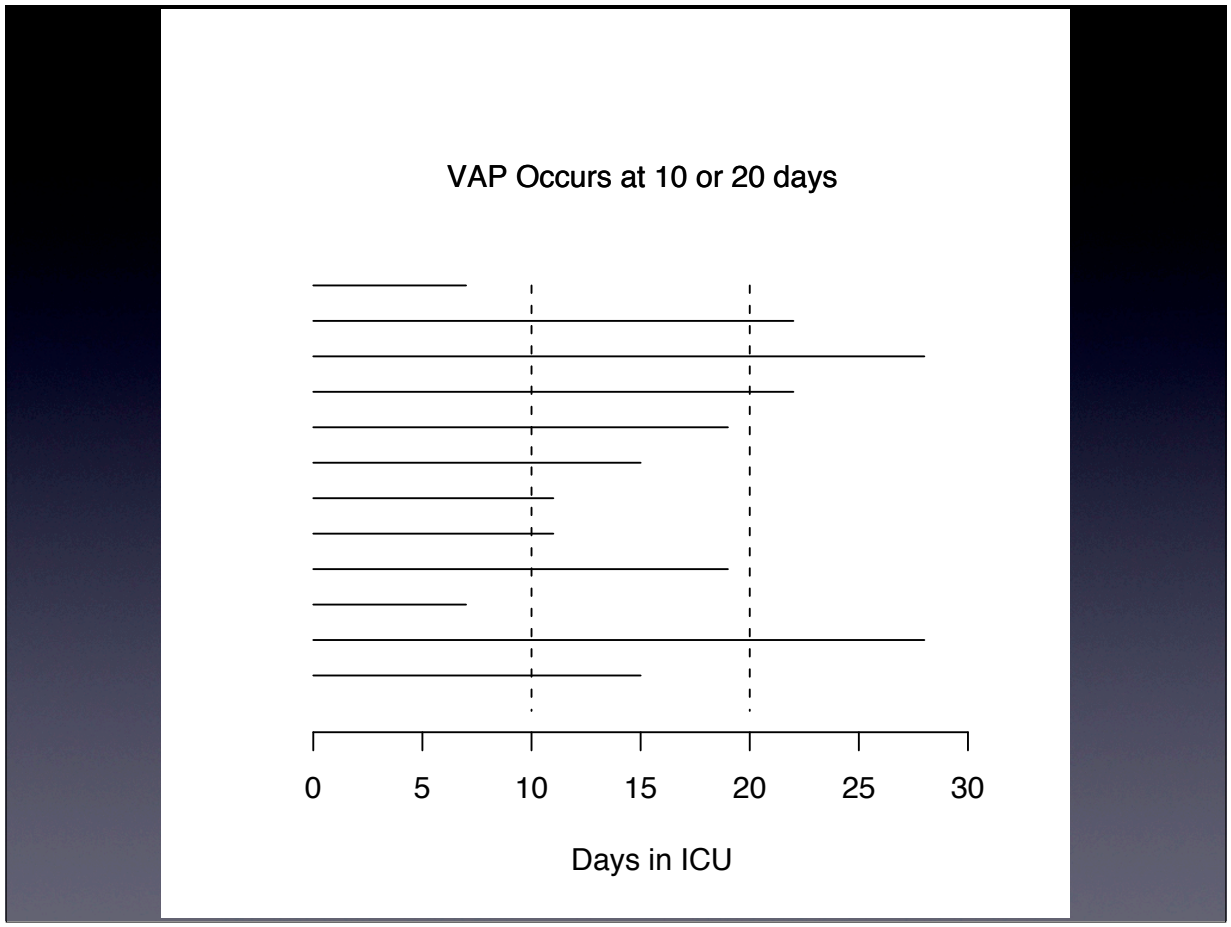
- Missing data in predictors, use MI
- Missing data in only in outcome
  - if MAR vars in stat. model, complete case
    as long as ML based method
  - MAR vars not in stat. model, use MI
  - If want to use GEE, use MI

Time Dependent Covariates

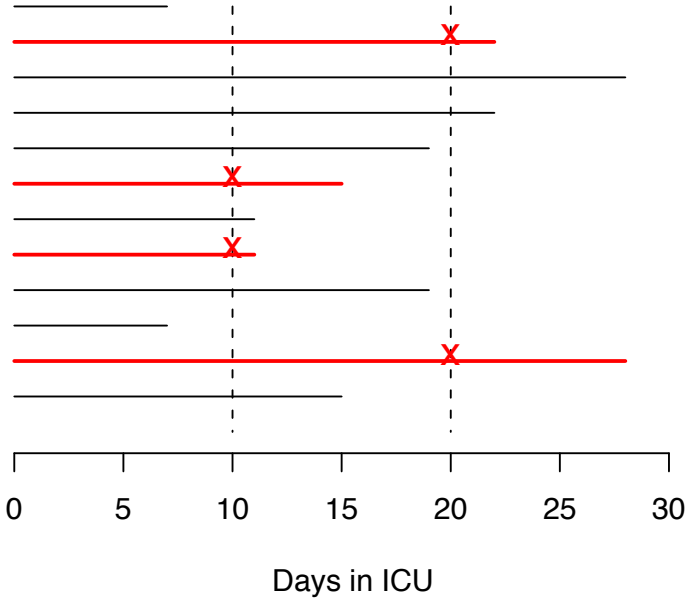
Data Example

- Collaboration with ICU investigators
- Cohort of ventilated patients
- Ventilator assisted pneumonia (VAP)
- Does it extend time on the ventilator?
  Does it increase risk of death?
- VAP develops over clinical course in ICU

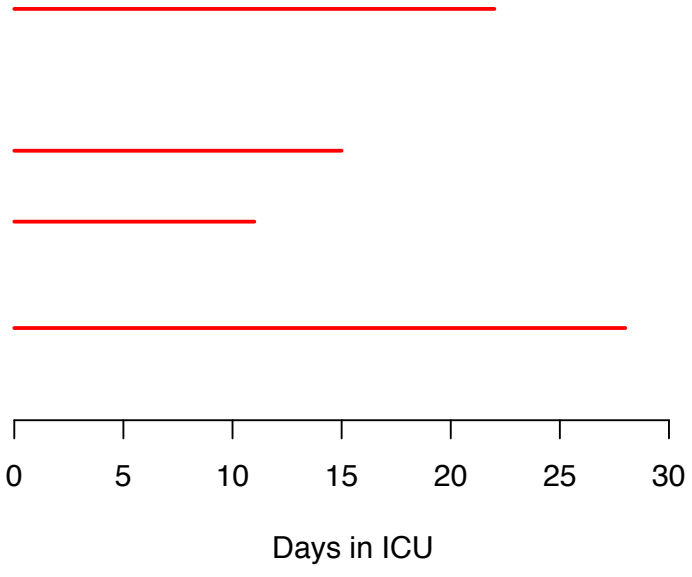

12 Ventilated ICU Patients
0
5
10
15
20
25
30
Days in ICU

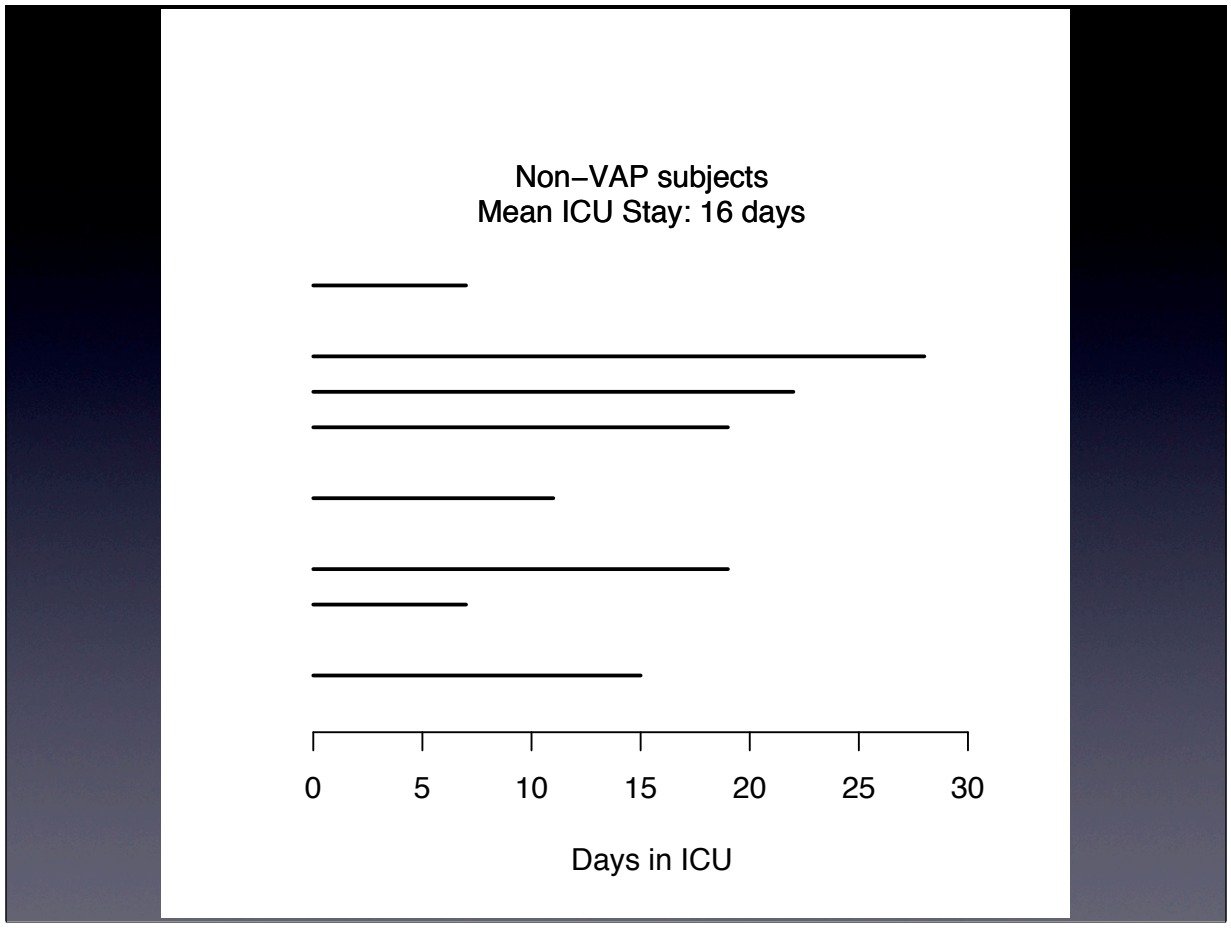

VAP Occurs at 10 or 20 days
0
5
10
15
20
25
30
Days in ICU

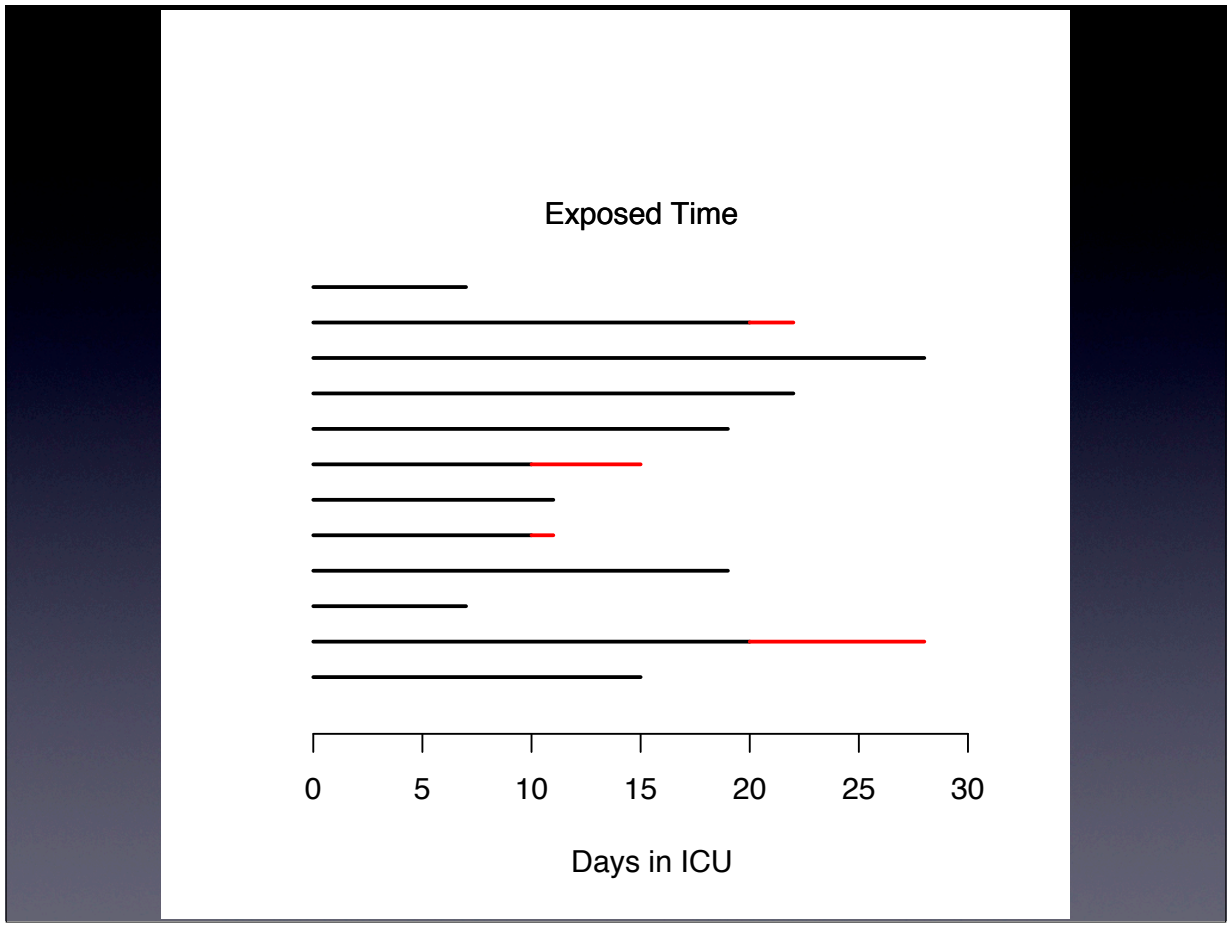
VAP Occurs at 10 or 20 days
Subjects Randomly Selected for VAP



VAP subjects
Mean ICU Stay: 19 days




Non–VAP subjects
Mean ICU Stay: 16 days
0
5
10
15
20
25
30
Days in ICU


Exposed Time
0
5
10
15
20
25
30
Days in ICU

The Bias

- Patients ever exposed to VAP vented longer
- VAP patients: two kinds of vent time
- Time before VAP: longer than non-VAP
  the so-called immortal time
- Time after VAP: same as VAP- subjects
- Should time before VAP count toward VAP+?

Other Examples

- PTLD increase risk of death in kidney tx?
- Does heart tx extend life in listed patients?
- Do OSCAR winners live longer?
- Common thread: exposure occurs sometime after follow-up
- PTLD, transplants, awards occur over time

SCC and voriconazole

- Solid organ transplant recipients
- Immune suppression
- Given voriconazole for fungal prophylaxis
- voriconazole introduced ~10 years ago
- Recent reports: may be associated with onset of squamous cell carcinoma (SCC)

UCSF Retrospective Cohort

- 380 Patients with Lung Transplant at UCSF
- Between 23 Oct '91 and 26 Dec '10
- Followed through mid 2011
- 55 patients developed SCC during FU
  17% in those with voriconazole
  9% in those with no voriconazole

Patient 1123

Example

- Transplanted 4/25/05
- SCC free and alive on 10/31/06
  554 days after
- 4/28/05 - 6/16/05: 49d @ 400mg/day =
  49*400/(400*365) = 0.135
- 9/7/05 -11/10/06: @ 400mg/day
  419 days thru 10/31 = 1.148
- Cumulative exposure = 1.282 year of 400mg
- Need this for every person

Analysis Framework

- Cox model for rate of SCC
- Predictor: exposure to voriconazole
- Bad way:
  - fixed predictor
- Better way
  - break up follow up: td predictor

Fixed Predictor

abpatientcode	time	event	evervori	totaldose	ageattp
1123	837	2	1	1,36	64

Ever Voriconazole

fixed time covariate

```
No. of subjects =          379                     Number of obs   =      8433
No. of failures =           54
Time at risk    =       413146
                                                    LR chi2(1)      =      3.47
Log likelihood  =   -250.63795                      Prob > chi2     =    0.0625

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
evervori_f~d |   1.787881   .5859601     1.77   0.076     .9405156    3.398689
------------------------------------------------------------------------------
```

ever voriconazole versus never

Dose of Voriconazole

fixed time covariate

```
Cox regression -- Breslow method for ties

No. of subjects =          379                     Number of obs    =      8433
No. of failures =           54
Time at risk    =       413146
                                                    LR chi2(1)       =      0.28
Log likelihood  =   -252.23301                      Prob > chi2      =    0.5970

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   totaldose |   1.051435   .0973257     0.54   0.588     .8769829    1.260589
------------------------------------------------------------------------------
```

per year at 400 mg dose

Time-Dependent Covariate

TDC

- Predictors whose value may vary with time
- And measured as a part of the study
- Longitudinal data in a predictor
- Outcome is usually longitudinal
  GEE or a mixed model
- Or as a part of a Cox model

Td predictor

abpatientcode	from	to	event	cumdose	ageattp
1123	0	91	0	0.13	64
1123	91	198	0	0.31	64
1123	198	263	0	0.48	64
1123	263	398	0	0.85	64
1123	398	405	0	0.87	64
1123	405	447	0	0.99	64
1123	447	545	0	1.26	64
1123	545	554	0	1.28	64
1123	554	590	0	1.31	64
1123	590	637	0	1.36	64
1123	637	638	0	1.36	64
1123	638	723	0	1.36	64
1123	723	761	0	1.36	64
1123	761	800	0	1.36	64
1123	800	817	0	1.36	64
1123	817	837	2	1.36	64

The Cox Model

with time-dependent covs

- $h(t|x_1,..,x_p) = h_0(t)\exp(\beta_1 x_1 + ... + \beta_p x_p)$
- Model hazard (like a rate) function
- baseline hazard not specified
- Predict effect acts multiplicatively on hazard
- $h(t) = h_0(t)\exp(\beta_1 x_1 (t) + .. + \beta_p x_p (t))$
  allows the predictors to vary with time

Why use time-dependent covariate?

- Factor or exposure is time varying
- Can increase predictive power
  e.g., viral load in HIV
- Aligning timing of exposures & outcomes
  avoiding immortal time bias

Defining Exposure

- Created exposure variable:
  x(t) dose-days of voriconazole through t
- Acknowledges that exposure changes
- Time prior to exposure, x(t)=0
- After exposure accumulates through t
- If exposure stops: x(t) doesn't increase

Time-Dependent Covariate

- Data management is biggest hurdle
- Need a long record dataset for each person
- Need know for series of time periods
  - Did person get SCC or get censored?
  - What was value of td cov for that period?
- Series of time points -- observed times of SCC

Td predictor

abpatientcode	from	to	event	cumdose	ageattp
1123	0	91	0	0.13	64
1123	91	198	0	0.31	64
1123	198	263	0	0.48	64
1123	263	398	0	0.85	64
1123	398	405	0	0.87	64
1123	405	447	0	0.99	64
1123	447	545	0	1.26	64
1123	545	554	0	1.28	64
1123	554	590	0	1.31	64
1123	590	637	0	1.36	64
1123	637	638	0	1.36	64
1123	638	723	0	1.36	64
1123	723	761	0	1.36	64
1123	761	800	0	1.36	64
1123	800	817	0	1.36	64
1123	817	837	2	1.36	64

Cox Model w/ TD Cov

- Patient #37: SCC 554d after transplant
  received 1.01 years of voriconazole
- Think of him as a "case"
  to be compared to a set of controls
  matched on follow-up
- More vori exposure than typical case?
- How much vori in those SCC-free 554 days?

stset

- stset to, failure(event==1) id(abpatientcode)
- "from" is not need but helps me keep track
- event: 0 censored, 1 SCC, 2 death
  if want Cox on SCC, set event to 1
- Must use id for multiple events data
- Then just apply Cox
- Not good for anything else!
  multiple records foul all else up

Ever Voriconazole

as a time-dependent covariate

```
stcox evervori_td

Cox regression -- Breslow method for ties

No. of subjects =          379                     Number of obs   =      8433
No. of failures =           54
Time at risk    =       413146
                                                    LR chi2(1)      =      8.53
Log likelihood  =    -248.1054                      Prob > chi2     =    0.0035

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 evervori_td |   2.444013   .8028329     2.72   0.007     1.283786      4.6528
------------------------------------------------------------------------------
```

Dose of Voriconazole

as a time-dependent covariate

```
stcox cumdose

No. of subjects =          379                     Number of obs   =      8433
No. of failures =           54
Time at risk    =       413146
                                                    LR chi2(1)      =      5.93
Log likelihood  =    -249.4083                      Prob > chi2     =    0.0149

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cumdose |   1.311188   .1346107     2.64   0.008     1.072205    1.603438
------------------------------------------------------------------------------
```

per year at 400 mg dose

TD Covs Eliminates ITB

- Behind the scenes it's implementing the matching approach
  comparing comparable follow-up
  looking at effects after exposure
- But using all the data in a flexible way
  no artificial matches
  all available data used

Sparsity

- HIV chemoprophylaxis
  with anti-viral drug emtricitabine+tenofovir
- Risk of HIV high associated with levels
  pill taking fluctuates, drugs levels per 24w
- Week 24: 10.5 fmol/10^6 cells
- Week 48: 10.5 fmol/10^6 cells
- What is the level at Weeks 32, 54?

Paradox

reverse causation in smoking exposure

- Coronary artery surgery study
  smoking status per 6 months
- Current smoking: reduced risk of death
  not statistically significant
- Most deaths among smokers
  but had stopped smoking just prior to death
- Hospitalization/CHF lead quitting before death
- Their soln: lag smoking by 6 month

Histories

- BCC study: voriconazole hx over many years
- Which part is relevant for risk of BCC?
  current voriconazole
  current dose
  total accumulated voriconazole
  accumulated exposure in last year
- Combine subject-matter knowledge
  and can see what fits better

Ever vs. Cumulative Voriconazole

```
. stcox cumd evervori_td

No. of subjects =          379                     Number of obs   =      8433
No. of failures =           54
Time at risk    =       413146
                                                    LR chi2(2)      =      9.72
Log likelihood  =   -247.51186                      Prob > chi2     =    0.0077

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cumdose |   1.155081   .1469683     1.13   0.257     .9001372    1.482232
 evervori_td |   2.026276   .7556136     1.89   0.058     .9756167     4.20841
------------------------------------------------------------------------------
```

ever vs. cumulative – ever appears stronger

Three Things to Think Of

1. May be gaps in exposure (sparsity)
2. Can introduce additional confounding
   some td covs can even be mediators
3. Have to distill a complex history
   rich modeling choices

Baseline Covariates

- Osteoporosis study
- Baseline BMD predicts fracture over 3 years
- That, in itself, is powerful
- Prognostic over many years powerful
- Baseline and td cov analyses complimentary
- A baseline value can be just fine

Recognizing ITB

- Carefully examine when and how people enter a "group" or accrue their exposure
- Is their a time-related requirement?
  if yes, ITB
- Based on classification @ time 0 or later?
  if later, ITB
- TD covs can account for this

TD Covs

- Clarifying exposure and events
- Addresses immortal time bias
- Dataset construction can be tedious
- Fitting model is easy
- Mindful of inverse associations and mediators
- Rich choices for incorporating hx