Virtual Twins method for estimating long-term treatment effects from a short-term clinical trial with an active-arm extension

- Background, trial extension designs
- Alternative analysis methods
- Virtual twins method
- Application to FIT/FLEX
- Model misspecification
- Simulation study
- Conclusions

Two disclosures: Chuck McCulloch’s idea, salary support from Amgen
Background

• Treatment shown to work well in short-term placebo-controlled trial

• But would it work as well, and benefits outweigh adverse effects, in long-term use?

• Long-term placebo-controlled trials infeasible for ethical, practical reasons

• Evidence restricted to short-term placebo-controlled trials with various longer-term extensions
Extension designs - I

<table>
<thead>
<tr>
<th>trial</th>
<th>extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>active</td>
<td>none</td>
</tr>
<tr>
<td>placebo</td>
<td>none</td>
</tr>
</tbody>
</table>

- IBIS, BCPT (tamoxifen for breast cancer prevention): blinded follow-up, no one remained on active treatment
- Captures long-term effects of short-term treatment, but not effects of long-term treatment
### Extension designs - II

<table>
<thead>
<tr>
<th>trial</th>
<th>extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>active</td>
<td>active</td>
</tr>
<tr>
<td>placebo</td>
<td>none</td>
</tr>
</tbody>
</table>

- MARS (lovastatin for atherosclerosis): double-blind extension, 58% participation; stopped by DSMB

- HERS (hormone therapy for secondary prevention of CHD):
  - in trial, no overall treatment effect – early harm, late benefit
  - participants unblinded, encouraged not to change treatment
  - 93% participation, little crossover in 2.8 year extension; no long-term benefit
Extension designs - III

<table>
<thead>
<tr>
<th>trial</th>
<th>extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>active</td>
<td>active</td>
</tr>
<tr>
<td>placebo</td>
<td>active</td>
</tr>
</tbody>
</table>

• Most common design

• Analyses focus on within-group changes from baseline in active→active group

• Examples: bisphosphonates (Bone et al., NEJM; Tonino et al., JCEM), clozapine in late Parkinson’s, galantamine for dementia, possibly iPrEx pre-exposure prophylaxis trial
Virtual Twins pre-cursor

• Galantamine for treatment of Alzheimer’s

• Placebo group crossed over to active for extension period

• Long-term placebo response projected forward, using baseline scores of placebo group and published prediction equations based on historical data
Virtual Twins pre-cursor
Extension designs - IV

<table>
<thead>
<tr>
<th>trial</th>
<th>extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>active</td>
<td>active/placebo</td>
</tr>
<tr>
<td>placebo</td>
<td>discontinued</td>
</tr>
</tbody>
</table>

- **FIT/FLEX**
- Randomized active/placebo comparison in extension shows whether it is better to continue on treatment than to stop
  - but not whether it is better to have started in the first place
- Active→placebo group not informative for long-term effects
Extension designs - summing up

• Most include long-term active treatment

• Almost none include long-term placebo
  – can’t control for long-term placebo, secular effects

• Extension usually requires re-consent
  – can entail considerable dropout, selection bias
Naive estimators of long-term treatment effects

- Assume short-term effect holds long-term

- Assume short-term effect holds long-term, provided trial and extension outcome rates equivalent in active arm

- FIT/FLEX design: estimate late treatment effect by active/placebo comparison in extension, combine with short-term effect from trial

- Project placebo forward: estimate late treatment effect by comparing active in extension to placebo in trial, combine with short-term effect from trial
Virtual Twins estimator

- For each active-treatment volunteer in the extension, model expected outcomes for a virtual twin with the same prognostic covariates, under counterfactual assignment to placebo.
- Parameters of outcome models estimated using data from placebo group.
- Expected outcomes for each twin calculated using those parameter estimates, covariate values for volunteer.
- Treatment effects estimated by average difference (or ratio) of observed outcomes for volunteers, expected outcomes for virtual twins.
- Bootstrap used to calculate confidence intervals.
Controlling selection effects

• Re-consent for extension opens door to selection bias

• Controlled by computing expected outcomes for twins using prognostic covariates for volunteers

• Assumes that measured covariates adequately capture selection effects

• Inference restricted to volunteer group

12
Controlling secular effects

- Placebo outcome rates might change during extension
- Controlled by updating covariates for twins:
  - use observed end-of-trial values for volunteers if unaffected by treatment
  - otherwise simulate end-of-trial values:
    1. fit models for end-of-trial values in placebo group
    2. simulate values for twins using model parameter estimates, baseline covariates for volunteers
## Counterfactual framework of extension data

<table>
<thead>
<tr>
<th>Group</th>
<th>Volunteer for Extension</th>
<th>Proportion of Group</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trial</td>
</tr>
<tr>
<td>Treatment</td>
<td>Yes</td>
<td>$p_{v,T}$</td>
<td>$Y_{1,T}^v$</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$1 - p_{v,T}$</td>
<td>$Y_{1,T}^{nv}$</td>
</tr>
<tr>
<td>Placebo</td>
<td>Yes</td>
<td>$p_{v,P}$</td>
<td>$Y_{1,P}^v$</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$1 - p_{v,P}$</td>
<td>$Y_{1,P}^{nv}$</td>
</tr>
</tbody>
</table>

- highlights show what we don’t observe in FIT/Flex design
If we had complete data ...

- Effect of treatment during trial is

\[ [p_{v,T}\mu_{1,T}^v + (1 - p_{v,T})\mu_{1,T}^{nv}] - [p_{v,P}\mu_{1,P}^v + (1 - p_{v,P})\mu_{1,P}^{nv}], \]

where \( \mu_{1,T}^v = E[Y_{1,T}^v] \), and so on

- Treatment effect during the extension is

\[ [p_{v,T}\mu_{2,T}^v + (1 - p_{v,T})\mu_{2,T}^{nv}] - [p_{v,P}\mu_{2,P}^v + (1 - p_{v,P})\mu_{2,P}^{nv}] \]

- Long-term effects of treatment estimated by (weighted) average of trial and extension differences
In absence of complete data ...

- Since extension responses for non-volunteers are never observed, we can at best estimate treatment effects in volunteers: $\mu_{1,T}^v - \mu_{1,P}^v$ and $\mu_{2,T}^v - \mu_{2,P}^v$

- Virtual Twins method estimates the expected counterfactual placebo responses of each observed active-treatment volunteer

- Two potential problems to resolve:
  - Selection effects: $\mu_{i,P}^v \neq \mu_{i,nv}^v$?
  - Secular, cohort effects: $\mu_{2,P}^v \neq \mu_{1,P}^v$?
Some notation

- \( \mathbf{X}^i_1 \): vector of prognostic baseline covariates that are fixed or not affected by treatment

- \( \mathbf{X}^d_1 \): other prognostic baseline covariates that are subsequently affected by treatment

- \( \mathbf{X}^i_2 \) and \( \mathbf{X}^d_2 \): updated values of baseline covariates, observed or counterfactual, at start of extension

- \( \mathbf{Z} \): additional baseline covariates that influence \( Y_{2,P}^v \) only through their effects on \( \mathbf{X}^d_2 \)
Model for virtual twin responses during the trial

- Assume placebo responses during trial arise from GLM with conditional mean

\[
E[Y_{1,P}|X^i_1, X^d_1] = g^{-1}[\beta_0 + \beta_1 X^i_1 + \beta_2 X^d_1]
\]  

- Model (1) assumed to hold for all placebo participants: i.e., \(X^i_1\) and \(X^d_1\) capture any dependence of response on volunteering for the extension (selection effects)

- In type III design, \(\beta\) could be estimated using data for placebo volunteers only
Complete data model for virtual twin responses during the extension

- Assume that during the extension, counterfactual placebo responses arise from the same GLM as (1), but with conditional mean

\[ E[Y_{2,P}^v | X_{2i}^i, X_{2d}^d, Z] = g^{-1}[\beta_0 + \beta_1 X_{2i}^i + \beta_2 X_{2d}^d] \]  (2)

- Link function, parameters shared by (1) and (2)

- Equivalently: differences in conditional means of \( Y_{1,P}^v \) and \( Y_{2,P}^v \) due to secular, cohort effects completely captured by changes in \( X^i \) and \( X^d \)
Prognostic variables at start of extension

• Use observed values of $X^i_2$ for volunteers in evaluating (2) for virtual twins (values are not affected by treatment)

• Model counterfactual values of $X^d_2$, assuming each element arises from GLM with conditional mean

$$E[X^d_{2j}|Z_j] = h^{-1}_j[\gamma_{0j} + \gamma_{1j}Z_j], \ j = 1, \ldots, J. \quad (3)$$

• $\gamma$ estimated using placebo data during trial

• $Z$ may include elements of $X^d_1$ (i.e., baseline values of the treatment-affected covariates used as predictors)
Observed data model for virtual twin responses during the extension

- Combining (2) and (3), we obtain

\[
E[Y_{2,v,P}|X_2^i, Z] = \int \cdots \int g^{-1}[\beta_0 + \beta_1 X_2^i + \beta_2 X_2^d] dF_1 \cdots dF_J, \tag{4}
\]

where \(dF_j\) is conditional density of \(X_{2j}^d\) given \(Z_j\), consistent with (3)

- (4) estimated using Monte Carlo integration:
  - sample \(X_2^d\) from (3) using \(\hat{\gamma}\)
  - compute \(E[Y_{2,v,P}|X_2^i, X_2^d, Z]\) using (2) and \(\hat{\beta}\)
  - repeat and average the results
Effects of treatment among volunteers

- If assumptions hold, estimate treatment effect during trial by averaging

\[ Y_{1,T}^v - E[Y_{1,P}^v | X_{i1}^i, X_{i1}^d] \]

over the volunteer sample

- Similarly, estimate treatment effect during extension by averaging

\[ Y_{2,T}^v - E[Y_{2,P}^v | X_{i2}^i, Z] \]

again over the volunteers

- Estimate long-term effect of treatment by weighted average of trial and extension effects
Bootstrap CIs

- Variability arises from
  - sampling of volunteers
  - sampling of placebo participants used to estimate model parameters

- Solution: resample with replacement from volunteers and placebo group, re-run procedure on each bootstrap sample

- Compute confidence bounds as percentiles of bootstrap effect estimates

- Compute point estimate as mean of effect estimates, averaging over simulations of end-of-trial covariates
Fracture Intervention Trial (FIT)

- Two large RCTs of Alendronate (ALN) for prevention of fractures

- Vertebral fracture trial:
  - 2027 post-menopausal women with existing vertebral fracture (VFx)
  - randomized 1-1 to ALN or placebo
  - 2.9-year average follow-up
  - primary endpoint: new morphometric VFx
  - results (N=1946): RR 0.53, 95% CI 0.41-0.68
FIT

• Clinical fracture trial:
  – 4432 post-menopausal women with low BMD
  – randomized 1-1 to ALN or placebo
  – 4.2-year average follow-up
  – primary endpoint: clinical fracture
  – results:
    * overall (N=4272): RR 0.86, 95% CI 0.73-1.01
    * T-score < −2.5: RR 0.64, 95% CI 0.50-0.82
FLEX - the extension

• At end of FIT, one year of ALN offered to all participants; outcomes not ascertained in this interval

• After interim open-label period (average 1.9 years), new trial of 3 additional years of ALN or placebo

• Eligibility for FLEX:
  – assigned to ALN in FIT
  – ≥ 3 years of ALN during trial and interim period
  – T-score > −3.5, BMD > FIT baseline

• Of 3236 assigned to ALN in FIT, 1099 randomized 3-2 in FLEX to ALN or placebo

• 662/1099 volunteers assigned to ALN included in analysis
FIT/FLEX Virtual Twins analysis

• Outcome: number of nonspine clinical fractures (NSFx)
  – Poisson, not-overdispersed
  – log-transformed volunteer person-years used as offsets

• Covariates unaffected by treatment: age, BMI, smoking

• Covariates potentially affected by treatment: history of NSFx, VFx, BMD

• NSFx, VFx, BMD modeled using Poisson, logistic, linear models
## Results for FIT/FLEX

Rates per 100 person-years (95% CI)

<table>
<thead>
<tr>
<th></th>
<th>Trial</th>
<th>Extension</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volunteers</td>
<td>3.6 (2.7-4.6)</td>
<td>5.1 (4.2-6.0)</td>
<td>4.4 (3.7-5.2)</td>
</tr>
<tr>
<td>Virtual twins</td>
<td>4.3 (3.8-4.7)</td>
<td>5.1 (4.4-5.8)</td>
<td>4.7 (4.2-5.3)</td>
</tr>
<tr>
<td>Rate ratio</td>
<td>0.85 (0.61-1.10)</td>
<td>1.00 (0.80-1.22)</td>
<td>0.94 (0.77-1.11)</td>
</tr>
</tbody>
</table>
Naive estimates of effect of long-term treatment for FIT/FLEX

1. Assume short-term results hold long-term: RR 0.83 (0.73-0.96)

2. Conditionally assume short-term results hold long-term: fracture rates in active arm increased from 3.6 to 5.1 per 100 p-y ($p < 0.0001$)

3. Use FLEX ITT result to estimate late treatment effect: 0.99 (0.77-1.26). Overall ITT FIT/FLEX RR: 0.87 (0.76-0.996)

4. Project placebo results: overall RR 0.94 (0.82-1.09)
Assumption checking: selection effects

- Outcome models assumed to hold for all placebo participants: covariates capture any dependence of response on volunteering for the extension.

- In FIT/FLEX, placebo participants were not asked to participate in extension, so we can’t estimate parameters using data for placebo volunteers only.

- In FIT, NSFx and VFx rates were lower, end of study BMD higher, among FLEX volunteers, compared to other ALN participants.

- “Offset model” didn’t clearly help in simulations.
Assumption checking: secular effects

- Overall placebo rates in trial could mask an increasing trend ("healthy volunteer effects"), so calculated twin rates in extension would be biased low.

- If interim outcomes during trial available, check for trend independent of time-dependent prognostic covariates.

- Parameters for calculating expected extension rates for twins could be estimated omitting early trial data, but extrapolation would be problematic.

- "Delay model" works for healthy volunteer effect in trial, fails with further changes in rates in extension.

- Also: consider plausibility of other secular effects during extension.
Simulation studies

- Assessed bias of RR estimate, CI coverage and width, relative efficiency

- Selection bias: correlated random effects used to link BMD change, fracture risk, volunteering for extension

- Secular bias: placebo event rates increased in second half of trial and extension, beyond what is predicted by baseline and end-of-trial covariates

- Relative efficiency: mean squared error compared to long-term placebo-controlled trial
Simulation study of selection effects

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Volunteering</th>
<th>Percent</th>
<th>Trial</th>
<th>Extension</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>True model</td>
<td>35%</td>
<td>-0.2</td>
<td>94.8</td>
<td>-0.5</td>
<td>97.6</td>
</tr>
<tr>
<td>Selection</td>
<td>35%</td>
<td>-3.5</td>
<td>92.4</td>
<td>-3.6</td>
<td>93.6</td>
</tr>
<tr>
<td>Selection</td>
<td>85%</td>
<td>-0.8</td>
<td>96.8</td>
<td>0.2</td>
<td>93.2</td>
</tr>
</tbody>
</table>

Bias - Percent bias of rate-ratio estimate
Cov - Coverage of 95% confidence interval for rate-ratio
## Simulation study of secular effects

<table>
<thead>
<tr>
<th>Model</th>
<th>Secular</th>
<th>Trial</th>
<th>Extension</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Increase</td>
<td>Bias</td>
<td>Cov</td>
<td>Bias</td>
</tr>
<tr>
<td>Standard</td>
<td>Trial only</td>
<td>0.8</td>
<td>93.2</td>
<td>28.4</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>0.7</td>
<td>95.6</td>
<td>62.4</td>
</tr>
<tr>
<td>Delay</td>
<td>Trial only</td>
<td>0.7</td>
<td>95.2</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>0.8</td>
<td>93.6</td>
<td>29.9</td>
</tr>
</tbody>
</table>

Bias - Percent bias of rate-ratio estimate  
Cov - Coverage of 95% confidence interval for rate-ratio  
Delay model uses second half of trial data to estimate parameters used in calculating expected twin outcomes in extension
Simulation study of relative efficiency

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Model</th>
<th>Trial MSE</th>
<th>Trial RE</th>
<th>Extension MSE</th>
<th>Extension RE</th>
<th>Overall MSE</th>
<th>Overall RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>True model</td>
<td>Full data</td>
<td>0.022</td>
<td></td>
<td>0.018</td>
<td></td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Twins</td>
<td>0.018</td>
<td>1.21</td>
<td>0.011</td>
<td>1.56</td>
<td>0.009</td>
<td>1.05</td>
</tr>
<tr>
<td>Selection (35%)</td>
<td>Full data</td>
<td>0.028</td>
<td></td>
<td>0.012</td>
<td></td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Twins</td>
<td>0.017</td>
<td>1.62</td>
<td>0.014</td>
<td>0.89</td>
<td>0.010</td>
<td>0.92</td>
</tr>
<tr>
<td>Selection (85%)</td>
<td>Full data</td>
<td>0.009</td>
<td></td>
<td>0.007</td>
<td></td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Twins</td>
<td>0.008</td>
<td>1.19</td>
<td>0.008</td>
<td>0.82</td>
<td>0.006</td>
<td>0.68</td>
</tr>
</tbody>
</table>

MSE - Mean Squared Error of rate-ratio estimate
RE - Relative Efficiency
Conclusions

• A method for estimating effects of long-term treatment from extension studies when the only placebo data is short-term

• Makes fewer and less onerous assumptions than naive methods; assumptions can be partially checked

• Simulations suggest selection bias benign, secular effects might cause trouble; relative efficiency surprisingly good

• Our view: preferable to alternative methods across the board

• Application to HIV/AIDS: possible extension of iPrEx trial of pre-exposure prophylaxis. Others?