Adjusting for selection bias due to missing data in EHR-based research

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

Sarah Peskoe, PhD
Duke University

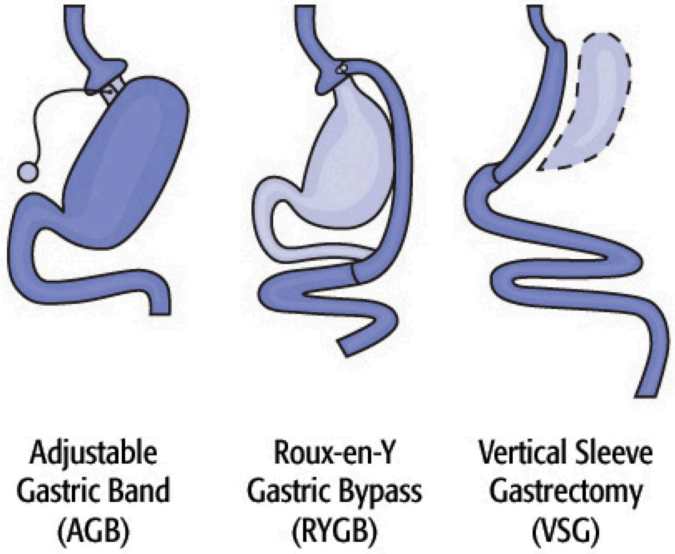
David Arterburn, MD
Kaiser Permanente Washington Health Research Institute

Michael Daniels, ScD
University of Florida

Long-term outcomes following bariatric surgery

- Approx. 300 million people worldwide have T2DM
  - number set to increase to 438 million by 2030
- Sustained metabolic control is difficult through first-line treatment options
  - i.e. lifestyle modifications, physical activity and pharmacotherapy
- Bariatric surgery is increasingly being accepted as a safe and effective alternative to conventional therapy for obese patients
  - recent endorsement by the American Diabetes Association
- Some controversy remains, especially around long-term outcomes
  - surgery vs. conventional therapy
  - across different procedures

- Three major bariatric procedures:
  - AGB: least-invasive
  - RYGB: current 'gold standard'
  - VSG: new, less-drastic procedure



- Existing evidence suggests that RYGB and VSG have better long-term T2DM outcomes than AGB

- Theoretically, RYGB should have the most profound effects on T2DM
  - glycemic control driven by non-weight-loss mediated changes in gut hormones
  - AGB and VSG are primarily restrictive

Q: Relative differences in long-term T2DM outcomes are due to weight-dependent or weight-independent mechanisms?

- Two EHR-based NIH-funded R-01s

- **PROMISE**
  - Examine the relationship between diabetes control, remission, and relapse after RYGB on:

    Aim 1: microvascular outcomes

    Aim 2: macrovascular outcomes
  - Four sites in the HMORN with $\approx$ 10,000 patients
  - Comparison with a matched non-surgical 'cohort'

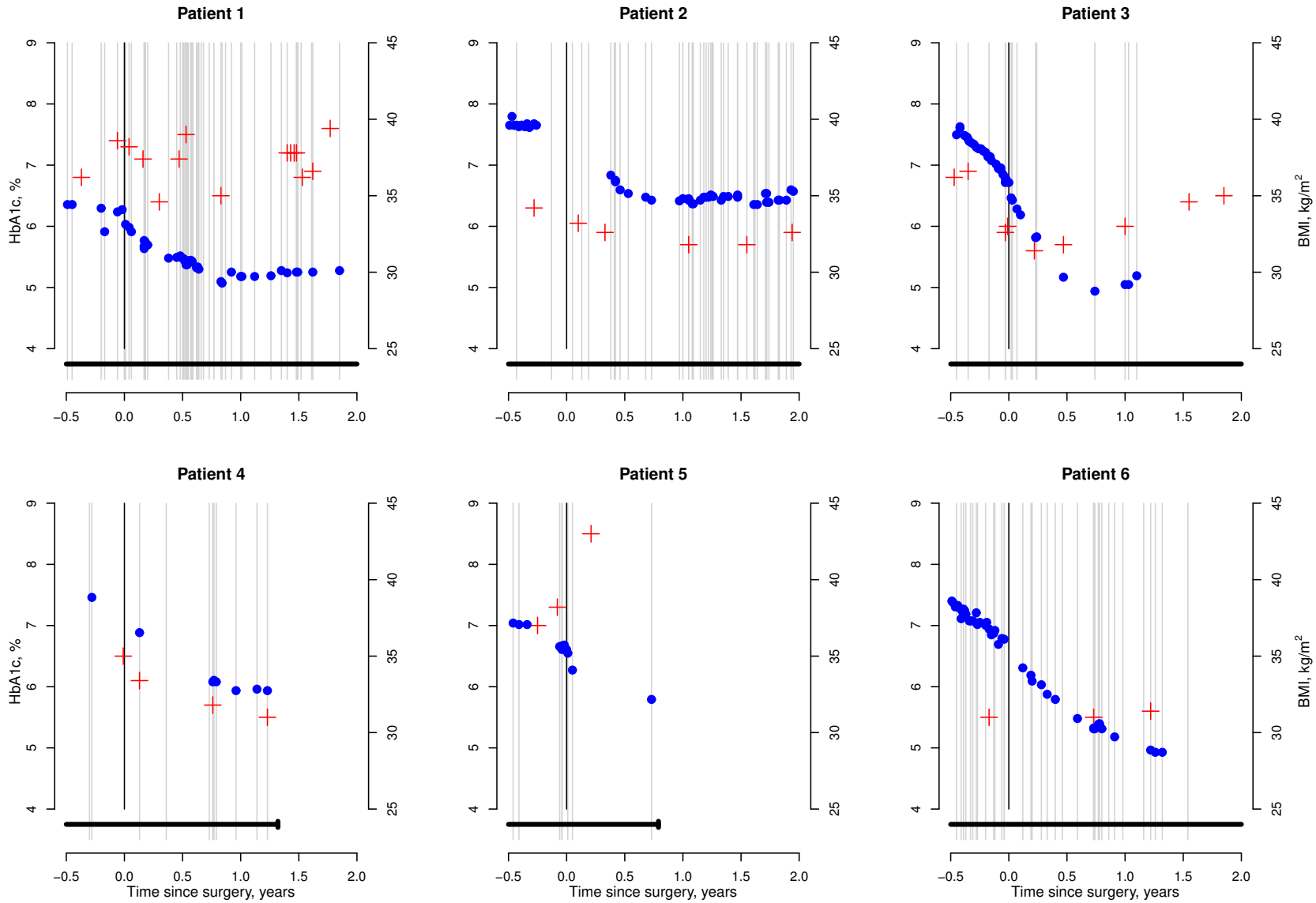
- **DURABLE**
  - Specific aims:

    Aim 1: long-term complications

    Aim 2: hypertension and chronic kidney disease

    Aim 3: exploration of potential bias
  - $\approx$ 45,000 patients who underwent RYGB, VSG or AGB
    - $\approx$ 28,000 patients with $\geq$ 5 years follow-up

- Data from a sample of 6 patients from DURABLE:

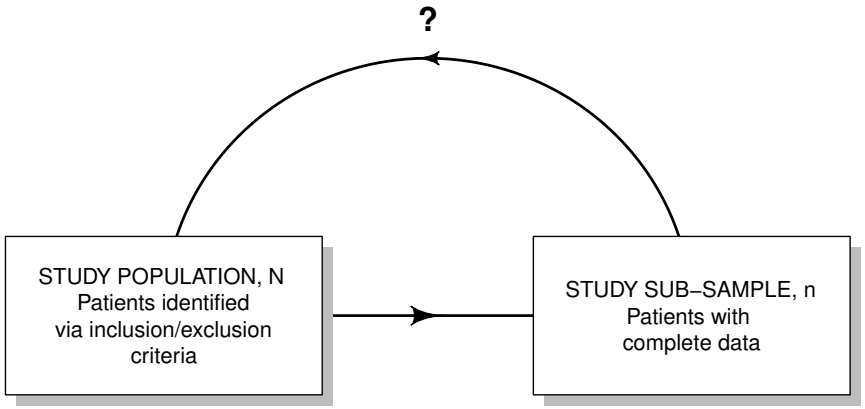


Selection bias due to missing data

- Data collected in EHRs are seldom (if ever) collected for research purposes
  - clinical and/or billing purposes
  - complex patterns of observed information
  - heterogeneity across patients
- May result in a substantial number of patients with insufficient information to be included in the analysis
- Consider a study of long-term change in BMI, comparing RYGB and VSG
  - data from DURABLE
  - focus on 5-year outcomes
  - 16,282 patients who underwent surgery between 1997-2010
  - only 6,206 (38%) BMI measurements at the time of surgery and 5-years post-surgery

Selection bias

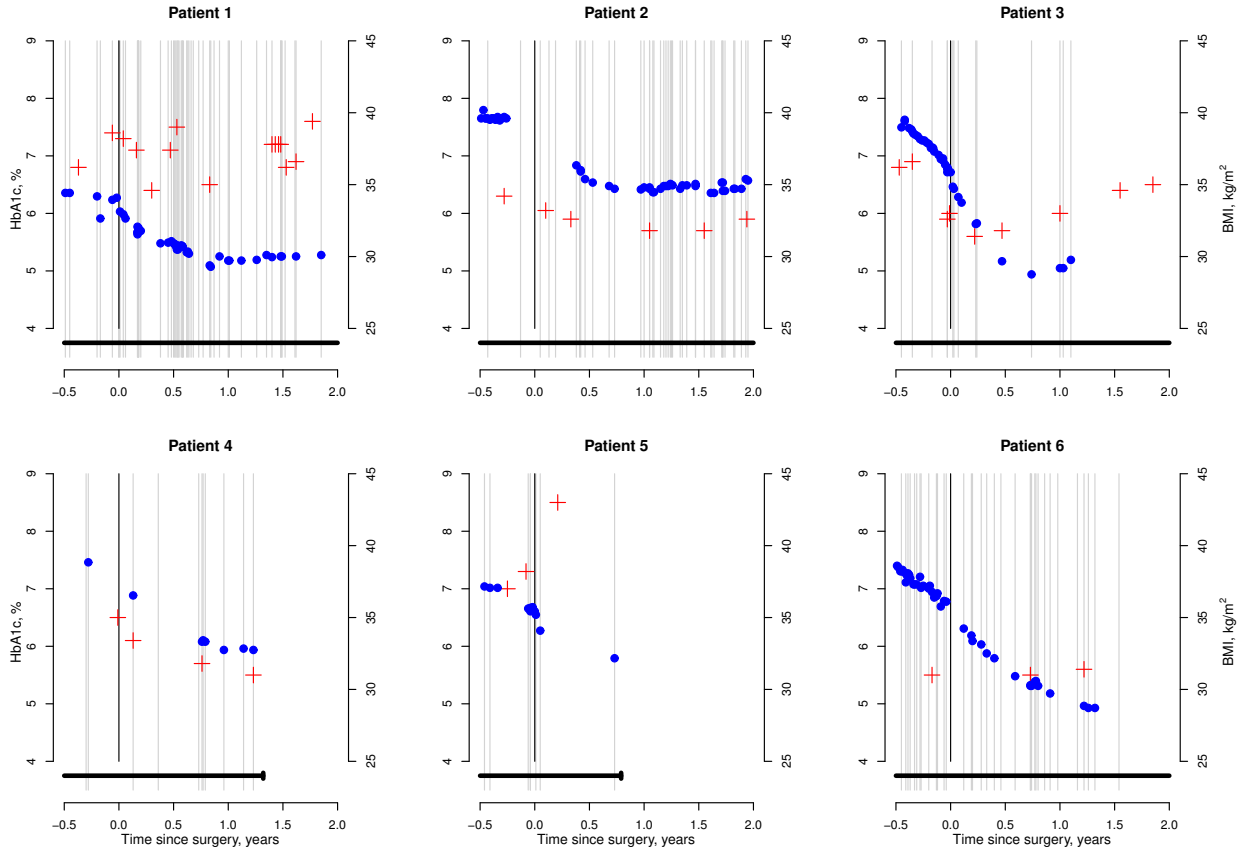
Q: If we restricted analyses patients with 'complete' data, how representative/ generalizable would the results be?



- Of concern is that a naïve analysis may be subject to *selection bias*
  - results may not be externally generalizable
  - Hernan et al (*Epidemiology*, 2004)
- Distinct phenomenon from confounding bias
  - speaks to internal validity
  - Haneuse (*Medical Care*, 2016)

- Viewing selection bias as a missing data problem one could appeal to the (huge) literature on methods for missing data
  - IPW, MI, DR, PMM
- In the sub-study of RYGB vs VSG, one natural way forward would be to model the BMI entire trajectory over the course of time
  - e.g. fit flexible hierarchical model of BMI as a function of time
- Appealing in the sense that one would:
  - make the 'most' use of all of the available data
    - i.e. use intermediate BMI measurements
  - borrow strength between and within patients
- One drawback of this approach, however, is that the model would likely be large and complex
  - challenging to specify and fit
  - sensitivity to functional form and/or distributional assumptions?

- A second drawback is that the notions of 'complete' data and 'missing' data are rendered unclear



- Implications for the assessing missing data assumptions?
  - validity of any given analysis relies, in one way or another, on the *missing at random (MAR)* assumption

The single mechanism approach

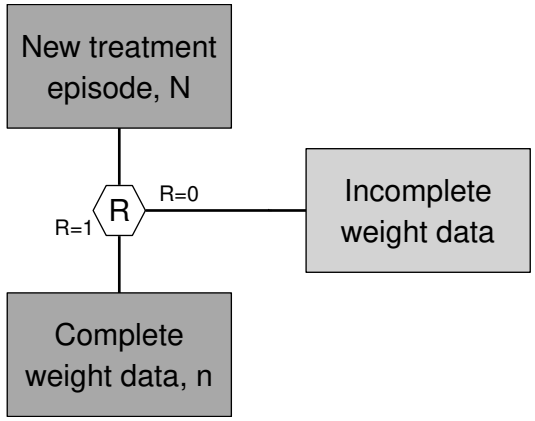
- Assessment of the plausibility of MAR typically proceeds by

  (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

- Operationally, this might be achieved by considering determinants of $R = 0/1$, an indicator of whether an individual study unit is observed to have 'complete' data



- In the EHR context, such a 'single mechanism' approach fails to acknowledge/recognize:

  (i) the inherent complexity of (most) clinical contexts
    - interplay between decisions made by patients and their health care providers

  (ii) the time-varying nature of many factors that influence decisions

  (iii) the heterogeneity within and between systems

  (iv) the motivation and incentives for the collection of data are not research-oriented

The proposed framework

- Moving forward, we propose that researchers initially consider and apply three 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 addressing selection bias with the question:

  what data are observed and why?

  * sometimes referred to as the *data provenance*
  * means of considering missing data assumptions

  (3) Apply appropriate statistical analysis methods

- (1) and (2) are laid out in Haneuse and Daniels (*eGEMS*, 2016)

1. 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
- Primary outcome in the DURABLE sub-study: *BMI change at 5 years*
  - arguably only need BMI information at two time points
- Note, an alternative scientific goal may have been to characterize the BMI trajectory of patients during the 5 years post-treatment initiation
  - intermediate BMI measurements, depending on the level of granularity

2. Consideration of data provenance

- The key benefit of going through the process of specifying the ideal study is that it renders meaningful the notions of 'complete' data' and 'missing' data
- Armed with this one 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 association

- Propose a general strategy based on *modularizing* the 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

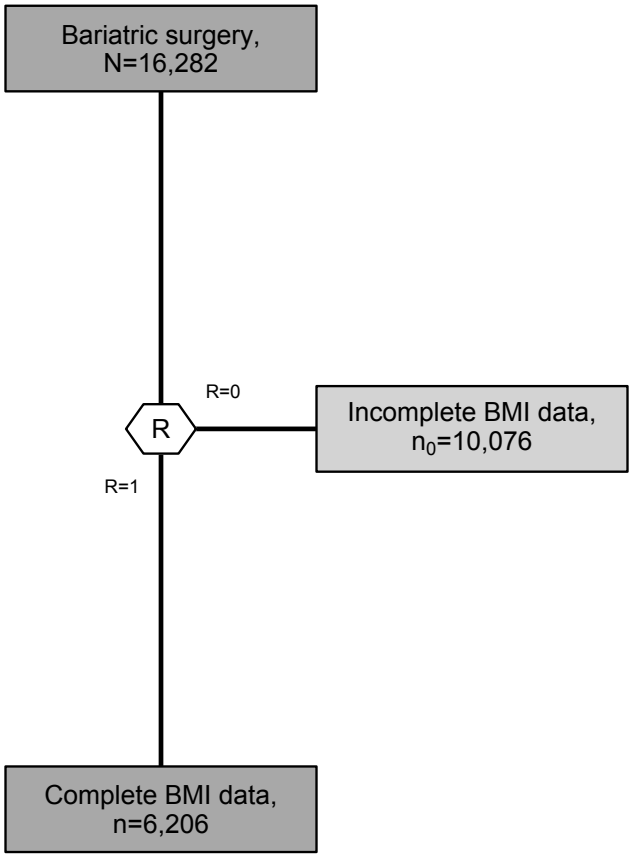
- In the example question on outcomes at 5 years post-surgery, for a patient to have 'complete' BMI data they must (at least):

  (i) have a BMI measurement recorded in the EHR at the time of surgery
  - or 'close' to it

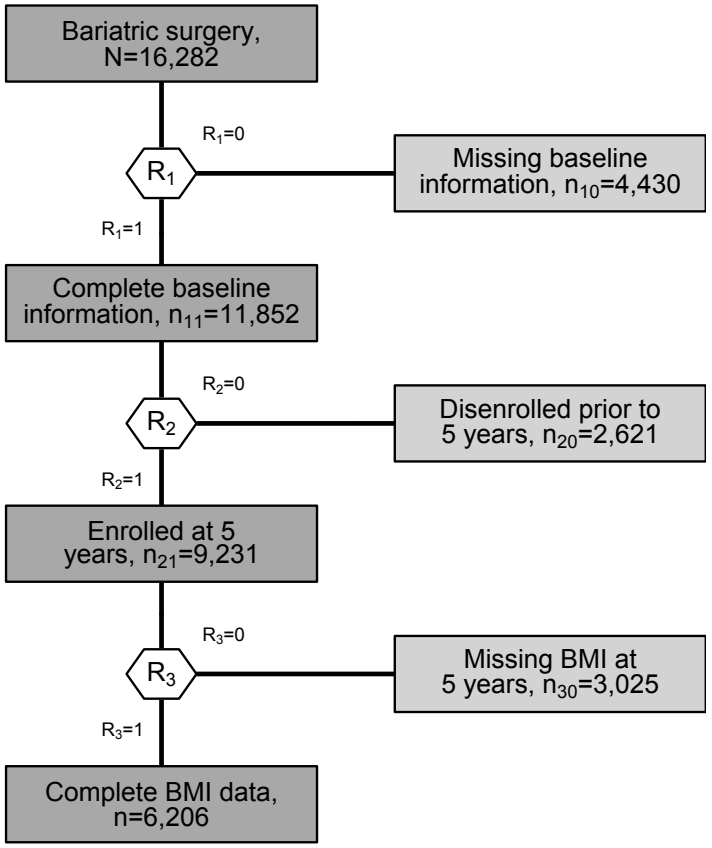
  (ii) be actively enrolled at 5 years

  (iii) had a BMI measurement recorded in the EHR during an encounter at 5 years
  - or 'close' to it

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



(a) Simple specification



(b) Modularized specification

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
  - EHR systems are incredibly heterogeneous
- Whatever structure is adopted, for each sub-mechanism one would need to consider a broad range of factors for each mechanism
- Should be open to the possibility that specific factors may differ across mechanisms in either the direction or magnitude of association
- Also should be open to the possibility that MAR does not hold for each sub-mechanism
  - i.e. some may be MNAR

Moving forward

- Conceptually, the proposed strategy provides a scalable framework 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

Estimation/inference

- Suppose K sub-mechanisms are specified
- Let $R_{k,i} = 0/1$ be an indicator of the 'positive' state required for the k^{th} sub-mechanism such that the observed data is complete for the i^{th} patient
- In the DURABLE sub- study, for example,
  - $R_{1,i} = 1$: BMI measurement at the time of surgery
  - $R_{2,i} = 1$: enrolled at 5 years post-surgery
  - $R_{3,i} = 1$: BMI measurement at 5 years
- Represent the probability that the i^{th} patient will have 'complete' data as:

$$\pi_i \;=\; \Pr(R_{1,i} = \ldots = R_{K,i} = 1)$$

  - i.e. all K have to be in the 'positive' state

- In some settings it will be reasonable to write:

$$\begin{aligned}\pi_i &= \prod_{k=1}^{K} \pi_{k,i} \\ &= \prod_{k=1}^{K} \Pr(R_{k,i} = 1 |\ R_{1,i} = \ldots = R_{k-1,i} = 1)\end{aligned}$$

- Hinges on *monotonicity* of the sub-mechanisms
  - view the K decisions as being sequential in time
  - in many instances it may hold naturally
  - in other instances, some restrictions might be needed
- Decomposition has its roots in the work of Robins et al (*JASA*, 1995)
  - dropout in longitudinal studies based on a fixed # of time points
- One could then proceed by specifying models for each $\pi_{k,i}$ as a function of sub-mechanism specific parameters, $\boldsymbol{\alpha}_k$

- Operationally, for some k, it will make sense to frame modeling $\pi_{k,i}$ as a problem in survival analysis

$$\pi_{k,i} \equiv \Pr(T_{k,i} > \tau_k | \boldsymbol{Z}_{k,i})$$

  * e.g. enrollment status as 5 years

- In other instances, it will make sense to use a GLM:

$$\pi_{k,i} = g_{\pi,k}^{-1}(\boldsymbol{Z}_{k,i}^T \boldsymbol{\alpha}_k)$$

  * e.g. whether the patient had a BMI measurement

- For each sub-mechanism, estimation of $\boldsymbol{\alpha}_k$ can proceed in the usual way
  * based on those patients for whom $R_{k-1,i} = 1$
  * solve

$$\mathbf{M}_k(\boldsymbol{\alpha}_k) = \sum_{i=1}^{N} R_{k-1,i} \mathbf{M}_{k,i}(\boldsymbol{\alpha}_k) = \mathbf{0}$$

    where $\mathbf{M}_{k,i}(\cdot)$ is an appropriate estimating function

- One could then estimate $\boldsymbol{\beta}$ by solving:

$$\mathbf{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_K) \;=\; \sum_{i=1}^{N} \Delta_i(\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_K)\mathbf{U}_i(\boldsymbol{\beta}) \;=\; \mathbf{0}$$

  where

$$\Delta_i(\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_K) \;\equiv\; \prod_{k=1}^{K} R_{k,i}\pi_{k,i}^{-1}(\widehat{\boldsymbol{\alpha}}_k)$$

  * denote the resulting estimator as $\widehat{\boldsymbol{\beta}}$

- Under suitable regularity conditions, one can use a Taylor series expansion and the CLT (van der Vaart, 2000) to show that:

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \;\longrightarrow\; \text{MVN}(\mathbf{0},\; \boldsymbol{J}^{-1}\boldsymbol{\Gamma}\boldsymbol{J}^{-1})$$

  * specific forms of $\boldsymbol{J}$ and $\boldsymbol{\Gamma}$ are relatively straightforward to obtain

Bias-variance trade-off

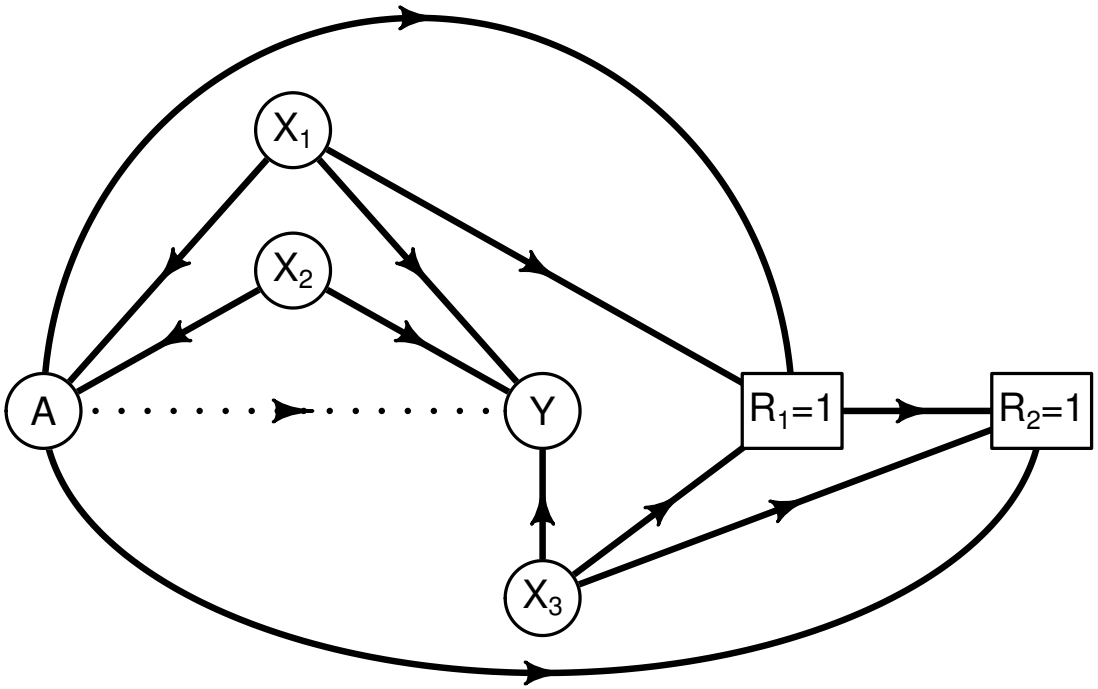
- An important caveat is that the extra 'work' involved may require researchers to possibly contend with a *bias-variance trade-off*

- 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
  - i.e. expect that asymptotic variance of $\widehat{\beta}$ to be larger than that of $\widetilde{\beta}$

- 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) $\quad E[Y] = \beta_0 + \beta_a A + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Model (2) $\quad \text{logit } P(R_1=1) = \alpha_{10} + \alpha_{1a} A + \alpha_{11} X_1 + \alpha_{13} X_3$

Model (3) $\quad \text{logit } P(R_2=1|R_1=1) = \alpha_{20} + \alpha_{2a} A + \alpha_{23} X_3$

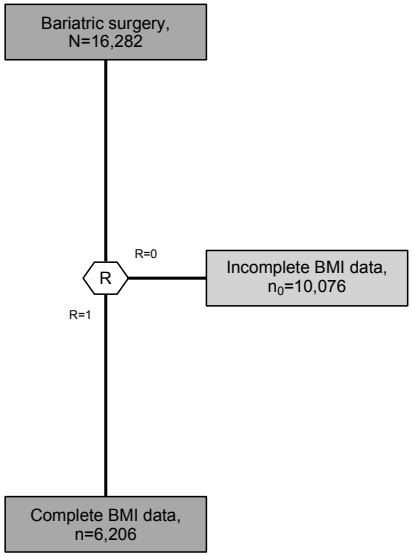
- Results for two scenarios:
  - #1: designed so that the naïve analysis would exhibit moderate bias
  - #2: designed so that the naïve analysis would exhibit little-to-no bias

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

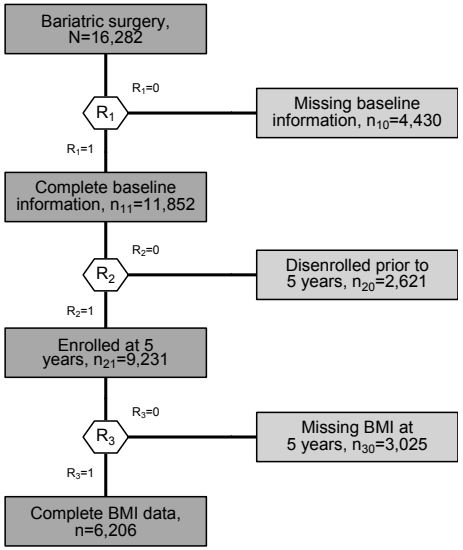
- An important avenue for future work, therefore, will be to characterize (to the extent possible) when the standard single mechanism strategy will be the 'optimal' way forward

DURABLE

- N=16,282 patients who underwent RYGB or VSG surgery between 1997-2010 at one of four Kaiser Permanente study sites
  - consider change in BMI at 5 years post-surgery
  - n=6,206 (38%) patients with 'complete' BMI data



(a) Simple specification



(b) Modularized specification

- Estimated log-ORs and log-HR from fits of missingness (sub-)mechanisms
  - boldface indicates that a 95% CI excluded 0.0

	Standard IPW	**Modularized IPW**		
		BMI at baseline, R_1	Enrolled at 5 years, R_2	BMI at 5 years, R_3
	N=16,262	N=16,262	N=11,852	N=9,231
VSG	-0.63	-1.11	**1.20**	0.66
Age at surgery	**0.08**	**0.03**	**-0.08**	**0.05**
Year of surgery	**0.27**	**1.00**	0.01	**-0.13**
Prior enrollment	**0.25**	0.00	**-0.44**	**0.09**
Male	**-0.26**	**-0.14**	**0.20**	**-0.20**
Site: KP-NC	-0.06	-0.13	0.03	-0.02
Site: KP-SC	**0.41**	**0.72**	0.00	**0.39**
Baseline BMI	–	–	0.02	-0.00
Pre-surgical BMI slope	–	–	**0.01**	**0.04**
Surgery × Site				
VSG & KP-NC	0.91	0.08	**-1.54**	-0.14
VSG & KP-SC	0.43	1.16	**-1.49**	-0.75

- Linear regression analysis with change in BMI from baseline to 5 years as the outcome

	Complete-case analysis		**IPW analysis**			
			Standard		Modularized	
	Est	95% CI	Est	95% CI	Est	95% CI
Model 1:						
Intercept	-9.5	(-9.6, -9.3)	-9.4	(-9.5, -9.2)	-7.9	(-8.4, -7.4)
VSG	2.1	(1.7, 2.5)	1.8	(1.5, 2.0)	0.4	(-0.1, 1.0)
Model 2†:						
Intercept	-12.8	(-13.3, -12.2)	-12.6	(-13.0, -12.2)	-10.9	(-11.8, -9.9)
VSG	2.9	(2.5, 3.3)	2.9	(2.6, 3.2)	3.1	(2.8, 3.5)
Male	0.6	(0.3, 1.0)	0.5	(0.2, 0.8)	0.4	(0.1, 0.7)
Site: KP-NC	1.5	(1.0, 2.1)	1.0	(0.7, 1.5)	-0.4	(-1.1, 0.4)
Site: KP-SC	1.5	(0.9, 2.0)	1.2	(0.8, 1.6)	-0.3	(-1.0, 0.5)
Year of surgery	0.0	(-0.1, 0.1)	0.0	(-0.1, 0.0)	-0.2	(-0.4, 0.0)
Age at surgery	0.3	(0.2, 0.4)	0.3	(0.3, 0.4)	0.1	(0.1, 0.3)
Pre-surgical BMI slope	-0.6	(-0.6, -0.5)	-0.6	(-0.7, -0.6)	-0.5	(-0.6, -0.4)

† Control for site is with KP-Washington as the referent; year of surgery was centered at 2008; age at surgery was standardized by centering at 45 years and dividing by 5 years.

Concluding remarks

- As EHRs become the norm in clinical practice, researchers will be increasingly drawn to the rich data they provide on large populations

- Statistical analyses can be guided by one of two philosophies:

  (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 appealing in that one can potentially gain statistical efficiency by borrowing strength across time and patients

- Potential drawbacks, however, are that:
  - likely requires the specification of a large, complex outcome model
  - notions of 'complete' data or 'missing' data are not clear
- The second is 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 BMI data?
  - may be a reasonable price to pay