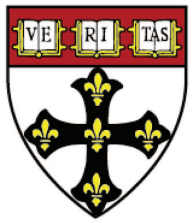
Addressing missing data in comparative effectiveness research using EHR data

Sebastien Haneuse, PhD



Joint with:

Sarah Peskoe, PhD

David Arterburn, MD

Michael Daniels, ScD

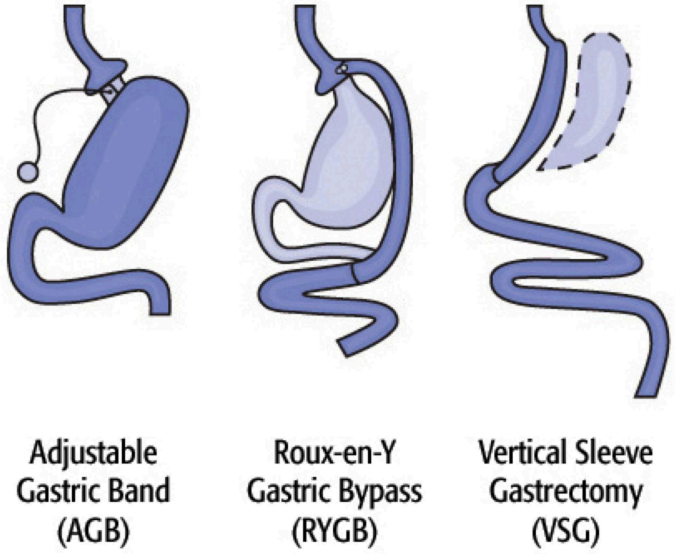
Tony Thaweethai, PhD

Alex Levis, PhD

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
  - endorsed by the American Diabetes Association
- Some controversy remains, especially around long-term outcomes
  - surgery vs. conventional therapy
  - across different procedures

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



- Collaboration over about a 10-year period using EHR data from three Kaiser Permanente (KP) health care systems
  - KP Washington (formerly Group Health Cooperative)
  - KP Northern California
  - KP Southern California

- Two NIH-funded R-01s that gave rise to 12+ publications looking at various long-term outcomes
  - PROMISE and DURABLE

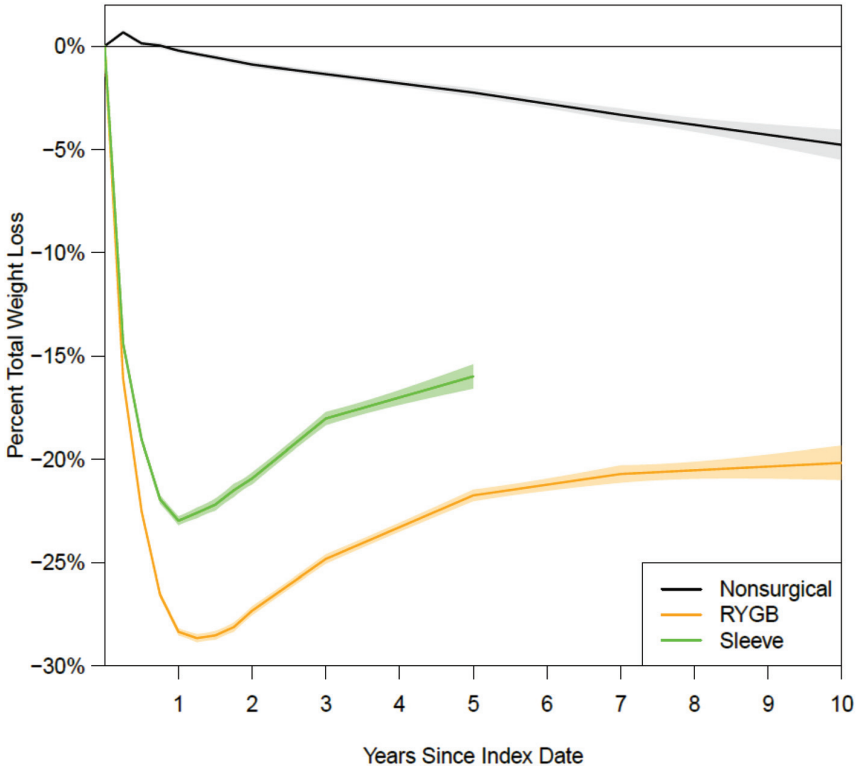
- Arterburn et al (2020, *Annals of Surgery*)

ORIGINAL ARTICLE

Weight Outcomes of Sleeve Gastrectomy and Gastric Bypass Compared to Nonsurgical Treatment

David E. Arterburn, MD,✉ Eric Johnson, MS,* Karen J. Coleman, PhD,† Lisa J. Herrinton, PhD,‡ Anita P. Courcoulas, MD,§ David Fisher, MD,‡ Robert A. Li, MD,‡ Mary Kay Theis, MS,* Liyan Liu, MS,‡ James R. Fraser, BA,* and Sebastien Haneuse, PhD||*

- Matched cohort study
  - $\sim$ 30,000 bariatric cases
  - $\sim$ 90,000 non-cases
- Analyses based on a hierarchical linear mixed model
  - normally-distributed patient-specific random intercepts



EHR data for research

- Numerous well-advertised benefits of using EHR data for research:
  - large patient populations
  - long time periods
  - huge amounts of information
  - readily-accessible and relatively cheap to obtain
- EHR data, however, are not collected for research purposes
- EHR systems are developed, primarily at least, to facilitate:
  - improved clinical care
  - improved tracking/processing of claims

Q: Are EHR data comparable in scope and quality to data that would have been collected by a dedicated study?

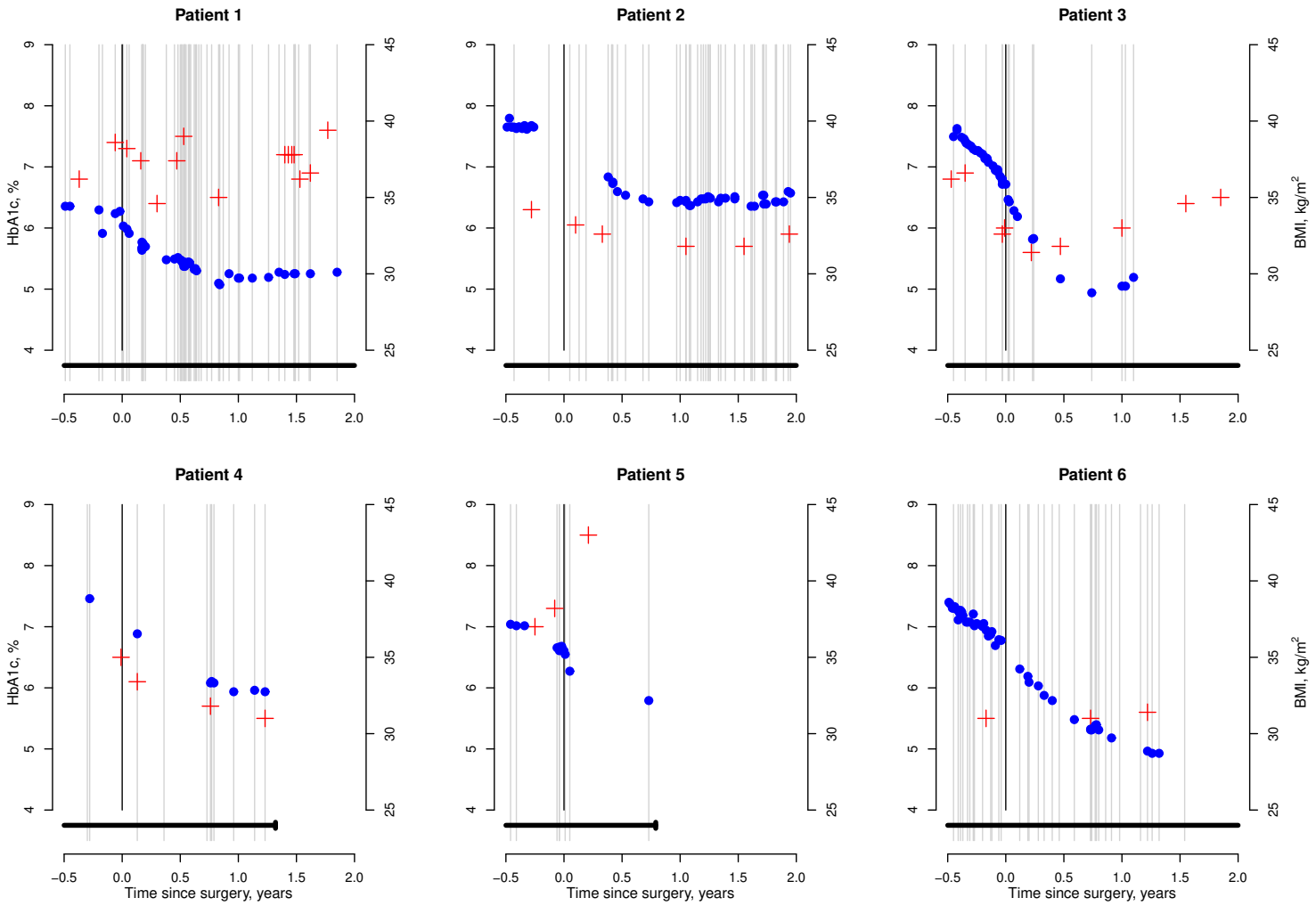
- Many challenges:
  - linkage of patient records across databases
  - extraction of text-based information
  - irregular and inconsistent measurements
  - inaccurate data (i.e. measurement error and misclassification)
  - confounding bias

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

- There is, of course, a massive biostatistical and epidemiologic methods literature, along with statistical software implementations, that we could appeal to

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

- Consider a (hypothetical) study of weight loss at two-years post-surgery
- Data from a sample of 6 patients from DURABLE:



- My sense is that, much of the time, existing, standard methods will, in one way or another, be inadequate and/or unsatisfactory
  - in part because they generally fail to acknowledge the scale, complexity and heterogeneity of EHR data
  - in part because methods are typically developed through the lens of focusing on a single challenge (i.e. in isolation of any other potential challenge)

- There is an emerging literature, on statistical methods that are specifically tailored to research in the EHR setting

- Much, if not most, of this literature has focused on methods towards resolving *confounding bias*
  - e.g. high-dimensional propensity scores

- Other areas that have received attention include
  - probabilistic linkage of records across databases
  - NLP for text-based notes

- The focus of our work is on what we believe to be an under-appreciated problem ... the potential for *selection bias* due to missing data

- How we handled missing data in Arterburn et al (2020, *Annals of Surgery*) involved combining various standard strategies:

  '... 449 (1.3%) were excluded due to missing pre-operative BMI data ...'

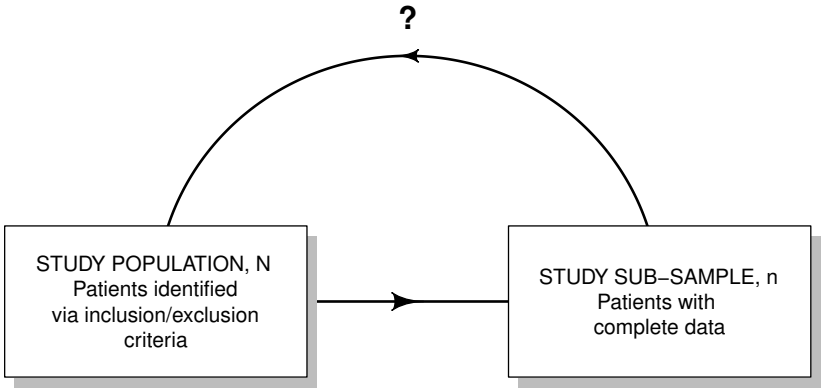
  'To account for the missing data seen at baseline in Table 1 for race/ethnicity, blood pressure, and smoking status, we used multiple imputation via chained equations ... '

- Seems reasonable but also ad-hoc

- Taking a closer look at what was done, it's not really clear what assumptions about the missing data this strategy invoked

Selection bias due to missing data

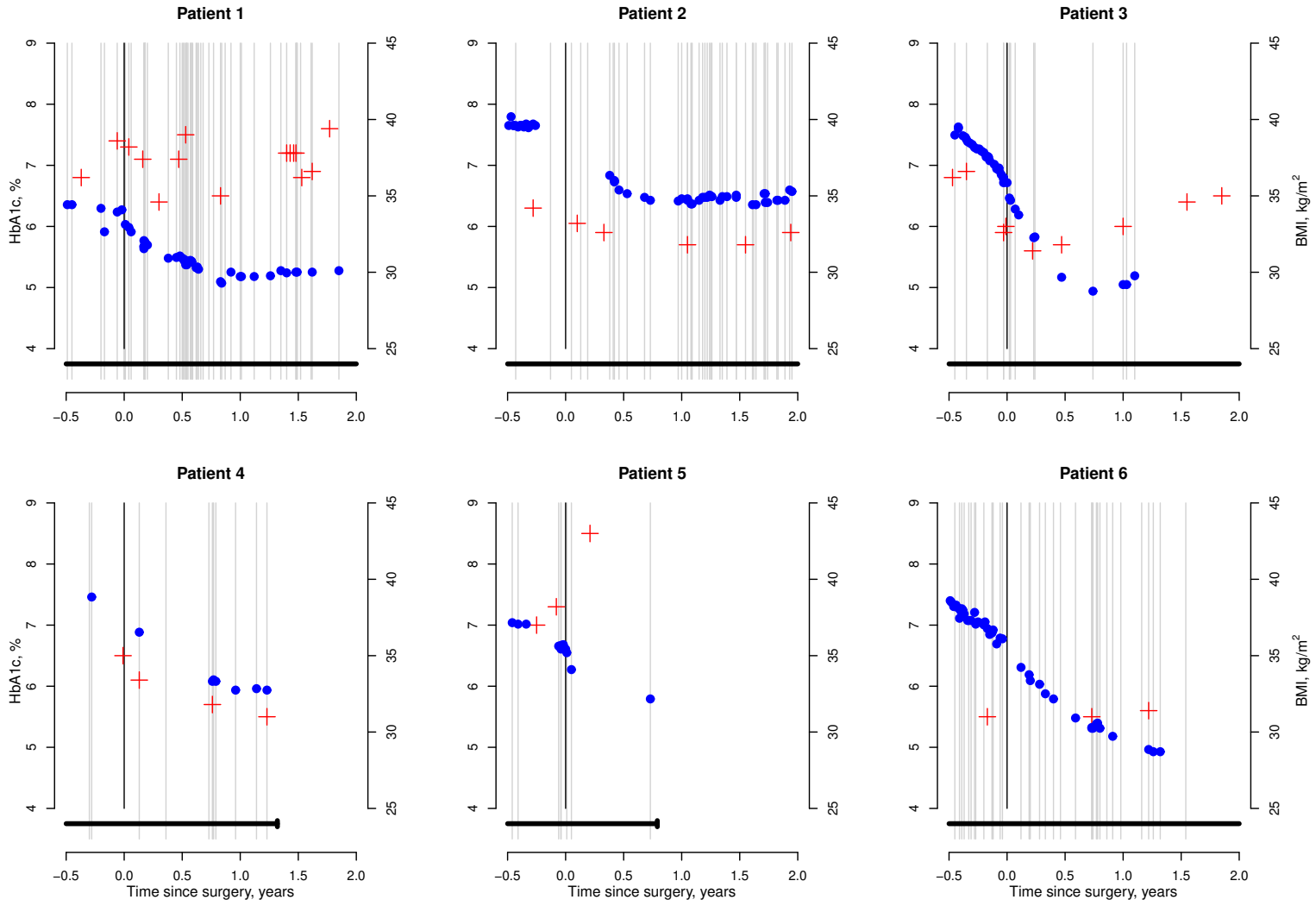
- Selection bias arises when the study sub-sample used in the analyses is not representative, in some way, of the study population to whom we intend to generalize the results



- Speaks to externally generalizable, or lack thereof, of the study results
- Distinct phenomenon from confounding bias
  - speaks to internal validity
  - Haneuse (*Medical Care*, 2016)

- In some cases the potential for selection bias may be evident
- Consider, for example, a 'complete case' analysis where patients with missing or incomplete data are excluded
  - a straightforward (path of least resistance) approach to 'dealing' with missing data
- Generalizabilty of the study results will depend on who is it that as complete data, and why
  - missing data 'assumptions'
- In some settings, however, asking who has 'complete' data may not be straightforward
- Suppose, for example, we wanted to characterize weight loss trajectories in the two-year window following bariatric surgery
  - in the same vein as Arterburn et al (2020, *Annals of Surgery*)

Q: What does 'complete data' mean in this context? Especially given how complex the raw data are?



- An alternative is to say *'let's not restrict in this way but, instead, let's try to make the most of the available data'*
- For example, fit a flexible hierarchical model of BMI as a function of time
  - as we did in Arterburn et al (2020, *Annals of Surgery)*
- Appealing in the sense that one would make the 'most' use of all of the available data
  - i.e. use intermediate BMI measurements
- 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 it is unclear what assumptions regarding the missing data are actually being invoked

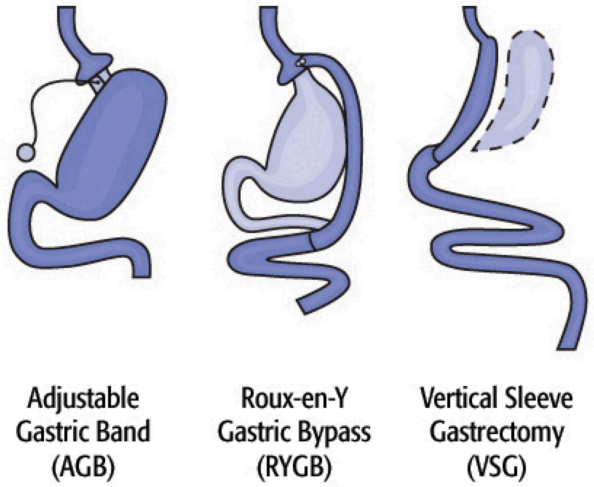
- Intermediate between these two extremes are an almost infinite number of ad-hoc approaches that one could employ, possibly involving combinations of:
  - restriction/exclusion
  - inverse-probability weighting (IPW)
  - multiple imputation (MI)
  - doubly-robust methods (DR)
  - pattern mixture models (PMM)

- Unfortunately, it is not always obvious how one can apply these standard methods in EHR-based studies
  - development is often in idealized settings where missing data is the only issue at-hand
  - unclear what assumptions are being invoked

- With this backdrop, while there is still lots to do (!), the rest of the talk will provide an overview of some of the progress we've made in a number of directions

Modularization of the data provenance

- Suppose interest lies in comparing VSG to RYGB on the basis of two-year weight loss using EHR data



- Formally, let A denote treatment and Y the outcome of interest
  - $A = 0/1 =$ VSG/RYGB
  - Y is weight change at two years
- Acknowledging this is an observational study, suppose a set of factors, $\mathbf{L}$, has been identified as being sufficient for the control of confounding
- The *full data* would consist of an i.i.d sample of size n, with information on $(\mathbf{L}, A, Y)$ for each patient

- Suppose we know what analyses we would perform if we had access to the full data

- Further suppose, however, that we find ourselves in the (admittedly simplified) scenario where A and $\mathbf{L}$ are always observed in the EHR but that Y is only sometimes observed

- Let $R \in \{0, 1\}$ be the indicator for observing Y

- Refer to $(\mathbf{L}, A, R, RY)$ as the *incomplete data*
  - if $R = 1$, then we observe $(\mathbf{L}, A, R = 1, Y)$
  - if $R = 0$, then we observe $(\mathbf{L}, A, R = 0, 0)$

- Given incomplete data, one way forward is to combine the use of whatever full data strategy we had in mind with some approach for 'dealing' with the missing data
  - e.g. use MI at the outset

- In addition to the usual causal assumptions, the validity of such a procedure will hinge on a missing at random (MAR) assumption, such as:

$$\boxed{R \perp\!\!\!\perp Y | \mathbf{L}, A} \tag{1}$$

  - intuitively, whether a value is missing is unrelated to the value itself
  - so MAR would rule out scenarios where patients who do poorly are less likely to disenroll from the health plan

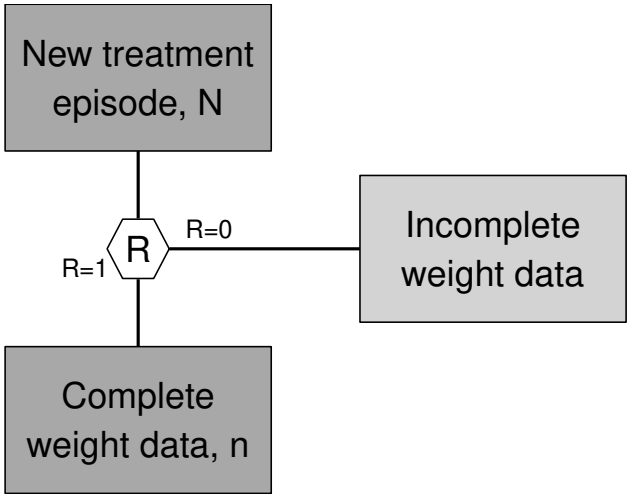
- Assessment of the plausibility of MAR typically proceeds by:

  (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



- In the EHR context, such a 'single mechanism' approach typically fails to acknowledge/recognize:
  - (i) the inherent complexity of (most) clinical contexts
    - $*$ interplay among decisions made by patients and 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

- 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:

  $$\boxed{\textit{what data are observed and why?}}$$

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

  (3) Apply appropriate statistical analysis methods

- 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 study: *BMI change at two 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 two years post-surgery
  - 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

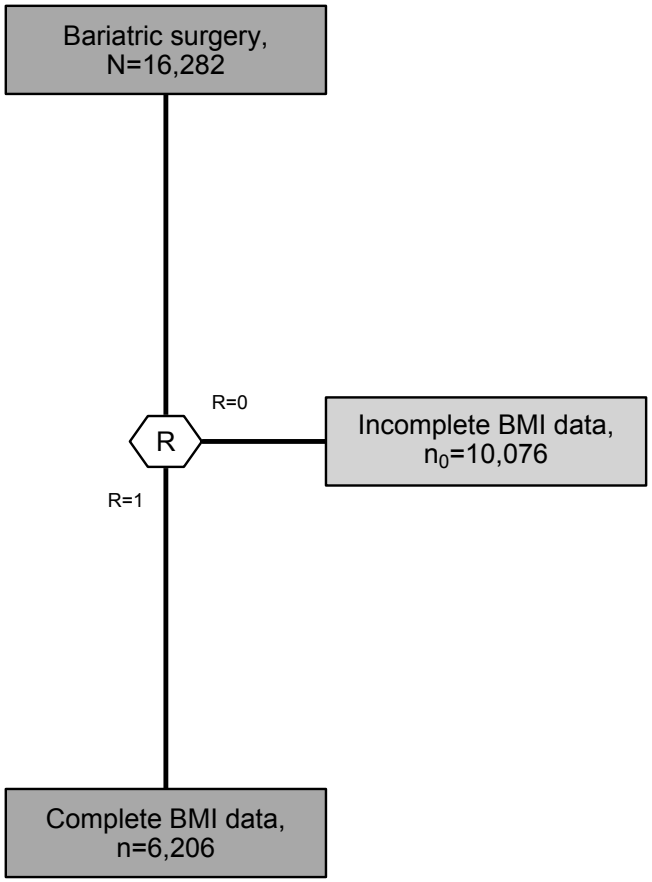
- In the running hypothetical study, for a patient to have 'complete' weight/ BMI data they must (at least):

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

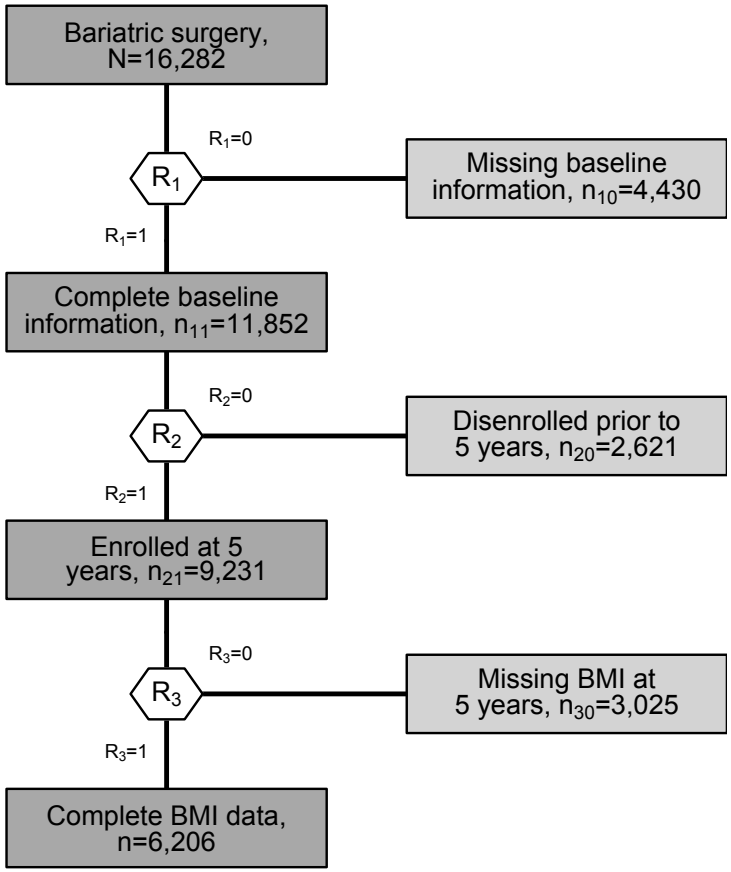
  (ii) be actively enrolled at two years

  (iii) had a weight/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:

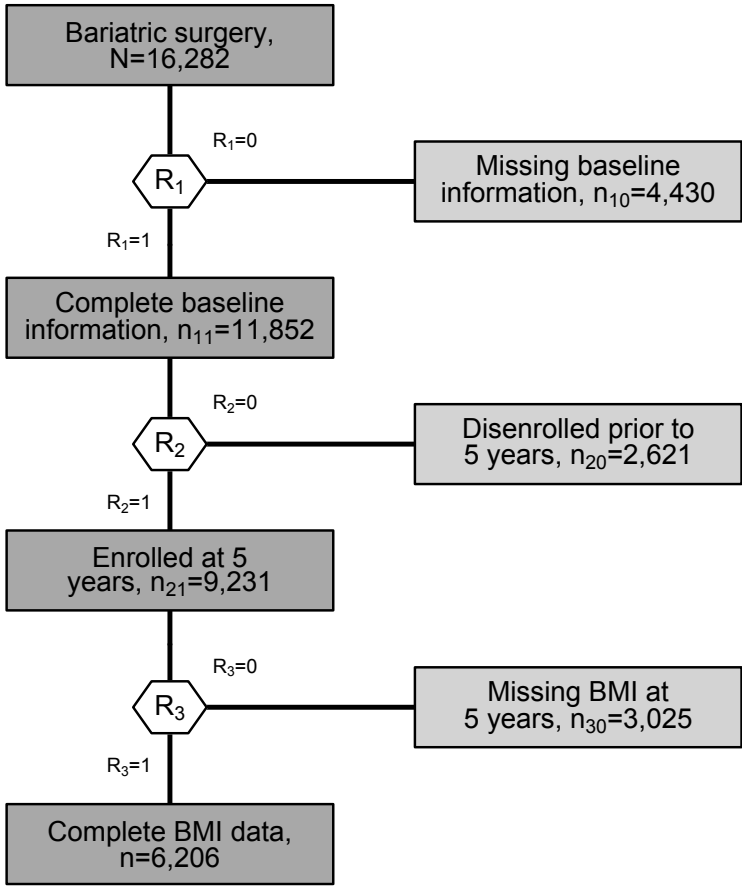
  (i) transparency of assumptions regarding missing data can be enhanced

  (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

- Inverse-probability weighting (IPW)
  - weights for each sub-mechanism
  - e.g. via logistic regressions
  - updated/tailored covariates for each
  - combine for 'overall adjustment
  - Peskoe et al (*SMMR*, 2021)
- Blended analyses
  - use IPW for some mechanisms and MI for others
  - e.g. might be natural to use MI for missing baseline BMI and IPW for dis-enrollment
  - Thaweethai (*PhD*, 2021)



Causal inference in the presence of missing data

- Of course, any EHR-based study will have to contend with confounding bias as well as selection bias
- Towards characterizing 'causation', one could proceed within the counterfactual framework that underpins modern causal inference
- Consider the counterfactual, $Y(a)$, that is the outcome that would be observed had (possibly contrary to the fact) treatment been set at $A = a$
- If, as before $A = 0/1 =$ VSG/RYGB and Y is weight change at the two year mark, we could focus on estimating the *average treatment effect*:

$$\text{ATE} \;=\; \text{E}[Y(1)] - \text{E}[Y(0)]$$

- Progress requires formal consideration of a set of causal inference assumptions:
  - consistency
  - positivity
  - no unmeasured confounding

- Using the full data (i.e. a sample of size n, $(\mathbf{L}, A, Y)$), we could estimate each $\mathrm{E}[Y(a)]$ using any of a number of well-established methods
  - g-formula
  - IPW
  - augmented IPW

- *Causal Inference: What If* (Hernán and Robins, 2020)

- Interestingly, while there are massive literatures on (formal) methods for causal inference and methods for missing data, there is very little at the intersection of the two

- Our broad goals are to see what can be done to establish methods that simultaneously address confounding and selection bias in a wide range of (realistic) settings, and that have desirable properties
  - assumptions are clear
  - optimal statistical efficiency/power
  - robustness

- While we have made progress in a number of scenarios, here I am going to present some recent work on settings where the outcome data are potentially missing not at random (MNAR)
  - also known as informative missingness

The potential for MNAR

- Suppose we find ourselves at the 'design stage', perhaps at the point where we are developing a proposal/grant
- Furthermore, suppose the possibility that the outcome data are MNAR is raised
  - discussion among the collaborators
  - critique from reviewers based on an initial submission
- If it truly is the case that the outcome data are MNAR, then analyses that assume MAR will be compromised/biased
- Unfortunately, by its nature, MNAR is not (typically) testable
- As such, much of the literature on methods for data that are MNAR tends to focus on sensitivity analyses

- A drawback of these *post-hoc* approaches, however, is that there are no guarantees that a 'concrete' conclusion will emerge
  - generally unsatisfying
  - although this may not be a bad thing!

- An alternative philosophy is to engage in additional data collection efforts
  - efforts preemptively tailored to being able (at least partially) to move 'beyond' MNAR

- Rich literature on the design and analysis of studies that involve the collection of additional information on a sub-sample
  - e.g. case-control, case-cohort, two-phase, etc.

- Much of this work has focused on settings where:
  - some particular variable is not readily-available and is 'expensive'
  - concern lies with the mitigation of confounding bias or bias due to missclassification/measurement error

Double-sampling

- Suppose resources are set aside with which the otherwise missing outcome on some sub-sample of patients with $R = 0$ are ascertained
  - e.g. via detailed chart review or a follow-up survey

- Koffman et al (2020, *Obesity Surgery*)

Obesity Surgery
https://doi.org/10.1007/s11695-021-05226-y

IFSO

ORIGINAL CONTRIBUTIONS

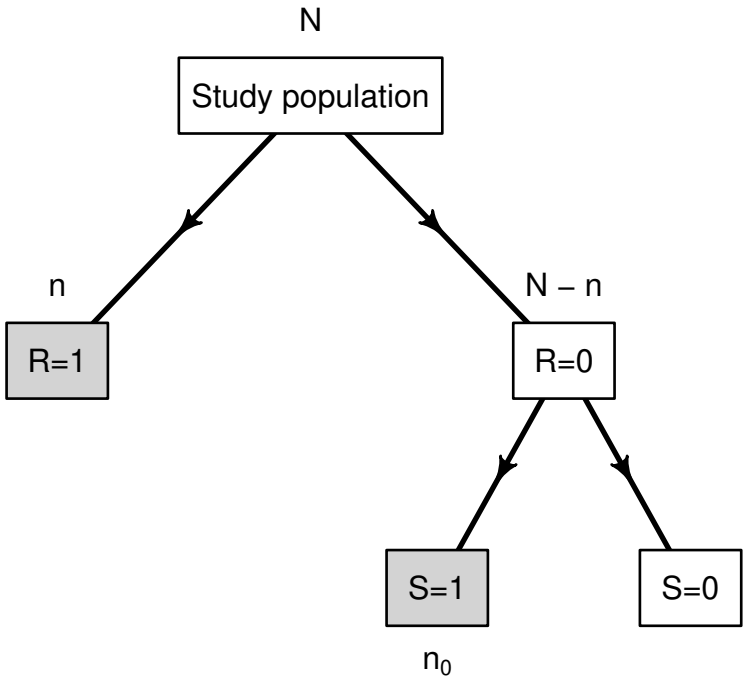
Check for updates

Investigating Bias from Missing Data in an Electronic Health Records-Based Study of Weight Loss After Bariatric Surgery

Lily Koffman[1] · Alexander W. Levis[1] · David Arterburn[2] · Karen J. Coleman[3] · Lisa J. Herrinton[4] · Julie Cooper[2] · John Ewing[2] · Heidi Fischer[3] · James R. Fraser[2] · Eric Johnson[2] · Brianna Taylor[3] · Mary Kay Theis[2] · Liyan Liu[4] · Anita Courcoulas[5] · Robert Li[4] · David P. Fisher[4] · Laura Amsden[4] · Sebastien Haneuse[1]

  - follow-up survey at KP Washington
  - patients who had disenrolled prior to their 5-year anniversary

- Let $S \in \{0, 1\}$ be an indicator of whether a given patient is in the sub-sample



- Expanded terminology/labels related to various data scenarios:

Full data	$(\mathbf{L}, A, Y)$
Incomplete data	$(\mathbf{L}, A, R, RY)$
Final observed data	$(\mathbf{L}, A, R, S, (R+S)Y)$

- Key is that we observe Y on:

  (1) those for whom the value was recorded in the EHR ($R = 1$)

  (2) those who are selected into the sub-sample ($S = 1$)

- Using these additional data requires additional assumptions regarding the double-sampling, in particular S

  * *assumption of non-informative double sampling*
    * selection into the sub-sample is independent of the value of Y
    * similar, in spirit, to the MAR assumption for R
  * *assumption of positivity of double sampling probabilities*
    * everyone for whom their outcome was initially missing has a chance of being 'selected'

Analytic strategies

- After having conducted the double-sampling, we have a wide range of analysis strategies that could be employed, depending on:
  - the data we use in the analysis
  - assumptions one is willing to make
  - the estimator that is employed
- Options include
  - forging ahead as if the original data really were MNAR
  - performing an analysis based on assuming MAR for R is actually ok
  - using an adaptive estimator that decides between the two
- A lot of (theoretical) detail but the long-story-short is that each of these is possible

Concluding remarks

- Key ideas:
  - missing data in EHR-based settings may not be 'business as usual'
  - *modularization of the data provenance*, as a means to better align analyses with the complexity of the data
  - *double-sampling*, as a means to mitigate concerns arising from the potential for data to be MNAR
- Double-sampling work is motivated, in part, by a recently published study in which we actually contacted a number of folks who had missing 5-year outcome data
  - Koffman et al (2021)
  - also collected data on some patients for whom 5-year data was not missing in order to investigate potential recall bias

- We aren't aware of many instances where this strategy is exploited in EHR-based settings
  - may be that researchers bank on the data being 'rich'
- Many practical issues including the potential for recall bias
- Our perspective is that if you are thinking about all this at the 'design stage' then you can proactively attempt to do something, and plan/devote resources accordingly