Big Data – Challenges & Opportunities

L. Miriam Dickinson, PhD
An Example from the Primary Care Setting

- NIDDK-funded pragmatic trial to test two approaches to improving care for stage 3 and 4 CKD patients in primary care (Chet Fox, PI)
  - Arm 1: Computer decision support (CDS) for CKD patients
  - Arm 2: TRANSLATE action plan: 9 point action plan based on the CCM (includes CDS) plus practice facilitation

• Cluster randomized trial: Unit of randomization (cluster) is the primary care practice
Potential Problems with Cluster Randomized Trials

- The number of clusters to be randomized is much smaller than trials in which individuals are randomized
- Heterogeneity among clusters to be randomized
- Simple, or even stratified randomization can result in study arms that are imbalanced
- Contextual effects and potential confounders (cluster level variables that differ between study arms) can influence intervention implementation and patient outcomes
- Covariate constrained randomization can help achieve balanced study arms
Opportunities from “Big” Data

- Able to obtain summary data from practice EHRs to allow us to use constrained randomization procedures
  - DartNET Institute: a 501©3 organization the hosts datasets for QI and research
- Practice-level data: variables that may affect implementation or outcomes
  - Structural and patient sociodemographic data obtained from practice survey
  - Clinical data
    - Eligible patients identified and baseline data obtained from EHR
    - Aggregated to the practice level
- Structural and sociodemographic data
  - # FTE clinicians, % African American, % Hispanic, % Medicaid or uninsured
- Clinical data
  - % of patients with HbA1c>9, % diabetic, % stage 4 CKD, % with systolic BP>130, % with systolic BP>140
  - Mean GFR, mean HbA1c, mean systolic BP
### CKD Study Baseline Data: Implementation Wave 1: 18 practices

<table>
<thead>
<tr>
<th>Practice Level Variables</th>
<th>Mean (SD)</th>
<th>Min, Max</th>
</tr>
</thead>
<tbody>
<tr>
<td># FTE clinicians</td>
<td>3.8 (3.1)</td>
<td>1, 12</td>
</tr>
<tr>
<td>% African American</td>
<td>2.9% (3.7)</td>
<td>&lt;1%, 15.0</td>
</tr>
<tr>
<td>% Hispanic</td>
<td>16.6% (18.5)</td>
<td>&lt;1%, 68.0</td>
</tr>
<tr>
<td>% Medicaid/ Uninsured</td>
<td>13.8% (8.6)</td>
<td>0%, 30.0</td>
</tr>
<tr>
<td>% Diabetic</td>
<td>34.9% (18.4)</td>
<td>13.0%, 100%</td>
</tr>
<tr>
<td>% with HgA1c&gt;9</td>
<td>9.1% (7.5)</td>
<td>0%, 25%</td>
</tr>
<tr>
<td>% Stage 4 CKD</td>
<td>6.7% (4.5)</td>
<td>0%, 15.6%</td>
</tr>
<tr>
<td>% with BP &gt;130/80</td>
<td>58.1% (14.2)</td>
<td>35.4%, 1.0%</td>
</tr>
<tr>
<td>% with BP&gt;140/90</td>
<td>31.2% (11.6)</td>
<td>0%, 51.9%</td>
</tr>
<tr>
<td>Mean HgA1c</td>
<td>7.02 (0.38)</td>
<td>6.63, 7.88</td>
</tr>
<tr>
<td>Mean eGFR</td>
<td>49.4 (3.2)</td>
<td>45.2, 59.4</td>
</tr>
<tr>
<td>Mean systolic BP</td>
<td>132.1 (3.8)</td>
<td>124.7, 138.6</td>
</tr>
</tbody>
</table>
Covariate Constrained Randomization Procedure

- Generated all possible randomizations using SAS Proc IML
  - 48620 possible combinations of 18 practices into 2 groups (more about this later)
- Stratification incorporated into the procedure (by limited possible randomizations to those that meet pre-set criteria)
  - Geographic region (East vs West)
  - Practice organization: 8 practices in 3 organizations
- Standardize randomization variables to (z-scores or 0 to 1 range)
- Variables weighted equally, i.e. \( w_i = 1 \)
- For each randomization we computed the balance criterion
  - \( B = (w_1(z_{11} - z_{21})^2 + w_2(z_{12} - z_{22})^2 + \ldots ), \) where \( z_{11} \) is the mean of arm 1 units on standardized variable 1 and \( z_{21} \) is the mean of arm 2 units on standardized variable 1, and so on
- Identified an “optimal set” of randomizations from which 1 will be randomly selected
Big Data Challenges – the data

• Challenges now that the intervention is complete
• All data extracted from EHRs retrospectively – some practices dropped out because of changes in EHRs
• Files included
  – Patient level file with 12269 patients in control or intervention practices
  – Clinical data file with 6,104,221 observations
    • 203913 GFRs
    • 47114 ACRs
    • 466078 BPs
    • 114972 LDLs
  – Drug exposure data: 5,412,770 observations
  – Visit data: 804,283 observations
Decisions and Trade-Offs

- Initial cohort – stringent or more relaxed criteria?
  - Misclassification of initial stage
- Different follow-up times for different clinics
- Different follow-up times for patients
  - Mechanisms of missingness (ignorable vs non-ignorable)
- Modelling issues
- Covariates
- Other outcomes
Challenge: Identifying baseline cohort

Patients in control or intervention practices with 2 GFRs <60 at least 90 days apart at any time in the past (for which we have data)
N=12269

Patient had at least two GFRs within 24 months prior to baseline
N=9356

Limit to stage 3 or 4 at baseline: average of last two GFRs >= 15 and <60
N=6699

No GFR in past two years
N=2913

Not stage 3 or 4 at baseline
N=2657
Big Data Challenges – follow-up time

• Initial ITT analysis: purpose is to compare the two groups on CKD progression
• Follow-up time varied by practice so we limited all practices to 24 months of follow-up
  – Less time to observe change from a practice level intervention (possibly less likely to observe a treatment effect)
  – Possible reduction in potential bias due to differences in maximum follow-up time
  – Less likely to have patient loss to follow-up due to rapid CKD progression (hopefully, less likely to have non-ignorable missingness)
• Follow-up time and number of GFRs varied by patient
  – Patients had an average of 12.8 months of follow-up time
  – 28% had less than 6 months of followup
  – 16% had 6 to 12 months of followup
  – 17% had 12 to 18 months of followup
  – 40% had 18 to 24 months of follow-up
• Patients had an average of 4.9 follow-up GFRs (range 2 to 60)
Big Data Challenges – follow-up time

• Careful investigation of mechanisms of missingness to be sure the appropriate covariates are included and missingness is (hopefully) ignorable
  – Our concern was that patients who progress to stage 5 are more likely to be lost to follow-up
• Significant associations with longer follow-up time (>= 12 months vs < 12 months)
  – Stage 4 ckd at baseline (vs stage 3)
  – Male gender
  – Baseline LDL < 100
  – HbA1c >9 at baseline
  – Lower GFR as of last measure and worse CKD stage as of last measure
• It seems that sicker patients may have more aggressive follow-up
• Relatively few patients (1.4%) progressed to stage 5 CKD
• Modeling challenges: Longitudinal mixed effects models (having some difficulty with random SAS memory)
  – Random effects include site of care and patient
  – We would also like random coefficients models (including random slopes) but models don’t converge
  – Sensitivity analyses?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef (SE)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>44.4 (.96)</td>
<td>-----</td>
</tr>
<tr>
<td>Age</td>
<td>-.10 (.01)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Gender</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>Female (ref)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2.14 (.18)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Current smoker</td>
<td>-.06 (.32)</td>
<td>.8518</td>
</tr>
<tr>
<td>Diabetes</td>
<td>-1.09 (.2561)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Ace/Arb</td>
<td>0.43 (.18)</td>
<td>.0192</td>
</tr>
<tr>
<td>Stage 4</td>
<td>-14.7 (.19)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>.29 (.20)</td>
<td>.1548</td>
</tr>
<tr>
<td>BP&lt;140/90</td>
<td>.60 (.20)</td>
<td>.0032</td>
</tr>
<tr>
<td>LDL&lt;100</td>
<td>.08 (.18)</td>
<td>.8518</td>
</tr>
<tr>
<td>HbA1c: less than 7 (ref)</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>7 or greater</td>
<td>-1.18 (.25)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Not done</td>
<td>-.23 (.25)</td>
<td></td>
</tr>
<tr>
<td>Intervention vs controls (at baseline)</td>
<td>-.53 (.71)</td>
<td>.4552</td>
</tr>
<tr>
<td>GFR change per 12 months in controls (slope)</td>
<td>-.83 (.11)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Difference in GFR slope for patients in TRANSLATE practices</td>
<td>0.83 (.14)</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Other issues

• On the subject of covariates .................
  – For now, we’re using baseline sociodemographic and clinical variables
    • Race/ethnicity generally missing
  – Counter-intuitive associations: e.g. use of NSAIDS associated with higher GFRs overall
    • This could reflect greater recognition of CKD and appropriate treatment in sicker patients so be careful about interpreting results
  – Misclassification is a problem, especially for comorbidities – ICD9, ICD10, SNOMED codes mixed

• Other clinical outcomes – systolic blood pressure
• Process of care outcomes – nephrology referral, use of NSAIDS, ACE/ARB, presence of CKD dx, smoking cessation