Selection bias in secondary analysis of electronic health record data

Sebastien Haneuse, PhD

Department of Biostatistics
Harvard School of Public Health
(1) Antidepressants and weight change study
(2) Selection bias in EHR-based studies
(3) Addressing the complexity for EHR data
(4) Remarks
Antidepressants and weight change

- For the most part, antidepressant drugs ‘work’ and the key to decision-making is understanding side-effect profiles and patient preferences

**Q:** Long-term impact of choice on weight change?
- some drugs are hypothesized to induce weight gain/loss
- independent of changes in behavior

- Setting is Group Health
  - integrated health insurance and health care delivery system
  - approx. 600,000 members in WA and ID

- Electronic administrative databases
  - EHR based on EpicCare as of 2005
  - pharmacy database since 1993
  - other databases that track demographic data, enrollment and claims
**Design**

- Retrospective longitudinal study
  - adults aged 18-65 years
  - at least 9 months of continuous prior enrollment and one post-initiation visit

**Weight information**

- The primary outcome of interest is weight change at 24 months post-treatment initiation
- Extract all relevant records for the 2-year interval prior to the start of the episode through to 11/2009
- Observed information consists of 519,344 records on 16,277 individuals
- Although weight is continuous and follows some smooth trajectory over time, the EHR only provides a series of ‘snapshots’ of a person weight over time
In some instances, these snapshots provide rich information:
In other instances, the information is ‘less rich’:
• Still others are ‘less rich’ but for different reasons:

![Graphs showing weight changes over time relative to treatment initiation.](image)
Selection bias in EHR-based studies

- When we proposed the antidepressants study, we emphasized the benefits of using Group Health EHR data:
  - large patient population
  - long time period
  - huge amounts of information
  - readily-available and accessible
  - relatively cheap to obtain

- Notwithstanding these advantages, EHRs are primarily developed to facilitate improved clinical care and improved tracking/processing of claims

- As we move forward we need keep in mind the fact that the data was not collected for the purposes of this study (or any other study)
Q: Are data obtained from the EHR comparable in scope and quality to data that would have been collected by a dedicated study?

* probably not

- From a methodologic perspective, challenges that we face when using EHR data for research include:
  - extraction of text-based information
  - irregular and inconsistent measurements
  - inaccurate data (i.e. measurement error and misclassification)
  - confounding bias

- While each of these will have to be considered in the antidepressant study, the focus in this talk is on the second challenge
  - particularly with respect to the measurement of weight
Selection bias

- With respect to the primary outcome, only \(\approx 15\%\) of patients identified in the EHR via the inclusion/exclusion have ‘complete’ data:

<table>
<thead>
<tr>
<th>Patients identified in the EHR</th>
<th>16,277</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight information at baseline*</td>
<td>14,570</td>
</tr>
<tr>
<td>Weight information at 24 months*</td>
<td>2,647</td>
</tr>
<tr>
<td>Weight information at both*</td>
<td>2,408</td>
</tr>
</tbody>
</table>

* based on a \(\pm 30\)-day window

Q: If we restricted our analyses to \(n=2,408\) patients with ‘complete’ data, how representative/generalizable would the results be?

- Depending on why some folks have complete data vs other not, a naïve analysis may be subject to selection bias
  
  * results are not externally generalizable to the population of interest
Towards the control of selection bias

- In principle one can cast the control of selection bias as a missing data problem and use established statistical methods
  - multiple imputation, IPW, doubly-robust methods, PMMs

- Validity of these methods relies on the *missing at random (MAR)* assumption

- In practice, consideration of the MAR assumption typically boils down to:

  (i) conceiving of a mechanism that drives missing vs. not

  (ii) identifying factors that are relevant to the mechanism

  (iii) hoping that all relevant covariates are measured
● This approach, however, fails to acknowledge three important features of EHR data settings:

(i) the inherent complexity of clinical contexts
(ii) the high-dimensional nature of EHRs
(iii) EHRs are not developed for research purposes

● Jointly these features can cause havoc:
  * standard practice likely represents an overly simple/unrealistic notion of ‘observance’
  * identifying relevant factors is often challenging
  * relevant information may not be available in the observed data
    * i.e. the data are MNAR

● Inappropriate treatment of these features can leave the analysis suffering from residual selection bias
Addressing complexity

For a patient to have ‘complete’ data from the EHR in the antidepressants study they must:

- be enrolled within the Group Health health plan at 24 months
- have initiated a clinical encounter at 24 months
- had a weight measurement recorded in the EHR during the encounter

Each of these is, in some sense, a distinct ‘decision’

- a distinct sub-mechanism

More generally, if residual selection bias is to be avoided, one needs to understand/determine

- which sub-mechanisms needs to be considered
- the interplay between the sub-mechanisms
- how risk factors influence the sub-mechanisms
May be helpful to visualize the flow of decisions with a diagram:

(a) Simple specification

(b) Detailed specification
Mechanism 1: Continuity/enrollment within the system

- One can generally conceive of a ‘health care system’ to which the EHR corresponds

- EHRs can only observe/record care to the extent that the patient is able to interact with the system
  * distinct from whether or not they do interact

- Within the context of the antidepressants study at Group Health:
  * we know if/when someone dies
  * at any given time point, we know their insurance/enrollment status

- Some people disenroll and then re-enroll
  * assume gaps $\leq 92$ days do not represent actual discontinuities in coverage
  * 5% of patients in the antidepressants data have more than one enrollment period
Enrollment patterns for a non-random sample of 12 individuals
Distribution of the gap of enrollment among 802 folks with only one gap:
• With respect to the primary outcome, among the 16,277 patients identified at the outset:
  * 3,405 (21%) disenroll prior to censoring or 24 months
  * 6,205 (38%) are censored prior to disenrolling or 24 months
  * 6,698 (41%) make it to the 24 month mark without disenrolling or being censored
Mechanism 2: Initiation of an encounter

- For an encounter to occur it must be initiated

- For some diseases/conditions, there are clear guidelines about when encounters should occur
  * e.g., HEDIS guidelines for treatment and follow-up of depression

- Intuitively, the intensity of interaction with the health care system (and EHR) will often depend on the underlying health state of the patient

- We’ve already seen that there can be substantial variation in the number and the timing of encounters across patients
- Observed number of encounters on or after the date of treatment initiation
  - 175 individuals with more than 100 encounters
Number of post-initiation encounters during a (standardized) 24-month period

- Median of 11 encounters/24-months
- 1,030 individuals with 2 or fewer encounters/24-months
- 724 individuals with more than 50 encounters/24-months
• Number of patients enrolled and not censored at 24 months who initiated an encounter
  * out of 6,698

<table>
<thead>
<tr>
<th>Window</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 months ± 0 days</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>24 months ± 7 days</td>
<td>1,736</td>
<td>25.9</td>
</tr>
<tr>
<td>24 months ± 14 days</td>
<td>2,665</td>
<td>39.8</td>
</tr>
<tr>
<td>24 months ± 30 days</td>
<td>3,882</td>
<td>58.0</td>
</tr>
</tbody>
</table>

• Number of individuals with a given number of encounters in 24 months ± 14 days:

<table>
<thead>
<tr>
<th>Number of Encounters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1463</td>
<td>626</td>
<td>314</td>
<td>143</td>
<td>143</td>
<td>54</td>
<td>26</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Mechanism 3: Measurement

- Given that an encounter took place, the measurement must be taken and recorded.

- Among the 2,665 individuals with at least one encounter in a 24-month ± 14 day window, \( n=1,431 \) have at least one weight measurement.
  * 1,234 (46%) have an encounter but no weight measurement.

- Number of individuals with a given number of weight measurements in 24 months ± 14 days:

<table>
<thead>
<tr>
<th>Number of Measurements</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>1234</td>
<td>1166</td>
<td>212</td>
<td>44</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

- There may be many reasons why a measurement is or is not taken:
  * reasons related/unrelated to the underlying value
  * decisions made by patients, providers and organizations.
Of all 4,990 encounters at 24 months, 1,760 (35.3%) had a recorded weight measurement

<table>
<thead>
<tr>
<th></th>
<th>Total-N</th>
<th>Total-%</th>
<th>Observed-N</th>
<th>Observed-%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>Yes</td>
<td>1792</td>
<td>35.9</td>
<td>1393</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>3198</td>
<td>64.1</td>
<td>367</td>
</tr>
<tr>
<td>Spec</td>
<td>Yes</td>
<td>2544</td>
<td>51.0</td>
<td>556</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2446</td>
<td>49.0</td>
<td>1204</td>
</tr>
<tr>
<td>MH</td>
<td>Yes</td>
<td>595</td>
<td>11.9</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>4395</td>
<td>88.1</td>
<td>1743</td>
</tr>
<tr>
<td>Sex</td>
<td>Male</td>
<td>1450</td>
<td>29.1</td>
<td>518</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>3540</td>
<td>70.9</td>
<td>1242</td>
</tr>
<tr>
<td>BP</td>
<td>Yes</td>
<td>1981</td>
<td>39.7</td>
<td>1670</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>3009</td>
<td>60.3</td>
<td>90</td>
</tr>
<tr>
<td>Pulse</td>
<td>Yes</td>
<td>1661</td>
<td>33.3</td>
<td>1398</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>3329</td>
<td>66.7</td>
<td>362</td>
</tr>
</tbody>
</table>
Proposed general strategy/framework

• Our work on the antidepressants study suggests the need for a more nuanced approach to selection bias:
  ∗ specification of the mechanism that drives whether or not we observe complete data
  ∗ how we learn about the mechanism
  ∗ how we perform statistical adjustments

• Focus on the notion of observance

• Breakdown the task of characterizing observance via a sequence of focused sub-mechanisms
  ∗ each sub-mechanism corresponds to some specific decision

• Anticipate that it will be easier to consider each in turn:
  ∗ conceptually
  ∗ practically
Beyond those considered here, there are many other decisions/sub-mechanisms that may need to be kept in mind:

- completeness at other time points
- receipt of care outside the system
- choice of encounter type
  - specialist visit, phone encounter, secure messaging
  - changing measurement standards and/or infrastructure

Not all of these will be relevant in any given EHR context

- ‘closed’ systems, such as Group Health and the VA
- ‘open’ systems, such as the one maintained at Harborview or at Brigham and Women’s Hospital
- claims data, such as Medicare
- registries, such as SEER
Some may require consideration of monotonicity

- does it make sense to think of an ‘encounter’ if a patient is not enrolled?
- does it make sense to think of ‘measurement’ if no encounter took place?
- flow-type diagrams will be useful

Whatever structure is adopted, for each sub-mechanism one would need to consider a broad range of factors for each mechanism

- patient-, encounter-, provider-, organization-level
- specific factors may differ across mechanisms
  - whether or not they are important
  - direction of association
  - magnitude of association
Concluding remarks

• When using EHR data for research purposes, one needs to choose a general philosophy about how to use the available information:

  (1) Do the best that we can with everything that is available
      * e.g. model the entire trajectory over the course of time

  (2) Ground the analysis within the context of an ‘ideal’ study
      * i.e. the study that would have been designed, had opportunity arisen

• The first is likely the position that most folks will take by default
  * gain statistical efficiency by borrowing strength across time and patients

• Potential drawbacks:
  * likely requires the specification of a large, complex outcome model
  * notions of ‘complete’ data or ‘missing’ data are not clear and may be obscured
• Raises two important questions:

  Q: do we want to model ‘everything’?
  Q: what is the population to which the results generalize?

• The second philosophy is the one that I’ve adopted

• Appealing because it forces explicit conceptual and operational definitions of:
  ∗ the target patient population of interest
  ∗ the what it means to have ‘complete’ data

• These are not trivial tasks because the richness of EHR data gives researchers much more flexibility and choice than they would normally otherwise have

• The philosophy focuses the science and statistical analyses but has the drawback of resulting in the ‘throwing away’ of information
  ∗ what do we do, if anything, with the 12-month weight data?
Terminology:

* **target population**
  * defined on the basis of scientific inclusion/exclusion criteria

* **study population**
  * defined on the basis of a combination of scientific and practical inclusion/exclusion criteria
  * results in $N$ patients identified in the EMR

* **study sub-sample**
  * $n$ patients with ‘complete’ data
  * function, in part, of the scientific question
In EHR-based studies there are two distinct forms of potential non-representativeness:

* depending on what you take to be the population of interest

TARGET POPULATION
Defined on the basis of Inclusion/exclusion criteria

STUDY POPULATION, N
Patients identified via inclusion/exclusion criteria

STUDY SUB–SAMPLE, n
Patients with complete data