A general framework for selection bias due to missing data in EHR-based research

Sebastien Haneuse, PhD
Harvard T.H. Chan School of Public Health

David Arterburn, MD
Group Health Research Institute

Michael Daniels, ScD
University of Texas - Austin

Outline

(1) Antidepressants and weight change study

(2) Selection bias as an under-appreciated challenge

- cast as a missing data problem
- assumptions
- limitations of standard approaches

(3) The proposed framework

- simple illustration using the antidepressants study
- general applications

(4) Moving forward

- opportunities for methods development

(5) Concluding remarks

Antidepressants and weight change

- For the most part, antidepressant drugs 'work'
- Keys to decision-making include side-effect profiles coupled with patient preferences

Q: Do different drugs or drug classes have different impacts on long-term weight change?

  * some drugs were hypothesized to induce weight gain/loss
  * independent of changes in behavior

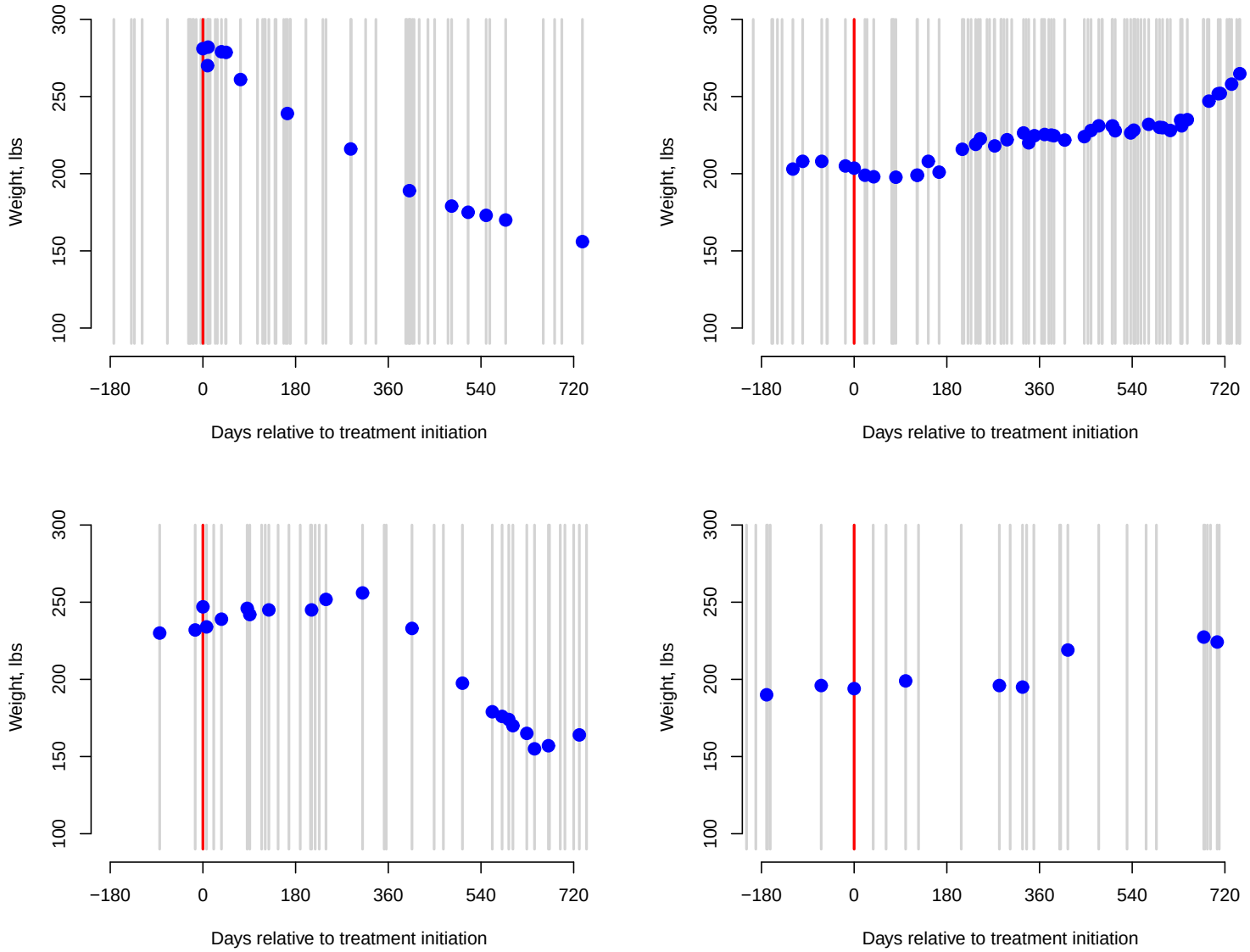
- R-01 funded comparative effectiveness study conducted at Group Health Cooperative
  * integrated health insurance and health care delivery system
  * approx. 600,000 members in WA and ID

- Electronic databases:
  - EHR based on EpicCare as of 2005
  - pharmacy since 1977
  - other databases that track:
    - demographic data
    - enrollment information
    - inpatient and outpatient claims
    - primary care visit appointments
- Design was a retrospective longitudinal study
- Inclusion/exclusion criteria define the population of interest:
  - adults aged 18-65 years
  - new treatment episode between 01/2006 - 12/2007
  - $\geq$ 9 months of continuous prior enrollment
- N=9,704 patients identified

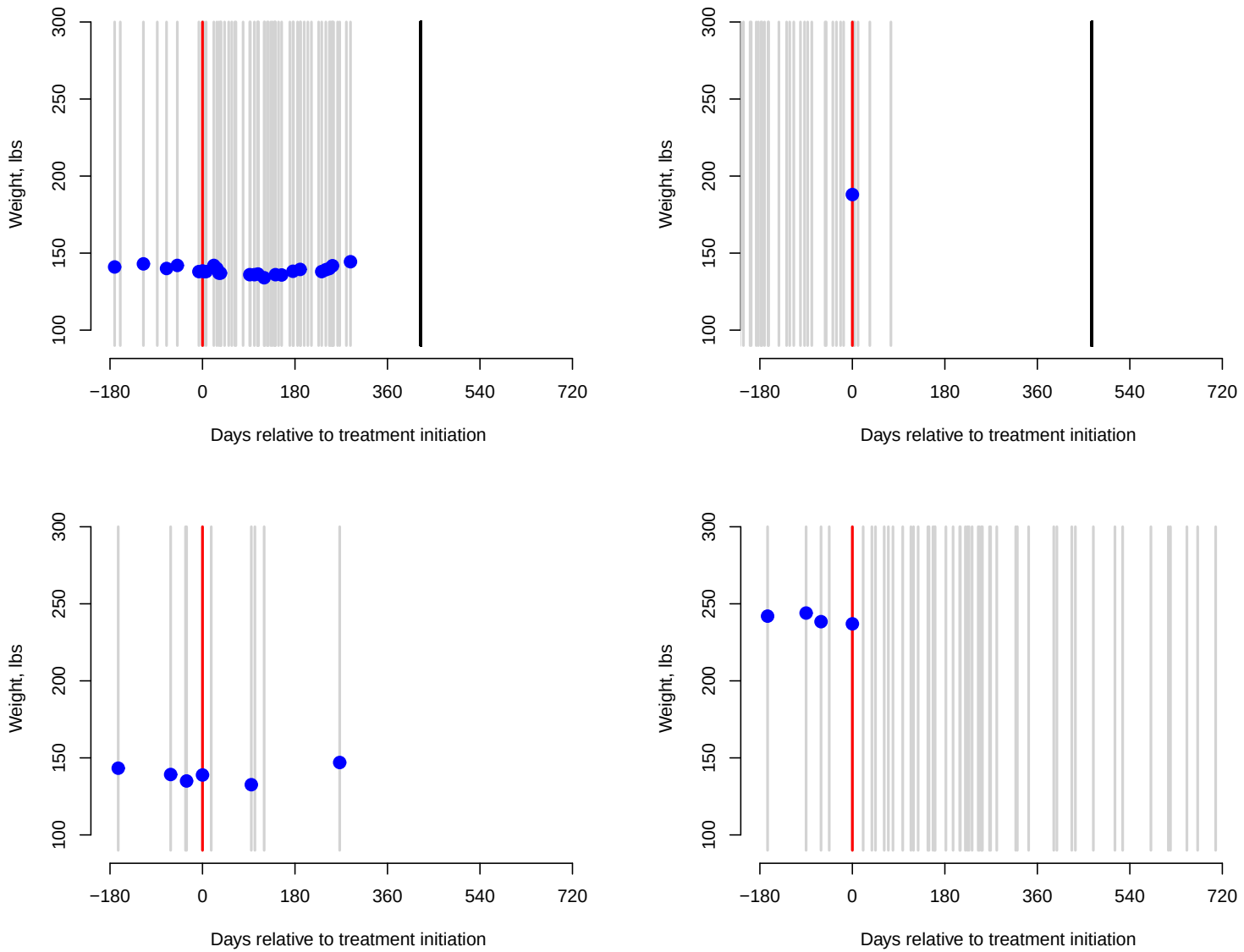
Weight information

- Primary scientific interest lies with long-term weight change
  - at 24 months following treatment initiation
- Extract all relevant records for the 2-year interval prior to the start of the episode through to 11/2009
- 354,945 records
  - weight
  - potential confounders
  - auxiliary variables
- Although weight is continuous and follows some smooth trajectory over time, the EHR only provides a series of 'snapshots'

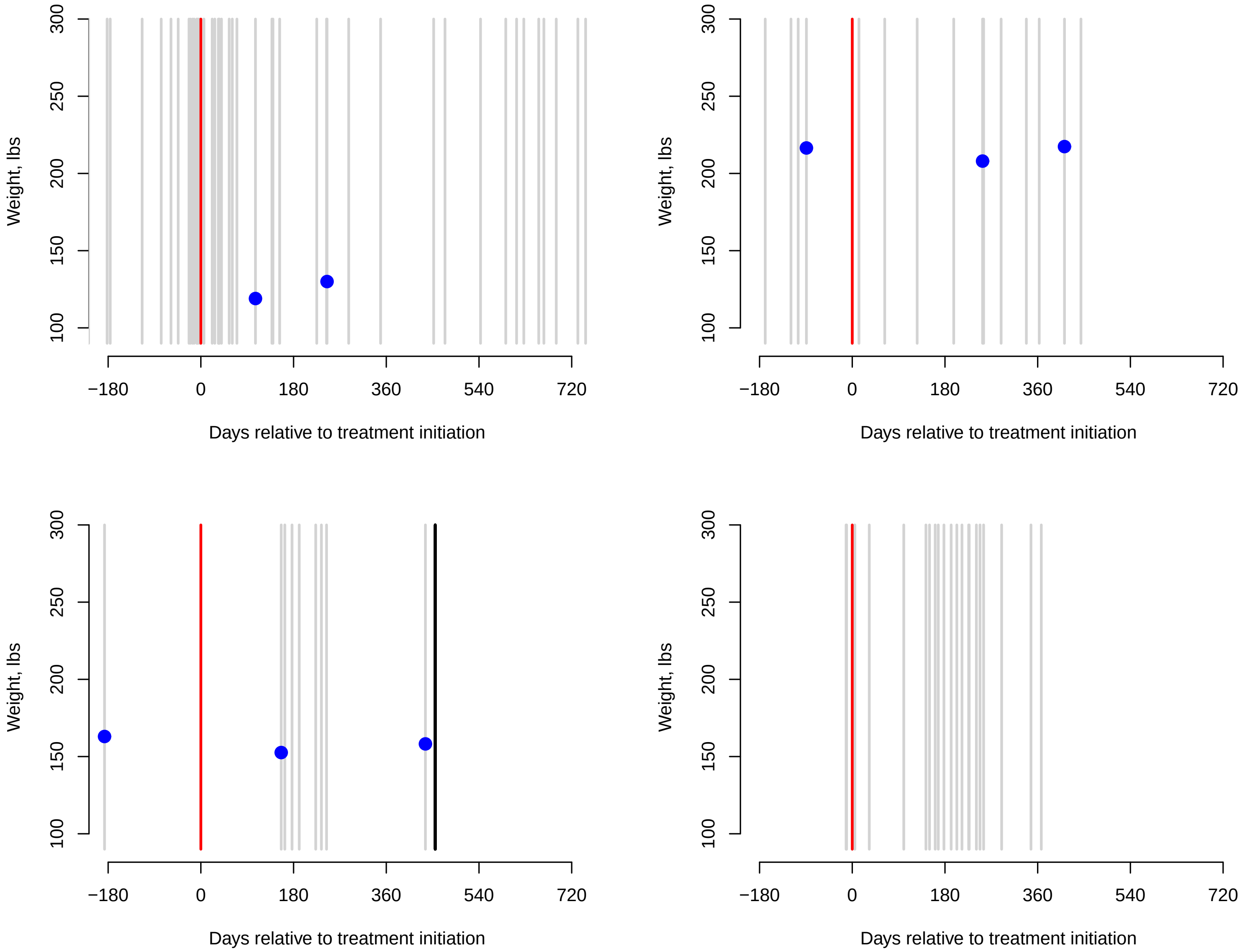
- In some instances, these snapshots provide rich information on weight:



- In other instances, the information is 'less rich':



- Still others are 'less rich' but for different reasons:



- The numbers/figures presented so far illustrate the numerous benefits that EHR data can enjoy:
  - large patient populations
  - long time periods
  - huge amounts of information
  - readily-accessible and relatively cheap to obtain

- Moving forward we need keep in mind the fact that EHR data is not collected for research purposes

- They are primarily developed to facilitate:
  - improved clinical care
  - improved tracking/processing of claims

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

- Specific challenges faced by EHR-based studies:
  - linkage of patient records across databases
  - extraction of text-based information
  - irregular and inconsistent measurements
  - inaccurate data (i.e. measurement error and misclassification)
  - confounding bias

- Most of these challenges are not new
  - manifest in more 'traditional' contexts

Q: Can we use existing methods to address these challenges in the EHR context?

- In many settings the answer will be 'no', in part because standard methods will often fail to acknowledge the scale, complexity and heterogeneity of EHR data

- There is an emerging literature, however, on statistical methods that are specifically tailored to comparative effectiveness research the EHR setting
- Much, if not most, of this has focused on methods towards resolving confounding bias
- Other areas that have received more recent attention include
  - probabilistic linkage of records across databases
  - NLP for text-based notes
- The focus here is on what we believe is an under-appreciated (potential) problem ... *selection bias*

Selection bias due to missing data

- In the antidepressants study, keeping in mind the primary scientific goal, the 'ideal' is that weight information is available at baseline and at 24 months for all N=9,704 identified by the inclusion/exclusion criterion

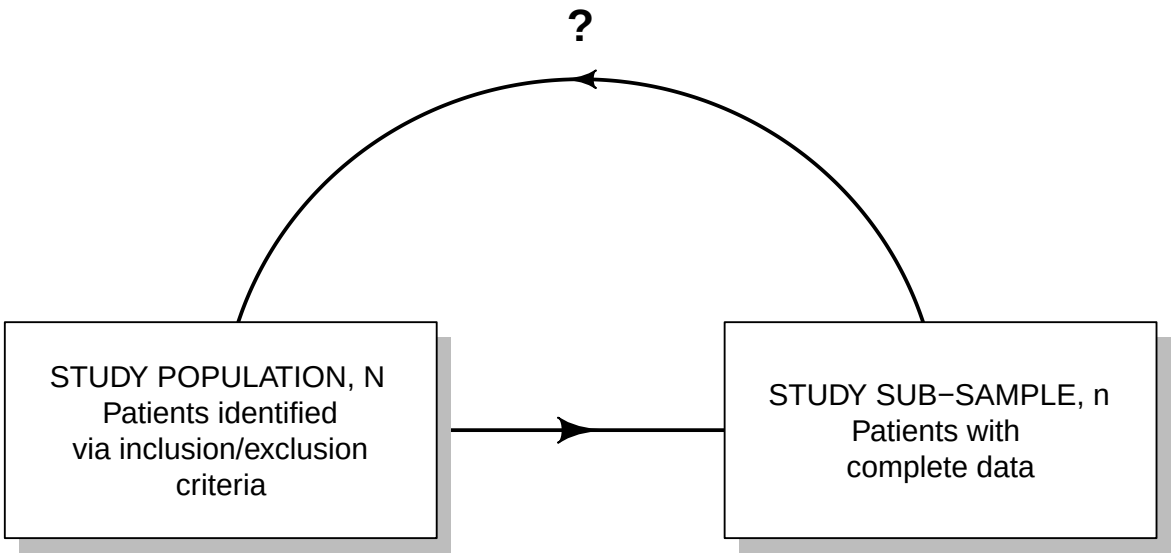
- Unfortunately, once data was abstracted from the relevant databases, it was discovered that there was substantial missing data:

Patients identified in the EHR	9,704	
Weight information at baseline*	8,631	88.9%
Weight information at 24 months*	2,647	27.3%
Weight information at both*	2,408	24.8%

* 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?

* to what extent can we draw valid conclusions for the population of interest?



- A naïve analysis may be subject to *selection bias*
  * results are not externally generalizable
  * distinct from confounding bias or internal validity

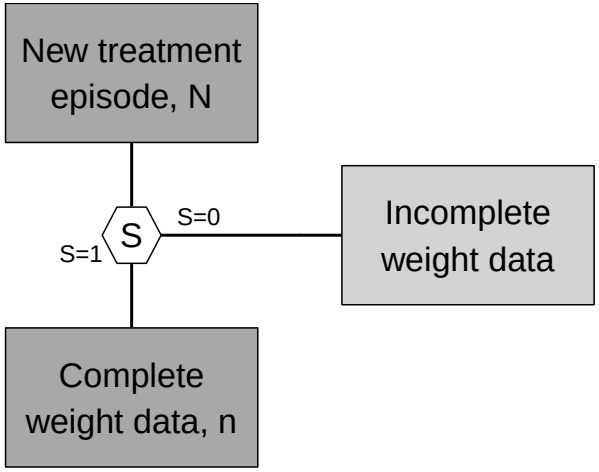
- Intuitively, selection bias can be cast a missing data problem
- Could therefore appeal to the (huge) literature on methods for missing data
- The validity of all methods for missing data rely, in one way or another, on the *missing at random (MAR)* assumption
  - missingness solely depends on variables one has access to
  - when MAR does not hold, we say that the data are *missing not at random (MNAR)*
- Typically, consideration of missing data assumptions boils down to:

(i) conceiving of a mechanism that drives whether or not data are missing

(ii) identifying factors that are relevant to the mechanism

(iii) hoping that all relevant covariates are measured

- Graphically, we might represent the single mechanism as:



- Such a 'single mechanism' approach, however, fails to acknowledge:
  - the inherent complexity of clinical contexts
  - the high-dimensional nature of EHRs
  - the heterogeneity within and between EHR systems

- Ultimately, failure to fully account for these issues may leave the analysis suffering from residual selection bias
  - compromise generalizability (and therefore utility) of the results

Proposed framework

- Given the complexity and heterogeneity of EHR systems it is unlikely that any single method will be universally applicable
  - cannot be prescriptive in the task of controlling selection bias
- Instead we are proposing that researchers consider and apply two key principles:
  1. Specify the structure of the data that would have been collected had the opportunity to conduct the 'ideal' study been an option
  2. Frame the task of controlling selection bias with the questions 'what data are observed and why?'
     - sometimes referred to as the *data provenance*

Consideration of the ideal study

- Will generally involve:
  - identifying all variables that would have been collected
  - indicating the timing of measurements
- Specific choices depend primarily on the scientific goals of the study
  - could be approached much in the same way that researchers approach data collection strategies in grant proposals
  - may be challenging given that there is (likely) so much choice
- In the antidepressants study we focused on weight change at 24 months
  - arguably only need weight information at two time points
- At alternative scientific goal may have been to characterize the weight trajectory of patients during the 24 months post-treatment initiation
  - intermediate weight measurements, depending on the level of granularity

Consideration of data provenance

- The key benefit of going through the process of specifying the ideal study is that it renders the notions of 'complete data' and 'missing data' as meaningful
- Armed with this we can begin to characterize why any given patient has complete/incomplete data
- Whether or not any given data element is observed could, for example, depend on decisions made by the patient, their provider(s) and the health care system
  - in many instances there will be a complex interplay between numerous such decisions
- It may also be that covariates have differential impact on different decisions
  - no impact vs some impact
  - positive association vs. negative assocation

- Propose a general strategy based on modularizing data provenance
  - breakdown the task of characterizing a complex process into a series of manageable sub-mechanism
  - each sub-mechanism corresponds to some specific decision

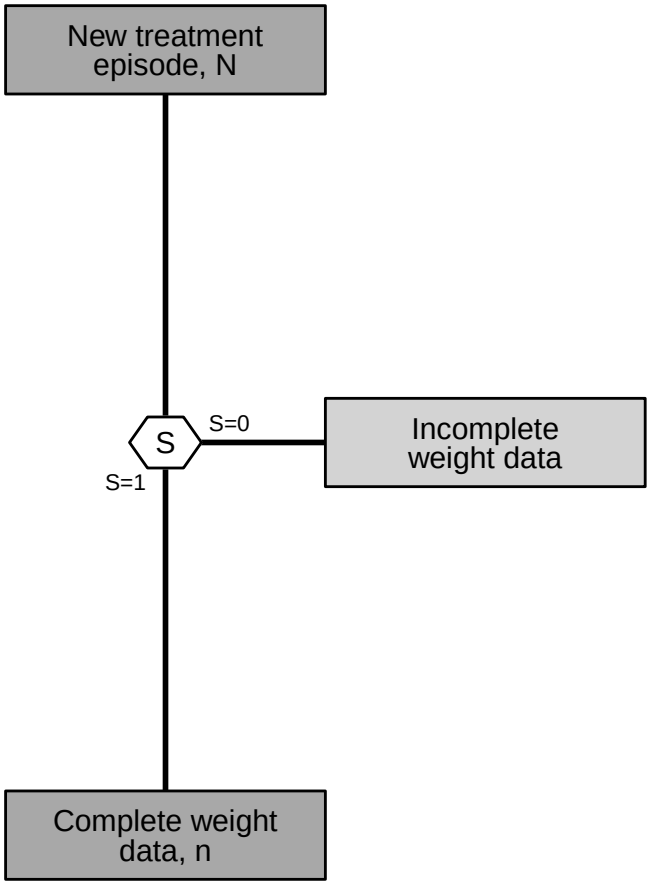
- In the antidepressants study, for example, for a patient to have complete weight data at 24 months in the Group Health EHR they must at least:

  (i) be actively enrolled at 24 months

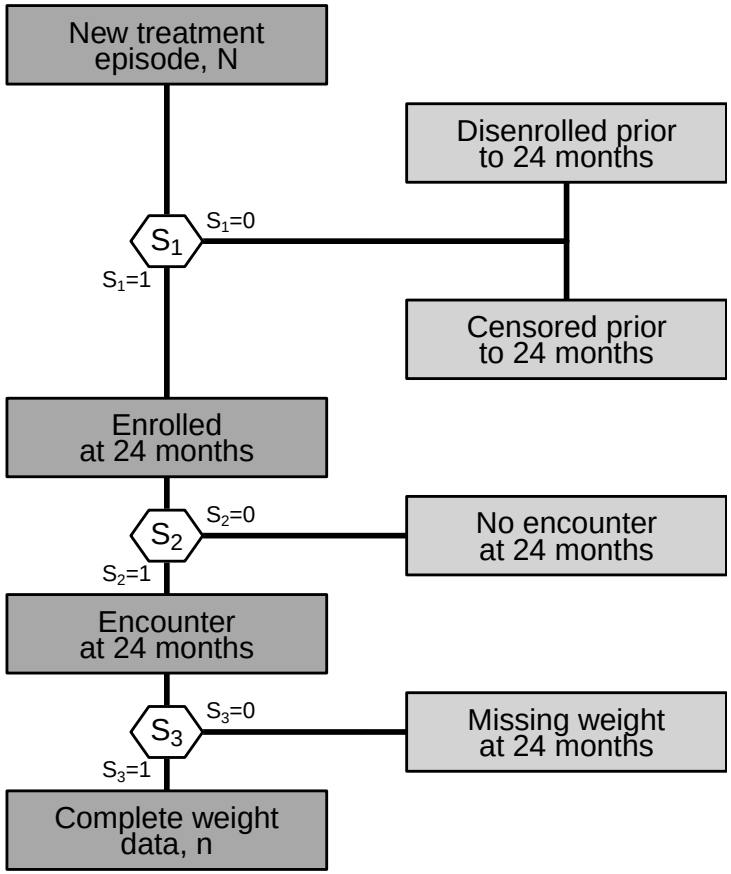
  (ii) have initiated a clinical encounter at 24 months

  (iii) had a weight measurement recorded in the EHR during the encounter

- Note, in the standard approach to missing data these three would be 'collapsed' into a single mechanism:



(a) Simple specification

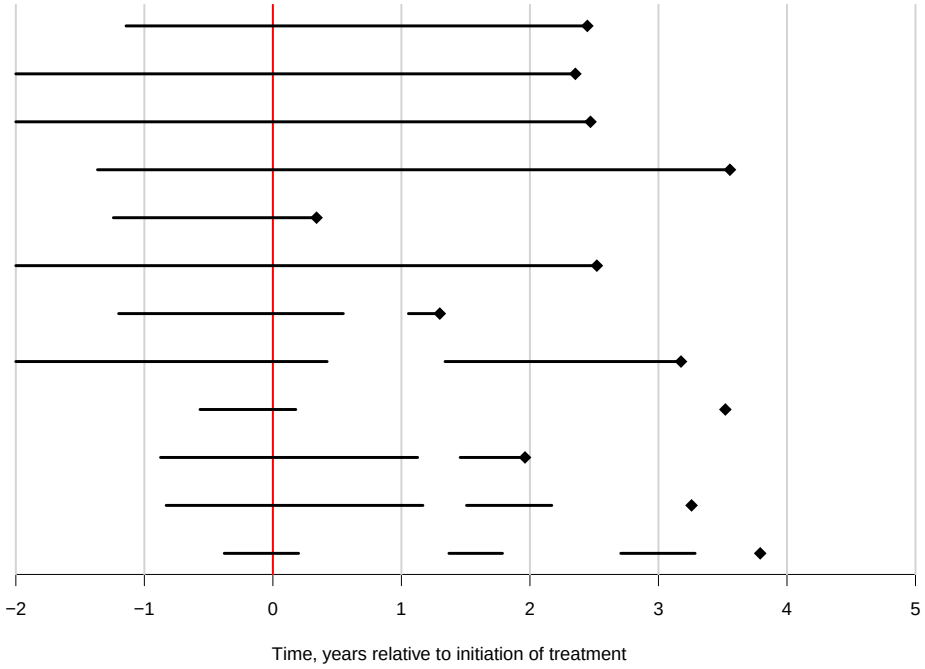


(b) Detailed specification

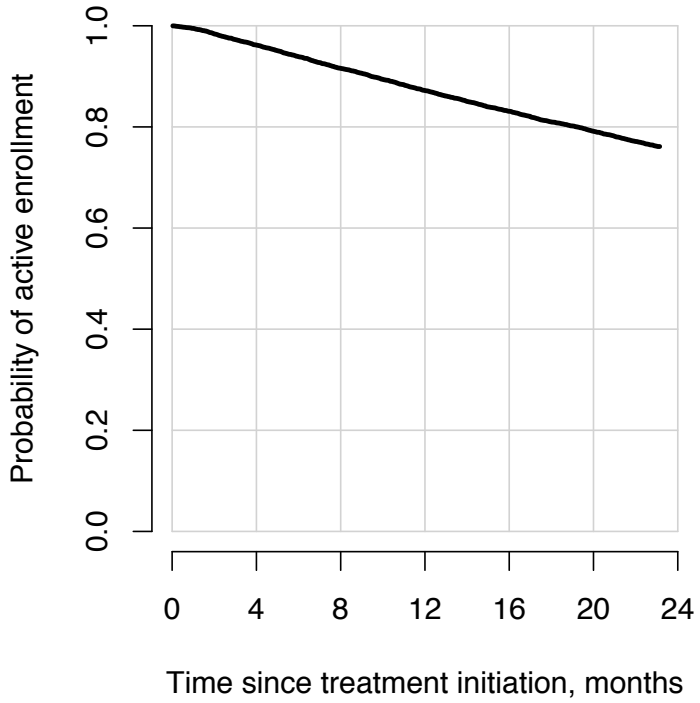
Sub-mechanism #1: Active enrollment status at 24 months

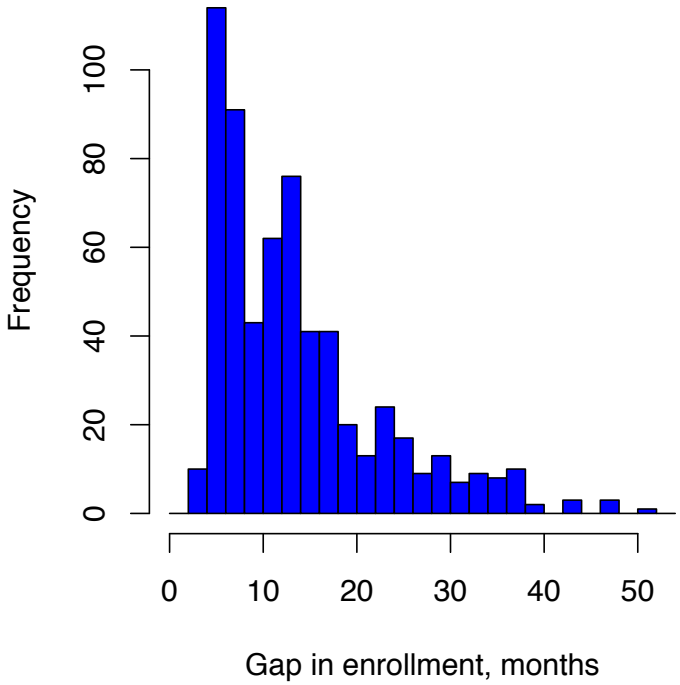
- EHRs can only observe/record care to the extent that the patient is *able* to interact with the health care system to which the EHR corresponds
  - distinct from whether or not they do interact

- Enrollment patterns for a non-random sample of 12 patients



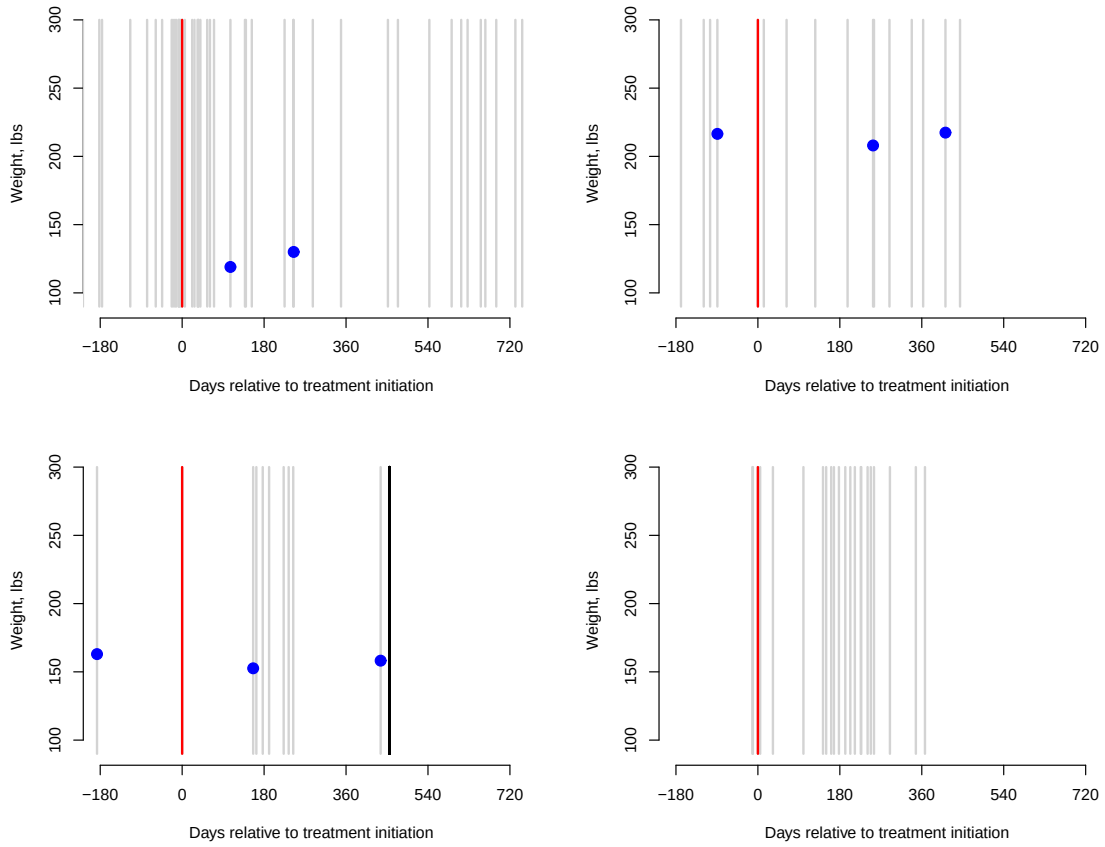
- Among the 8,631 patients with weight information at baseline:
  - 2,061 (23.9%) disenrolled at some point in the first 24 months
  - 617 patients had at least two periods of enrollment





Sub-mechanism #2: Initiation of an encounter at 24 months

- That a patient is enrolled does not mean that they engage with the health care system at points in time that are of interest to the research
  - for a measurement to be recorded in the EHR an encounter must be initiated



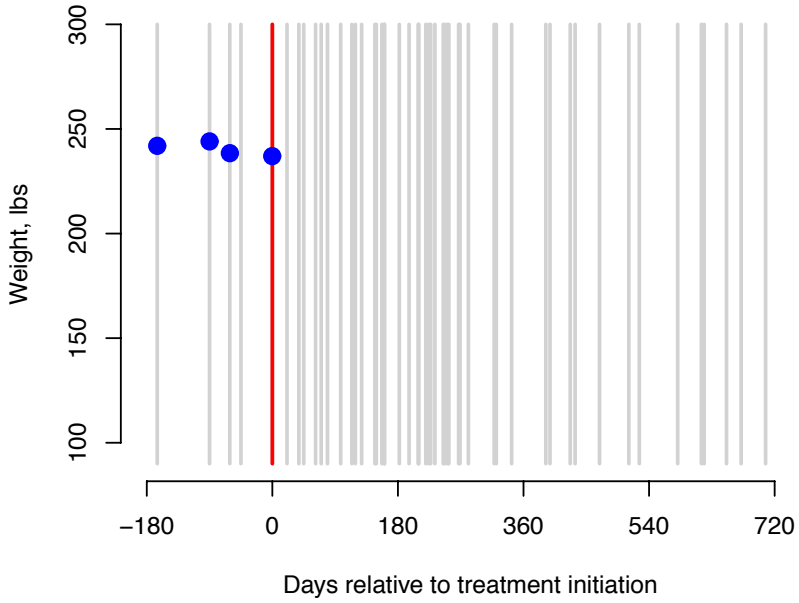
- Among the 6,570 patients actively enrolled at 24 months:

Window around the 24-month mark	Number of patients (%)
$\pm$ 7 days	1,604 (24.4%)
$\pm$ 14 days	2,485 (37.4%)
$\pm$ 30 days	3,688 (56.1%)

- 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

Sub-mechanism #3: Measurement of weight at 24 months

- Even if a patient is (i) enrolled and (ii) initiated an encounter, they may not have a weight measurement recorded



- Among the 3,688 patients enrolled and with at least one encounter in a 24-month $\pm$ 30 day window, n=2,408 (65.3%) have at least one weight measurement

Odds ratio estimates from model fits

	Single mechanism (N=8,631)	**Sub-mech #1: enrollment** (N=8,631)	**Sub-mech #2: encounter** (N=6,570)	**Sub-mech #3: measurement** (N=3,688)
Female	**1.33**	**1.11**	**1.30**	**1.20**
Age	**1.16**	**1.41**	**1.10**	**0.97**
Weight at baseline	1.05	1.02	1.03	1.05
Antidepressant				
Fluoxetine	1.00	1.00	1.00	1.00
Buproprion	1.05	1.01	1.14	0.92
Mirtazapene	1.29	0.94	1.18	1.54
Paroxetine	**1.09**	**1.27**	**1.19**	**0.83**
SSRI	0.90	0.87	1.08	0.79
SARI	1.51	1.36	1.59	1.09
Tryciclics	1.80	1.33	1.93	1.28

The framework in more general contexts

- Beyond those already considered, there are many other decisions/sub-mechanisms that may need to be kept in mind:
  - completeness at other time points
    - e.g., baseline weight
  - completeness in other variables
    - e.g., confounders such as depression type/severity
  - receipt of care outside the system
    - e.g., mental health visits with a specialist
  - choice of encounter type
    - e.g., specialist visit, phone encounter, secure messaging
  - changing measurement standards and/or infrastructure
    - e.g., ICD coding systems

- Not all sub-mechanisms 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 Brigham and Women’s Hospital
  - claims data, such as Medicare
  - disease registries, such as SEER
- Some may require consideration of monotonicity
  - does it make sense of 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-, system-level
  - note that specific factors may differ across mechanisms in either the direction or magnitude of association

Moving forward

- Conceptually, the proposed strategy provides structure within which:

  (i) transparency of assumptions regarding missing data can be enhanced

  (ii) factors relevant to each decision can be more easily elicited

  (iii) statistical methods and sensitivity analyses can be better aligned with the complexity of the data

- In regard to (iii), moving beyond the standard single mechanism framework provides opportunity for the development of novel statistical methods
  - specifically tailored to the high-dimensional, complex nature of EHR data

Opportunities

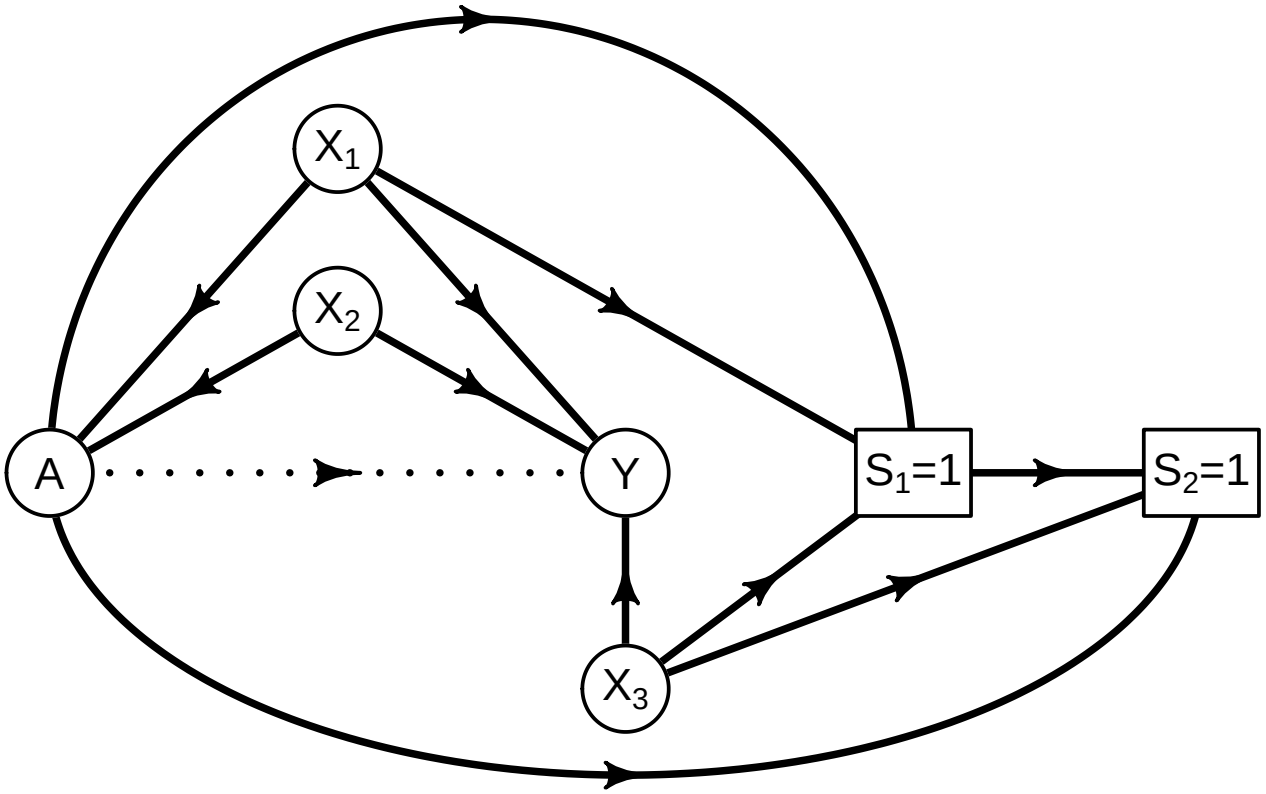
- The extension of existing methods to accommodate the structure identified by the proposed framework
  - IPW, MI, PMM
- The complementary use of these methods in a single analysis
  - i.e. the use of IPW for some sub-mechanisms and MI for others
- The development of data-driven model/variable selection algorithms
  - may not be feasible to incorporate all sub-mechanisms into an analysis
  - for any given sub-mechanisms they may be hundreds of potential inputs
- The development of methods for data integration
  - databases within a single health care system
  - EHRs from different health care systems
  - EHR data with administrative claims data

- Methods for the valid quantification of uncertainty
  - propagate all sources through to the final inference
  - e.g., those due to estimation of models for the various sub-mechanisms
- Folding in proactive targeted data collection efforts to 'fill in' missing data or towards evaluating assumptions that might otherwise not be testable
  - chart reviews
  - patient surveys
  - interviews with health care providers and systems maintainers
- Note, the last of these could be geared towards both confounding bias and selection bias

Bias-variance trade-off

- An important caveat is that the extra 'work' involved may require researchers to possibly contend with a bias-variance trade-off
- Specifically, in some settings the additional detail may be unnecessary/unwise
  - may not actually reflect the 'true' data provenance
  - may not have sufficient information to characterize certain sub-mechanisms
- Forging ahead dogmatically with the proposed framework may result in an increase in the variance of the estimator
- Small simulation study to illustrate the point
  - consider the association between some response Y and a treatment A
  - two component selection sub-mechanisms

- Graphical representation of the set-up:



Model (1) $E[Y] = \beta_0 + \beta_a A + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Model (2) $\text{logit } P(S_1=1) = \alpha_{10} + \alpha_{1a} A + \alpha_{11} X_1 + \alpha_{13} X_3$

Model (3) $\text{logit } P(S_2=1|S_1=1) = \alpha_{20} + \alpha_{2a} A + \alpha_{23} X_3$

- Results for two scenarios:
  - #1: designed so that the naïve analysis would exhibit moderate bias
  - #2: designed so that the naïve analysis would exhibit little-to-no bias

Simulation scenario	Selection bias adjustment	Percent bias	Standard error	Relative uncertainty*
#1	None	-35.9	0.51	0.87
	Single†	-38.8	0.59	1.00
	Modularized‡	-1.9	1.09	1.85
#2	None	-21.1	0.27	0.29
	Single†	-6.7	0.93	1.00
	Modularized‡	-0.3	1.22	1.30

- An important avenue for future work, therefore, is to characterize (to the extent possible) when the standard single mechanism strategy will be the 'optimal' way forward

Concluding remarks

- As EHRs become the norm in clinical practice, researchers will be increasingly drawn to the rich data they provide on large populations
- In the future, EHR systems may be designed with (secondary) research agendas in mind
- In the meantime, as EHR data is used for research purposes, statistical analyses can be guided by one of two philosophies regarding how the available information should be used:

(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 are that:
  - likely requires the specification of a large, complex outcome model
  - notions of 'complete' data or 'missing' data are not clear
- 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 we have adopted
- Appealing because it forces explicit conceptual and operational definitions of:
  - the target patient population of interest
  - 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
- While the philosophy focuses the science, it has the drawback of possibly 'throwing away' of information
  - what do we do, if anything, with the 12-month weight data?
  - may be a reasonable price however