Modelling the association between patient characteristics and the change over time in a disease process using observational cohort data

Linda Harrison, David Dunn, Hannah Green and Andrew Copas

HIV Group and Hub for Trials Methodology Research, MRC Clinical Trials Unit, London, UK
Objective

- **outcome** $Y$
  - measured at baseline and then repeatedly over follow-up in an observational study

- **patient characteristic** $X$
  - $X$ may be related to $Y$ at baseline
  - Subsequent change in $Y$ is likely to be related to $Y$ at baseline

Is $X$ related to change in $Y$ conditional on baseline $Y$?
The Data

• outcome $Y = CD4$ count
  ▪ measured regularly from start of HIV therapy (baseline) in an observational study

• patient characteristic $X = HIV$ subtype:
  ▪ B [n=1550]
  ▪ C [n=272]

• to study effect of subtype on $CD4$ to starting HIV therapy

• Subtype may be related to $CD4$ at start of HIV therapy (baseline). Subsequent change in $CD4$ is likely to be related to $CD4$ at start of HIV therapy (baseline).
Plot of average change by subtype

\( \sqrt{\text{CD4 (count/mm}\^3)} \) vs. time (months)
Plot of average change by subtype

\[ \sqrt{CD4 (count/mm^3)} \]

\[ \sqrt{time (\sqrt{month})} \]

- **B**
- **C**
Baseline CD4 by Subtype
Confounding: Effect of Correlation between Baseline and Change

\[ \sqrt{\text{CD4 (count/mm}^3\text{)}} \text{ vs. } \sqrt{\text{time (month)}} \]

Lines B and C represent different correlation scenarios.
Confounding: Effect of Correlation between Baseline and Change
Confounding: Effect of Correlation between Baseline and Change

\[
\sqrt{CD4} \text{ (count/mm}^3) \quad \sqrt{\text{time (month)}}
\]

- Line B
- Line C
Confounding: Effect of Correlation between Baseline and Change
Notation & Initial Model

\[ Y_{it} = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_{i0} + t b_{i1} + \varepsilon_{it} \]

- patient \( i \), time \( t \)
- random effects
  \[
  \begin{pmatrix}
  b_{i0} \\
  b_{i1}
  \end{pmatrix}
  \sim N
  \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix},
  \begin{pmatrix}
  \sigma_0^2 & 0 \\
  0 & \sigma_1^2
  \end{pmatrix}\right)
  \]
- measurement error
  \( \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2) \)

- \textit{xtmixed in Stata}
- \( \beta_3 \) is most interesting
  - does \( X \) affect the change in \( Y \) over time?
Model for Change Based on Observed $Y_0$

- Formulate a model for change in $Y$, and condition on observed $Y_0$

$$ Y_{it} - Y_{i0} = t \beta_2^C + X_i t \beta_3^C + Y_{i0} t \beta_5^C + b_{i0}^C + t b_{i1}^C + \varepsilon_{it}^C $$
Model for Change Based on Observed $Y_0$

- Formulate a model for change in $Y$, and condition on observed $Y_0$

$$Y_{it} - Y_{i0} = t \beta_2^C + X_i t \beta_3^C + Y_{i0} t \beta_5^C + b_{i0}^C + t b_{i1}^C + \varepsilon_{it}^C$$
• Formulate a model for change in $Y$, and condition on observed $Y_0$

$$Y_{it} - Y_{i0} = t\beta_2^C + X_it\beta_3^C + Y_{i0}t\beta_5^C + b_i^C + tb_{i1}^C + \epsilon_{it}^C$$
Model for Change Based on Observed $Y_0$

- Formulate a model for change in $Y$, and condition on observed $Y_0$

$$Y_{it} - Y_{i0} = t \beta_2^C + X_i t \beta_3^C + Y_{i0} t \beta_5^C + b_{i0}^C + t b_{i1}^C + \epsilon_{it}^C$$

- Should get approximate inference for $\beta_3^C$ conditional on $Y_0$
Adjusted Parameters from Initial Model

- Initial model:

\[ Y_{it} = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_{i0} + t b_{i1} + \epsilon_{it} \]
Adjusted Parameters from Initial Model

• ‘underlying’ model without measurement error:

\[ Y_{it}^* = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_{i0} + tb_{i1} \]
Adjusted Parameters from Initial Model

• ‘underlying’ model without measurement error:

\[ Y_{it}^* = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_{i0} + tb_{i1} \]

• Look at ‘underlying change’:

\[ Y_{it}^* - Y_{i0}^* = t \beta_2 + X_i t \beta_3 + tb_{i1} \]
Adjusted Parameters from Initial Model

• ‘underlying’ model without measurement error:

\[ Y_{it}^* = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_{i0} + t b_{i1} \]

• Look at ‘underlying change’:

\[ Y_{it}^* - Y_{i0}^* = t \beta_2 + X_i t \beta_3 + t b_{i1} \]

• Condition on ‘underlying baseline’ to remove the confounding:

\[ E(Y_{it}^* - Y_{i0}^* | X_i, Y_{i0}^*) = t \beta_2 + X_i t \beta_3 + E(tb_{i1} | b_{i0} = Y_{i0}^* - \beta_0 - X_i \beta_i) \]
Adjusted Parameters from Initial Model

\[ E(Y_{it}^* - Y_{i0}^* | X_i, Y_{i0}^*) = t(\beta_2 - \beta_0 \gamma) + X_i t(\beta_3 - \beta_1 \gamma) + Y_{i0}^* t \gamma \]

where \( \gamma = \text{cov} (b_{i0}, b_{i1}) / \text{var} (b_{i0}) \)
Adjusted Parameters from Initial Model

\[
E(Y_{it}^* - Y_{i0}^* \mid X_i, Y_{i0}^*) = t(\beta_2 - \beta_0 \gamma) + X_i t(\beta_3 - \beta_1 \gamma) + Y_{i0}^* t \gamma
\]

where \( \gamma = \text{cov} (b_{i0}, b_{i1}) / \text{var} (b_{i0}) \)

- ‘adjusted parameter’ \( \beta_3^A = \beta_3 - \beta_1 \gamma \) is of real interest
- So we estimate \( \beta_1, \beta_3 \) and \( \gamma \) from the initial, and then calculate
  \[
  \widehat{\beta}_3^A = \widehat{\beta}_3 - \widehat{\beta}_1 \gamma
  \]
- Use the delta method to get standard error
- Perform Wald tests and generate confidence intervals (nlcom in Stata)
- Produce plots of predicted trajectory for patients by \( X \) with common \( Y_0 \)
Plot of average change by subtype

√CD4 (count/mm³) vs. √time (month)

- B
- C
## Application to Our Data

<table>
<thead>
<tr>
<th>Initial Model</th>
<th>coefficient (se)</th>
<th>p-value</th>
<th>Change Model</th>
<th>coefficient (se)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (const)</td>
<td>15.13 (0.14)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$ (subtype)</td>
<td>-2.32 (0.36)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_2$ ($\sqrt{\text{time}}$)</td>
<td>1.16 (0.03)</td>
<td>&lt;0.001</td>
<td>$\beta_2^C$ ($\sqrt{\text{time}}$)</td>
<td>2.73 (0.05)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$\beta_3$ (interact.)</td>
<td><strong>0.19</strong> (0.07)</td>
<td>0.007</td>
<td>$\beta_3^C$ (interact.)</td>
<td><strong>-0.07</strong> (0.06)</td>
<td>0.20</td>
</tr>
<tr>
<td>$\beta_3^A$</td>
<td><strong>-0.02</strong> (0.06)</td>
<td>0.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>se ($b_0$)</td>
<td>5.13 (0.10)</td>
<td></td>
<td>se ($b_0^C$)</td>
<td>3.90 (0.09)</td>
<td></td>
</tr>
<tr>
<td>se ($b_1$)</td>
<td>0.87 (0.02)</td>
<td></td>
<td>se ($b_1^C$)</td>
<td>0.91 (0.02)</td>
<td></td>
</tr>
<tr>
<td>corr ($b_0$, $b_1$)</td>
<td>-0.52 (0.02)</td>
<td></td>
<td>corr ($b_0^C$, $b_1^C$)</td>
<td>-0.57 (0.02)</td>
<td></td>
</tr>
<tr>
<td>se ($\varepsilon_t$)</td>
<td>2.85 (0.01)</td>
<td></td>
<td>se ($\varepsilon_t^C$)</td>
<td>2.78 (0.01)</td>
<td></td>
</tr>
</tbody>
</table>
Predicted Trajectories

Initial Model Trajectories

[count/mm³]

time (months)
Predicted Trajectories

Initial Model Trajectories

Initial Model Trajectories, Same Baseline
Predicted Trajectories

Initial Model Trajectories

Initial Model Trajectories, Same Baseline

'Change Model' Trajectories
Predicted Trajectories

- **Initial Model Trajectories**
  - B
  - C

- **Initial Model Trajectories, Same Baseline**

- **'Change Model' Trajectories**

- **Trajectories Based on Adjusted Parameters**
Simulation study - setup

- 10,000 simulations of the dataset
- In each dataset; 1,000 patients $X=0$, 1,000 patients $X=1$
- Each patient 7 equally spaced observations of $Y$ (at $t=0, 8, 16, 24, 32, 40$ and 48 months)

- Model: $Y_{it} = \beta_0 - \beta_1 X_i + \beta_2 \sqrt{t} + \beta_3 X_i \sqrt{t} + b_{i0} + b_{i1} \sqrt{t} + \varepsilon_{it}$
  $se(b_0)=5$, $se(b_1)=1$, $corr(b_0,b_1)=-0.5$, $se(\varepsilon_t)=2.8$

- Scenario 1: $\beta_3=0.24 \Rightarrow \beta_3^A = 0$
- Scenario 2: $\beta_3=0.44 \Rightarrow \beta_3^A = 0.2$
## Simulation study - results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Initial Model</th>
<th>Change Model</th>
<th>Adjusted Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_3^A = 0$</td>
<td><strong>mean estimate (SE)</strong> 0.24 (0.05)</td>
<td><strong>-0.04 (0.04)</strong></td>
<td><strong>0.0005 (0.04)</strong></td>
</tr>
<tr>
<td></td>
<td><strong>coverage</strong> 0.2%</td>
<td><strong>86.6%</strong></td>
<td><strong>95.3%</strong></td>
</tr>
<tr>
<td></td>
<td><strong>mse</strong> 0.06</td>
<td><strong>0.003</strong></td>
<td><strong>0.002</strong></td>
</tr>
<tr>
<td>$\beta_3^A = 0.2$</td>
<td><strong>mean estimate (SE)</strong> 0.44 (0.05)</td>
<td><strong>0.16 (0.04)</strong></td>
<td><strong>0.20 (0.04)</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Coverage</strong> 0.2%</td>
<td><strong>86.1%</strong></td>
<td><strong>95.1%</strong></td>
</tr>
<tr>
<td></td>
<td><strong>mse</strong> 0.06</td>
<td><strong>0.003</strong></td>
<td><strong>0.002</strong></td>
</tr>
</tbody>
</table>
Conclusions 1

• Confounding by baseline value can be substantial. e.g. in our data the corr(b₀,b₁) was strong -0.5

• The difference in the average CD4 trajectories over time by subtype was partly due to differences in CD4 at the start of therapy

• Our proposed method is relatively simple [xtmixed followed by nlcom in Stata] leads to confidence intervals, tests and plots
Conclusions 2

• The ‘change model’ with $Y_0$ as a covariate is simpler and performs much better than the initial model without adjustment

• But, its performance was not as good

• Our proposed method can be extended to included
  ▪ multiple covariates and time components
  ▪ non-linear link functions
Thank you for listening

- Any Questions?
Adjusted Parameters from Initial Model

\[ Y_{it} = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_i 0 + tb_{i1} + \varepsilon_{it} \]

• ‘underlying \( Y \) without measurement error:
  \[ Y^*_{it} = \beta_0 + X_i \beta_1 + t \beta_2 + X_i t \beta_3 + b_i 0 + tb_{i1} \]
  \[ Y^*_{i0} = \beta_0 + X_i \beta_1 + b_i 0 \]

• Look at ‘underlying change’ without measurement error:
  \[ Y^*_{it} - Y^*_{i0} = t \beta_2 + X_i t \beta_3 + tb_{i1} \]

• Condition on ‘underlying baseline’ to remove the confounding:
  \[ E(Y^*_{it} - Y^*_{i0} | X_i, Y^*_{i0}) = t \beta_2 + X_i t \beta_3 + E(tb_{i1} | b_i 0 = Y^*_{i0} - \beta_0 - X_i \beta_i) \]
Adjusted Parameters from Initial Model

\[ E(Y_{it}^* - Y_{i0}^* | X_i, Y_{i0}^*) = t\beta_2 + X_i t \beta_3 + E(tb_{i1} | b_{i0} = Y_{i0}^* - \beta_0 - X_i \beta_i) \]

- By a property of the normal distribution

\[ E(tb_{i1} | b_{i0} = Y_{i0}^* - \beta_0 - X_i \beta_i) = (Y_{i0}^* - \beta_0 - X_i \beta_i) t \gamma \]

where \( \gamma = \text{cov}(b_{i0}, b_{i1}) / \text{var}(b_{i0}) \)

\[ E(Y_{it}^* - Y_{i0}^* | X_i, Y_{i0}^*) = t\beta_2 + X_i t \beta_3 + (Y_{i0}^* - \beta_0 - X_i \beta_1 ) t \gamma \]

\[ = t(\beta_2 - \beta_0 \gamma) + X_i t(\beta_3 - \beta_1 \gamma) + Y_{i0}^* t \gamma \]

\[ \beta_3^A = \beta_3 - \beta_1 \gamma \]
Multiple Covariates & Time Effects

- vector of covariates $X$, vector of time components $t$

$$Y_{it} = \beta_0 + X_i'\beta_1 + t'\beta_2 + (X_i' \otimes t)'\beta_3 + b_{i0} + t'b_{i1} + \varepsilon_{it}$$

- adjusted parameter of interest

$$\beta_3 - (\beta_1 \otimes \gamma)$$
Inference Conditional on Observed $Y_0$

- to assess prognosis for patients given their $X$ and $Y_0$, we are interested in $E(Y_{it}^* - Y_{i0}^* | X_i, Y_{i0})$

- we can follow earlier arguments to see that:

$$E(Y_{it}^* - Y_{i0}^* | X_i, Y_{i0}) = t(\beta_2 - \rho \beta_0 \gamma) + X_i t(\beta_3 - \rho \beta_1 \gamma) + \rho Y_{i0}^* t \gamma$$

where $\rho = \sigma_0^2 / (\sigma_0^2 + \sigma_\varepsilon^2)$
More complex covariance structures

• assumed that \( \text{var}(Y_i) \) can
  a. be ‘captured’ by random intercepts and a limited number of random ‘time effects’
  b. covariance is independent of \( X \)

• If not true - there is no longer a single adjusted parameter \( \beta_3^A \) to determine whether \( X \) is linked to underlying change

• But, plots of the trajectory can be examined for different \( X \) and \( Y_0^* \)