EDM Forum
EDM Forum Community

eGEMs (Generating Evidence & Methods to improve patient outcomes) Publish

8-31-2016

A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why?

Sebastien Haneuse
Harvard T.H. Chan School of Public Health, shaneuse@hsph.harvard.edu

Michael Daniels
The University of Texas at Austin, mjdaniels@austin.utexas.edu

Recommended Citation

Haneuse, Sebastien and Daniels, Michael (2016) "A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why?," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 4: Iss. 1, Article 16.
DOI: http://dx.doi.org/10.13063/2327-9214.1203
Available at: http://repository.edm-forum.org/egems/vol4/iss1/16

A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why?

Abstract

Electronic health records (EHR) data are increasingly seen as a resource for cost-effective comparative effectiveness research (CER). Since EHR data are collected primarily for clinical and/or billing purposes, their use for CER requires consideration of numerous methodologic challenges including the potential for confounding bias, due to a lack of randomization, and for selection bias, due to missing data. In contrast to the recent literature on confounding bias in EHR-based CER, virtually no attention has been paid to selection bias possibly due to the belief that standard methods for missing data can be readily-applied. Such methods, however, hinge on an overly simplistic view of the available/missing EHR data, so that their application in the EHR setting will often fail to completely control selection bias. Motivated by challenges we face in an on-going EHR-based comparative effectiveness study of choice of antidepressant treatment and long-term weight change, we propose a new general framework for selection bias in EHR-based CER. Crucially, the framework provides structure within which researchers can consider the complex interplay between numerous decisions, made by patients and health care providers, which give rise to health-related information being recorded in the EHR system, as well as the wide variability across EHR systems themselves. This, in turn, provides structure within which: (i) the transparency of assumptions regarding missing data can be enhanced, (ii) factors relevant to each decision can be elicited, and (iii) statistical methods can be better aligned with the complexity of the data.

Acknowledgements

The work presented in this paper was funded, in part, by National Institutes of Health grants R01-MH083671 and R01-183854.

Keywords

2014 Group Health Seattle Symposium, Comparative Effectiveness Research (CER), Electronic Health Record (EHR), Methods, Selection Bias, Missing Data

eGEMs
Generating Evidence & Methods
to improve patient outcomes

A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why?

Sebastien Haneuse, PhD;[i] Michael Daniels, ScD[ii]

ABSTRACT

Electronic health records (EHR) data are increasingly seen as a resource for cost-effective comparative effectiveness research (CER). Since EHR data are collected primarily for clinical and/or billing purposes, their use for CER requires consideration of numerous methodologic challenges including the potential for confounding bias, due to a lack of randomization, and for selection bias, due to missing data. In contrast to the recent literature on confounding bias in EHR-based CER, virtually no attention has been paid to selection bias possibly due to the belief that standard methods for missing data can be readily-applied. Such methods, however, hinge on an overly simplistic view of the available/missing EHR data, so that their application in the EHR setting will often fail to completely control selection bias. Motivated by challenges we face in an on-going EHR-based comparative effectiveness study of choice of antidepressant treatment and long-term weight change, we propose a new general framework for selection bias in EHR-based CER. Crucially, the framework provides structure within which researchers can consider the complex interplay between numerous decisions, made by patients and health care providers, which give rise to health-related information being recorded in the EHR system, as well as the wide variability across EHR systems themselves. This, in turn, provides structure within which: (i) the transparency of assumptions regarding missing data can be enhanced, (ii) factors relevant to each decision can be elicited, and (iii) statistical methods can be better aligned with the complexity of the data.

[i]Harvard T.H. Chan School of Public Health, [ii]University of Texas—Austin

Introduction

Electronic health records (EHR) are playing an increasingly prominent role in comparative effectiveness research (CER) with key benefits including that they often contain rich information on large populations over long time frames, are relatively inexpensive to obtain, and can be updated in near real time.[1-4] Recognizing these, the Institute of Medicine recently released called for increased use of EHR data for research.[5] Notwithstanding their huge potential, however, since EHRs are typically developed for billing and/or clinical purposes, and not with any specific research agenda in mind, researchers must ask whether or not the available EHR data is "research quality."[6-8] This includes consideration of whether all covariates relevant to the research goals are routinely collected in clinical care, whether covariates that are collected are done so consistently across patients and time, and, whether the available data is accurate and error free. Without consideration of these issues, naïve analyses may be subject to numerous biases, the most commonly cited form being confounding bias.[9]

Another important type of bias is selection bias, which arises when some patients identified as being eligible for inclusion in the study are found to have insufficient data to be included in the analyses.[10] Some patients may, for example, have missing baseline treatment, missing clinical information, missing laboratory measurements during follow-up, or have disenrolled from the health plan prior to the end of planned follow-up. Unfortunately, in contrast to confounding bias,[11-21] the control of selection bias in EHR-based settings has received virtually no attention in the literature. This may be due, in part, to the notion that selection bias can be cast as a missing data problem and that statistical methods for missing data are well established[22,23] and can be readily applied to EHR-based CER.[24]

Regardless of the specific method used, a critical step in any analysis involving missing data is the consideration of whether the data are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). In the context of an EHR-based study, this corresponds to addressing the question of why some patients have complete data and others do not. In practice, this is typically operationalized by first conceiving of a so-called missingness mechanism that drives whether or not a patient has complete data, and secondly by determining which factors influence it. This approach may be reasonable in many settings, especially those in which the data collection scheme and design are under the control of the research team. Adopting this approach in the EHR setting, however, likely corresponds to an unrealistically simple view of the missing data. In particular, restricting one's focus to a single mechanism for missingness masks the complex interplay of the numerous decisions made by patients, health care providers, and health systems that, collectively, give rise to the observed data. Naïvely moving forward with this standard approach, therefore, will likely result in a failure to completely control selection bias.

To resolve this challenge, we propose a new, general framework for consideration of selection bias in the complex setting of EHR-based CER. Central to the proposed framework is a shift away from asking "what data are missing and why?" to asking instead "what data are observed and why?". As we will elaborate upon, this shift provides researchers with a natural and intuitive approach to determining and understanding the sequence of decisions that must be made in order for a measurement to be recorded in the EHR and, ultimately, ensuring as thorough a control for selection bias as possible.

Volume 4 (2016) Issue Number 1

Antidepressants and Weight Change

The proposed framework is motivated by challenges we currently face in an ongoing comparative effectiveness study of antidepressant treatment and long-term weight change. Here we briefly describe the study, as well as the potential for selection bias.

Study Setting

The study is being conducted at Group Health, an integrated insurance and health care delivery system in Washington State. As part of its clinical systems, Group Health maintains numerous electronic databases including an EHR based on EpicCare (Epic Systems Corporation of Madison, WI), and a pharmacy database that has recorded all prescriptions dispensed at Group Health–owned pharmacies since 1977. Additionally, electronic databases track demographic information; inpatient treatment and outpatient encounter claims; insurance and enrollment status; and primary care visit appointments.

Study Population

To investigate the relationship between antidepressant choice and weight change, we identified all adults ages 18–65 years with a new episode of treatment for depression between January 2006 and November 2007. New treatment episodes were identified on the basis of a dispensing for an antidepressant medication without the occurrence of any treatment (including psychotherapy) in the prior nine months; thus, only subjects with at least nine months of continuous enrollment prior to the index date were included. Applying these criteria, we identified 9,704 patients in the Group Health EHR.

Primary Outcome

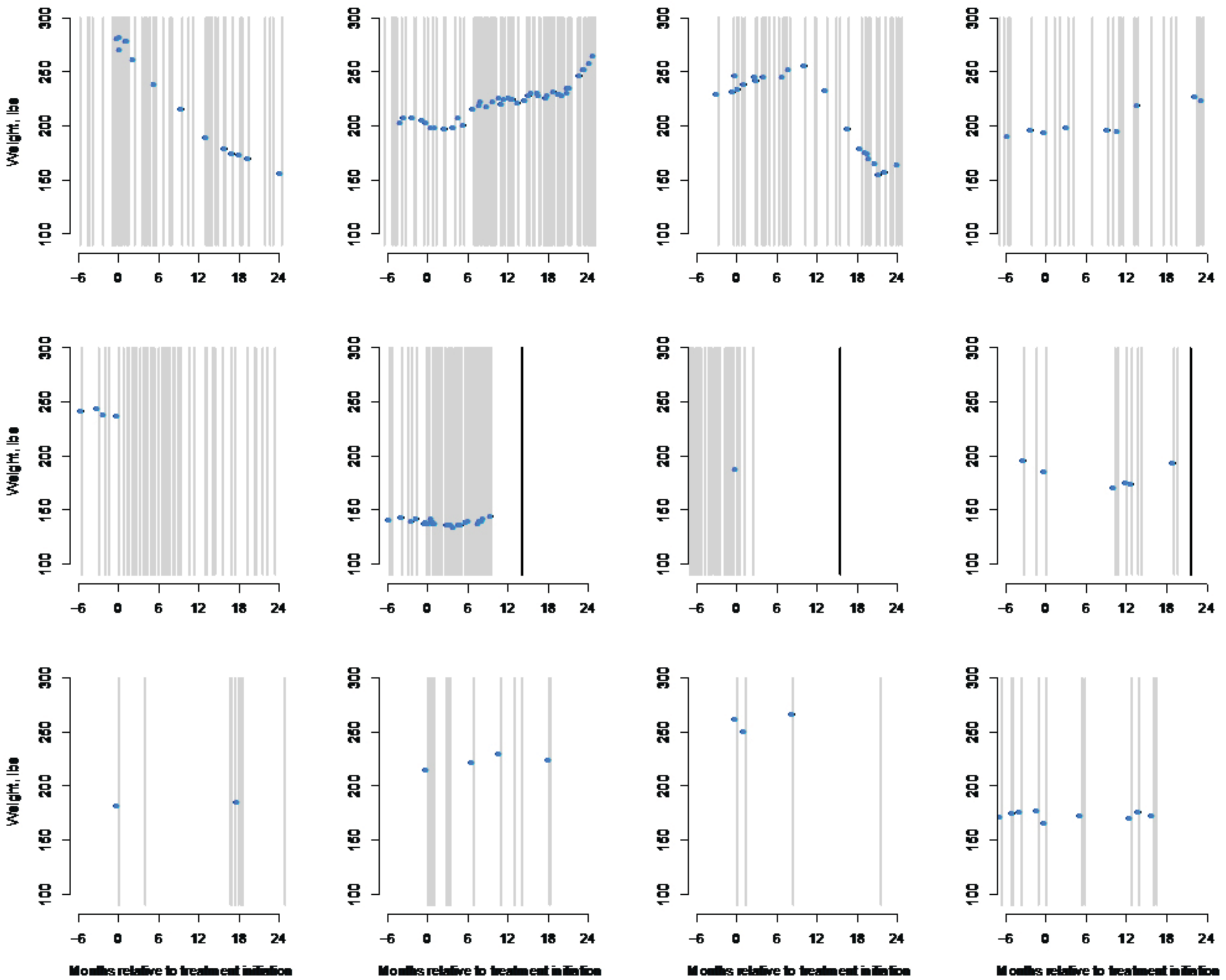
While previous studies indicate that certain antidepressants may reduce body weight (e.g., fluoxetine and bupropion) and others may increase body weight (e.g., paroxetine and mirtazapine), the existing literature is limited in that it focuses on short-term outcomes of 12 months or less.[25-29] Given the lack of evidence regarding long-term outcomes we took weight change at 24 months post-treatment initiation to be the primary outcome.

Weight Information in the Electronic Health Record (EHR)

For each of the 9,704 patients identified by the inclusion- and exclusion criteria, we extracted all records of an outpatient visit for the interval starting 24 months prior to the start of the treatment episode through November 2009. This resulted in 354,945 records from which information—on weight; potential confounders; and auxiliary variables such as age, gender, smoking history, and comorbid conditions—was extracted. Focusing on weight, despite a patient's underlying trajectory following some smooth path over time, the EHR provides only a series of "snapshots" of a patient's weight post-treatment initiation. To illustrate this, Figure 1 provides a graphical summary of the available weight-related information for 12 select patients during the interval starting 180 days prior to the start of the treatment episode and ending at the 24-month marks. Across the panels, gray lines indicate times when the patient had an encounter with the clinical system, the blue dots indicate that a weight measurement was recorded, and the black lines indicate that a patient disenrolled from Group Health.

From the first row of Figure 1, we see that some patients have rich weight-related information in the EHR with numerous encounters with the health system over the 24 month follow-up interval, as well as numerous weight measurements. In contrast, the eight patients in the second and third rows of Figure 1 have relatively sparse weight-related information, with substantial variation in the number of clinical encounters and weight measurements.

Figure 1. Summary of Weight-Related Information for 12 Patients in the Group Health EHR–Based Study of Treatment for Depression and Weight Change



Gray lines indicate when an encounter occurred; blue dots indicate a weight measurement; and black lines indicate that the patient disenrolled prior to the 24-month mark.

The Potential for Selection Bias

With respect to the primary outcome of weight change at 24 months, the "ideal" data scenario would be that complete weight information is available in the EHR at baseline and at 24 months post-treatment initiation for all 9,704 patients identified by the inclusion- and exclusion criteria. To simplify the illustration of potential selection bias we focus on 8,631 patients (88.9 percent) with complete weight information at treatment initiation based on a ±30 day window. Among these patients, only 2,408 (27.9 percent) patients have a valid weight measurement recorded in the EHR at 24 months, (again based on a ±30 day window). Given this significant missingness, and notwithstanding missingness in other important covariates such as confounders, there is clear potential for selection bias. Specifically, naïve analyses based solely on the 2,408 patients with complete data will be biased if they are not representative of the population defined by the inclusion/exclusion criteria.

Cast as a missing data problem, the standard approach to selection bias first conceives of a *missingness mechanism* that drives whether or not a patient has complete data.[10] Figure 2(a) provides a graphical representation, with S=0/1 indicating that a patient has incomplete- or complete-weight information at 24 months. Intuitively, this mechanism can be thought of as corresponding to a particular decision, made by the study participant, such as the decision to drop out from the study. Toward investigating the potential for selection bias and its impact, one could explore determinants of this (binary) decision via logistic regression analyses. Focusing on the 8,631 patients with complete data at baseline, the first three columns of Table 1 indicate that female patients have significantly higher estimated odds of complete weight information at 24 months, as do older patients. Furthermore, whether a patient has complete data at 24 months also depends, in part, on the choice of treatment; and patients with a higher baseline weight are estimated to have higher odds of complete data—odds ratio (OR) 1.05 for a 20-lb. increase in weight; 95 percent confidence interval (CI): (1.03, 1.07).

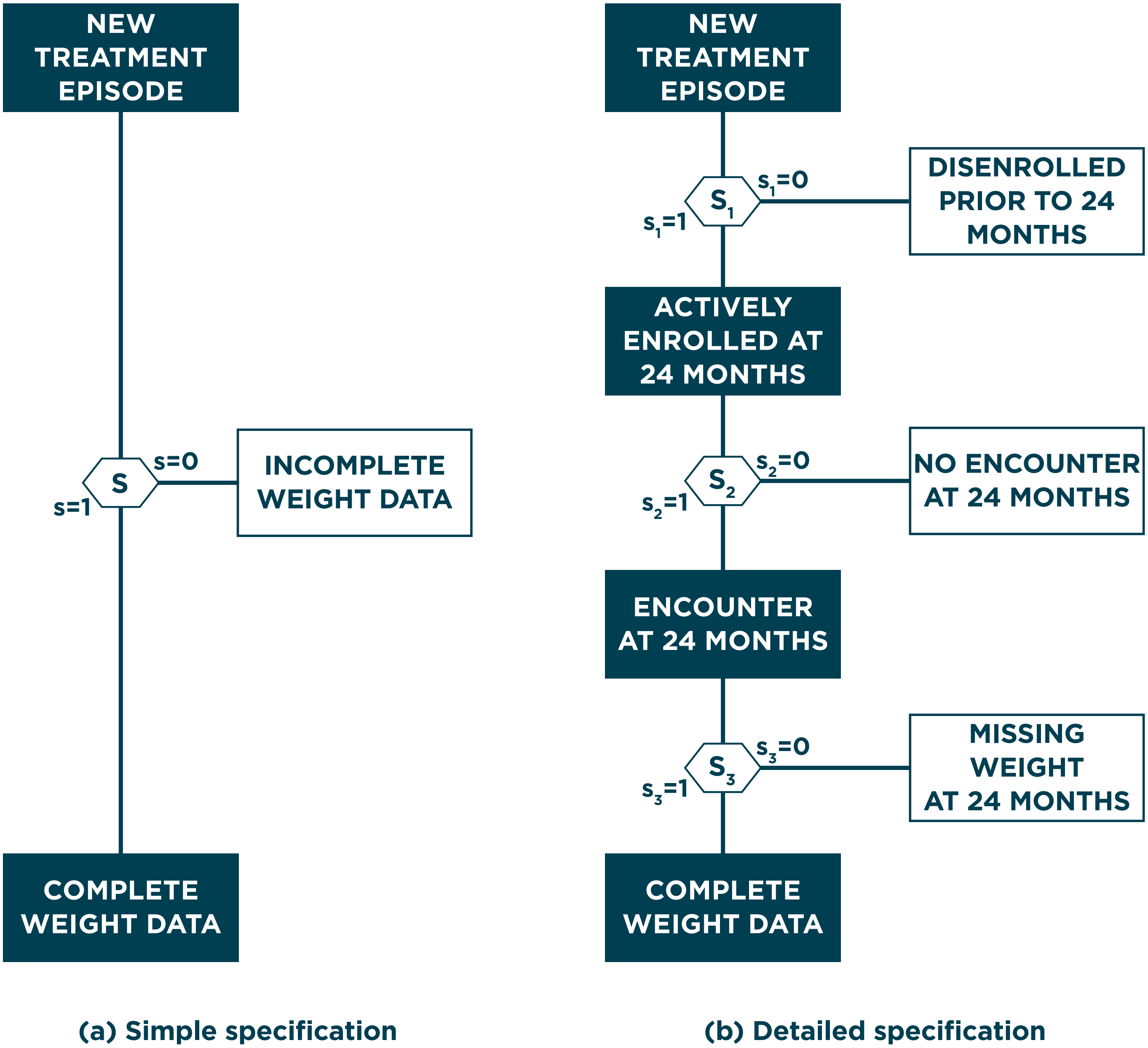
A General Framework

Overall, the preliminary evidence presented so far strongly suggests that a naïve (unadjusted) analysis of the relationship between treatment choice and 24-month weight change will suffer from selection bias. To resolve this one could use multiple imputation,[30,31] inverse-probability weighting,[32] or both.[33] The validity of each of these approaches, however, hinges on the appropriateness of the single mechanism and decision strategy as an approach to evaluating assumptions and performing necessary adjustments. In the clinical contexts that EHRs represent, however, a single mechanism is unlikely to fully characterize the complex set of decisions—made by the patient, their health care provider, and the health care system—that give rise to complete data in the EHR. As such, use of a single mechanism, as illustrated in Figure 2(a), will be unrealistically simple. To resolve this we propose a new framework for addressing selection bias in EHR-based CER, one that acknowledges and integrates the complex set of decisions that give rise to complete data. In the following we provide a detailed discussion of each aspect of the framework; Figure 3 provides an overview of the framework in the form of a process flow.

Consideration of the “Ideal” Study

Central to the proposed framework are two key principles. The first is that researchers initially specify the structure of the data that would have been collected had they had the opportunity to conduct an “ideal” study.[34,35] This specification will depend primarily on the scientific goals of the study, which will, in turn, determine which specific covariates are needed as well as their timing. In the antidepressants study, given the primary interest in weight change at 24 months, such a data structure would at a minimum include weight at baseline and at 24 months. If the primary interest lay with understanding a patient’s trajectory over the first 24 months following treatment initiation, the ideal data structure would additionally include intermediary weight measurements, the timing of which would depend on the desired level of granularity. Beyond outcome information, the data structure for the ideal study would also include covariates necessary for

Figure 2. Alternative Specifications for Observance of Complete Weight Data at 24 Months Post-treatment Initiation



Panel (a) corresponds to the traditional, single mechanism approach to selection bias; panel (b) corresponds to one possible implementation of the proposed framework that acknowledges the complexity of EHR data.

Table 1. Results from Logistic Regression Analyses Examining the Association Between Select Patient-Level Characteristics and Whether or Not a Patient Has Complete Weight Information at 24 Months

	SINGLE MECHANISM: WEIGHT DATA AT 24 MONTHS			SUB-MECHANISM NO. 1: ACTIVE ENROLLMENT AT 24 MONTHS			SUB-MECHANISM NO. 2: ENCOUNTER AT 24 MONTHS ±30 DAYS, GIVEN ENROLLMENT AT 24 MONTHS			SUB-MECHANISM NO. 3: WEIGHT MEASURED AT 24 MONTHS ±30 DAYS, GIVEN AN ENCOUNTER AT 24 MONTHS ±30 DAYS		
	(N=8,631)			(N=8,631)			(N=6,570)			(N=3,688)		
	OR	95% CI	P	OR	95% CI	P	OR	95% CI	P-VALUE	OR	95% CI	P
Female	1.33	(1.19, 1.48)	<0.001	1.11	(0.99, 1.24)	0.070	1.30	(1.16, 1.45)	<0.001	1.20	(1.03, 1.40)	0.022
Age*	1.16	(1.12, 1.21)	<0.001	1.41	(1.35, 1.46)	<0.001	1.10	(1.05, 1.14)	<0.001	0.97	(0.92, 1.03)	0.277
Antidepressant												
Fluoxetine	1.00	REF	<0.001	1.00	REF	<0.001	1.00	REF	<0.001	1.00	REF	0.014
Buproprion	1.05	(0.89, 1.22)		1.01	(0.86, 1.19)		1.14	(0.97, 1.33)		0.92	(0.74, 1.15)	
Mirtazapene	1.29	(0.68, 2.45)		0.94	(0.46, 1.93)		1.18	(0.59, 2.33)		1.54	(0.55, 4.31)	
Paroxetine	1.09	(0.83, 1.42)		1.27	(0.95, 1.72)		1.19	(0.91, 1.55)		0.83	(0.58, 1.19)	
SSRI	0.90	(0.78, 1.03)		0.87	(0.76, 1.00)		1.08	(0.95, 1.24)		0.79	(0.65, 0.95)	
SARI	1.51	(1.26, 1.80)		1.36	(1.09, 1.69)		1.59	(1.32, 1.93)		1.09	(0.85, 1.39)	
Tricyclics	1.80	(1.53, 2.11)		1.33	(1.09, 1.63)		1.93	(1.61, 2.32)		1.28	(1.01, 1.61)	
Weight at baseline*	1.05	(1.03, 1.07)	<0.001	1.02	(1.00, 1.04)	0.068	1.03	(1.01, 1.05)	0.003	1.05	(1.02, 1.08)	<0.001

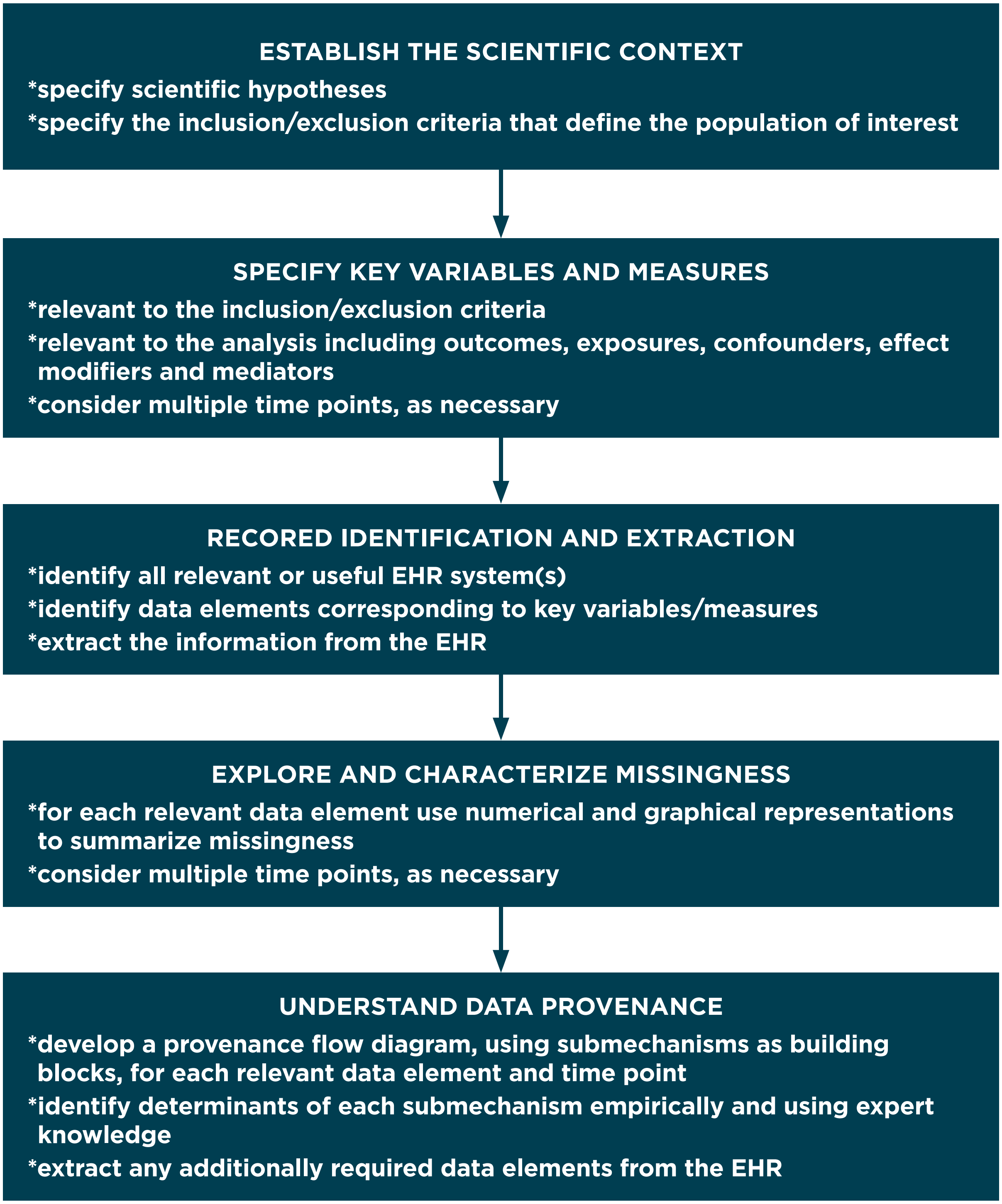
Note: *Age odds ratio is for a 10-year contrast; weight at baseline OR is for a 20 lb. contrast. OR = odds ratio; CI = confidence interval.

the implementation of inclusion/exclusion criteria, covariates relevant to characterizing treatment choice, and potential confounders and effect modifiers. Practically, specification of this structure could be approached much in the same way that researchers approach detailing data collection strategies in grant proposals.

Consideration of Data Provenance

Given an ideal data structure, the second key principle is that researchers frame the task of controlling selection bias with the question "what data are observed and why." Crucially, in doing so researchers can readily breakdown the complex process that governs whether or not a patient has complete data, referred to as the "provenance" of the data,[36] into a series of more manageable components or sub-mechanisms. Toward this, given the wide variation in study questions addressed in CER, as well as the heterogeneity in EHR systems, one cannot, unfortunately, be prescriptive in this task; no single set of sub-mechanisms will be appropriate or sufficient for all studies. Nevertheless, Table 2 provides a list of sub-mechanisms that researchers could consider, each of which is accompanied by one or more contextual questions that could be used to determine its relevance for a particular study. Prior to describing them, we emphasize that each sub-mechanism could be considered for any given data element identified as being relevant for the ideal study (i.e., the outcome, treatment, and confounders, possibly measured at different time points).

Figure 3. Process Flow Representation of the Proposed Framework for Selection Bias in EHR-Based Studies That Can Be Used in Conjunction with Table 2



ESTABLISH THE SCIENTIFIC CONTEXT

*specify scientific hypotheses

*specify the inclusion/exclusion criteria that define the population of interest

SPECIFY KEY VARIABLES AND MEASURES

*relevant to the inclusion/exclusion criteria

*relevant to the analysis including outcomes, exposures, confounders, effect modifiers and mediators

*consider multiple time points, as necessary

RECORED IDENTIFICATION AND EXTRACTION

*identify all relevant or useful EHR system(s)

*identify data elements corresponding to key variables/measures

*extract the information from the EHR

EXPLORE AND CHARACTERIZE MISSINGNESS

*for each relevant data element use numerical and graphical representations to summarize missingness

*consider multiple time points, as necessary

UNDERSTAND DATA PROVENANCE

*develop a provenance flow diagram, using submechanisms as building blocks, for each relevant data element and time point

*identify determinants of each submechanism empirically and using expert knowledge

*extract any additionally required data elements from the EHR

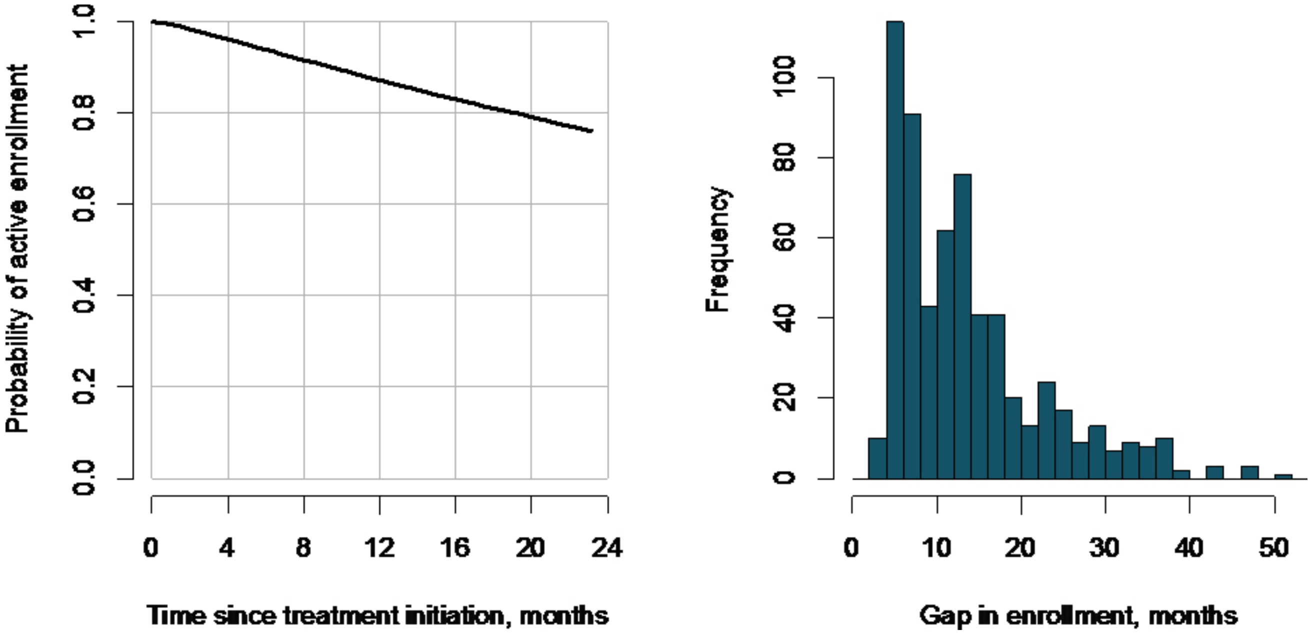
The first sub-mechanism in Table 2 refers to a patient's enrollment status. In some settings, for a measurement to be recorded in the EHR, the patient must have been actively enrolled in some specific health plan system. In other settings, enrollment status may be less relevant or not at all. For example, since all individuals 65 years and older in the United States are automatically enrolled in Medicare, studies using Medicare claims data need not consider enrollment status once an individual is 65 years of age. If enrollment status is relevant, one may find that some patients have multiple periods of enrollment during the observation period. In the antidepressants study, for example, 617 of the 8,631 patients with complete data at baseline have at least two periods of enrollment; assuming that gaps of under 92 days do not represent actual discontinuities in coverage, Figure 4(b) summarizes the distribution of the first such gap across these patients. A second, related phenomenon is that when EHR data is extracted it is typically subject to administrative censoring. In the antidepressants study, this date was November 2009. To simplify the exposition, we restricted the study to patients with at least two years of potential follow-up (i.e., we included only patients with a new treatment episode prior to November 2007). An alternative would have been to include all patients who initiated a treatment episode prior to November 2009 and to use survival analysis methods to explore enrollment status at 24 months while accommodating censoring.

Regardless of enrollment status, a patient may choose to receive their health care at different facilities and institutions. At Group Health, while 70 percent of patients receive mental health care within an integrated group practice that uses the same EpicCare EHR, 30 percent receive mental health care from an external network of providers. For these patients, although billing information

Table 2. Sub-mechanisms Potentially Relevant to Whether or not a Data Element Is Recorded in the EHR

SUB-MECHANISM	CONTEXTUAL QUESTIONS
1. Enrollment status	Was the patient continuously enrolled in the health plan or system during the time frame of interest or, at least, at the time points of interest?
2. Multiple facilities and institutions	Did the patient potentially receive care at multiple facilities and institutions? If so, do they maintain comparable and compatible EHR systems, and are they linked?
3. Encounters with the health system	Did the patient initiate an encounter with the health systems that the EHR corresponds to? If the patient initiated an encounter, was it of a type that would reasonably be expected to generate an entry in the EHR?
4. Measurement	If a relevant encounter type was initiated, was the measurement of interest recorded in the EHR? Are there any clinical reasons or contraindications why a patient would not have had a measurement taken? Could clinical priorities during an encounter have had an impact on whether or not a measurement was taken and recorded?
5. Structural changes	Have standards of care changed over time in a way that impacts which measurements are recorded and how? Has the EHR system evolved over time, either in terms of structure or of coding policies and procedures? If so, could these changes have influenced which measurements could have been recorded and how?

Figure 4. Summary information Regarding Disenrollment and Censoring of Patient Follow-Up During the 24-Month Interval Post-treatment Initiation among the 8,631 patients in the Antidepressants and Weight Change Study Who Had an Observed Weight Measurement at Baseline



The right-hand panel shows estimates of the cumulative probability of being disenrolled and being either disenrolled or censored. The left-hand panel shows 617 patients with more than one distinct period of enrollment during follow-up, specifically the distribution of the length of the first gap in enrollment.

is readily available, clinical information (including weight) may not be routinely collected using the same standards of care. Beyond Group Health–like settings, EHR-based studies conducted in tertiary care hospital settings may have detailed clinical information relevant to the condition that led to the admission and hospital stay but may not have broader information.

The third sub-mechanism in Table 2 considers the timing and nature of the clinical encounter. Clearly, for a measurement to be recorded in the EHR at a particular time point, an encounter must have been initiated. As highlighted in Figure 1 there can be substantial variation in the timing and intensity of encounters across patients. There can also be substantial variation in the type of encounter, with patients interacting with their health care providers via a primary care or specialty care visit, an inpatient or outpatient visit, or an urgent care or routine care visit. Furthermore, patients increasingly have the option to interact with their providers virtually, via telephone encounters or secure messaging systems. Since measurements of interest, such as weight in the antidepressants study, may not be collected during all encounter types—either at all or routinely, it would be important to identify (1) the types of care options that patients have, and (2) which types are captured in the EHR.

The fourth sub-mechanism speaks to the actual measurement and recording of information. From

Figure 1 it is clear that measurements may not be taken during all encounters. Critical to this sub-mechanism is that whether or not a measurement is taken may be dictated by decisions made by the patient, health care provider, and health system. For example, a patient may decide not to be weighed, or a physician may decide there is insufficient time to weigh the patient or may decide not to record a measurement in the light of extenuating circumstances (e.g., blood glucose may not be measured if the patient is known not to have fasted).

Finally, structural features of the EHR and the broader health system may result in information being less likely to be recorded. Even if the EHR system routinely collects clinical information, changing practice standards or an evolving internal structure of the EHR may result in differential completeness of some data elements over time. One ubiquitous example of this is the International Classification of Disease (ICD) coding system developed by the World Health Organization (WHO). First published in 1946, the current revision, ICD-10, came into use in 1994. In 2017 the WHO is planning on releasing the eleventh revision, which will be based on an updated standardized structure for disease definitions.

Application to the Antidepressants Study

To illustrate the proposed framework we return to the antidepressants study. As outlined above, the initial task in applying the framework is to specify the scientific question of interest and the corresponding ideal study and data structure. Since the primary analyses for the antidepressants study are ongoing, we focus on a simple question of whether choice of antidepressant medication at treatment initiation is associated with weight change at 24 months. That is, we focus on a question based on the intent-to-treat principle, which ignores changes in treatment choice postinitiation. This question is, arguably, most relevant to the clinical decision at the time of treatment initiation since one cannot know whether and how a patient will change treatment in the future. With this in mind, key variables that would be collected in an ideal study design would include the following: initial treatment choice; baseline; 24-month weight; and potential confounders such as gender, age, and comorbid conditions.

Given this list of variables, the next step is to consider the extent of missing data and its nature. In principle, any and all variables with missing values should be considered in this way; here, for simplicity, we focus on missingness in the 24-month weight measurement. Furthermore, we focus on three specific sub-mechanisms: (1) whether the patient was actively enrolled in Group Health, (2) whether they initiated an encounter with the health system at 24 months, and (3) whether their weight was measured during the encounter and recorded in EHR. Figure 2(b) provides a flow-type diagram to help visualize these sub-mechanisms and their interaction with each other.

Sub-mechanism 1: Active Enrollment Status at 24 Months

Returning to Figure 1, 3 of the 12 patients can be seen to have disenrolled from Group Health prior to the 24-month mark. Analogous to a dropout in a typical research setting, if a patient disenrolls, the EHR cannot be expected to have a weight measurement recorded. In practice, there are many reasons why an individual might disenroll from their health plan including cost increases, changes in employment status or employer coverage options, reaching eligibility for Medicare, and dissatisfaction with their coverage or provider access.

Among the 8,631 patients with complete weight data at baseline, 2,061 (23.9 percent) disenrolled at some point during the first 24 months following treatment initiation (Figure 4). From the second set of columns in Table 1, in contrast to the results for the single mechanism, gender does not appear to be associated with enrollment status at 24 months. Age, however, is positively associated with enrollment status, with older patients again estimated to have higher odds although the strength of the association is much greater (OR 1.41 for a 10-year increase in age; 95 percent CI: (1.36, 1.47)). With respect to treatment choice, the results are generally consistent with those based on single missingness mechanism although the strongest associations, specifically for serotonin antagonist and reuptake inhibitor (SARIs) and tricyclics, are somewhat attenuated. Finally, patients with higher baseline weight are estimated to have somewhat higher odds of active enrollment at 24 months (OR 1.02 for a 20-lb. increase in weight; 95 percent CI: (1.00, 1.04)), although the association is not statistically significant despite the sample size being the same as in the single missingness mechanism model.

Sub-mechanism 2: Initiation of an Encounter at 24 Months

Returning to Figure 1, despite being actively enrolled, none of the last three patients in the third row had initiated a clinical encounter at or around the 24-month mark. Clearly, for a weight measurement to be recorded in the EHR, however, an encounter must have taken place. In practice, encounters are initiated either because standards of care within the health system dictate a schedule of patient-provider interactions or because the patient is seeking care for a new or ongoing medical problem.

Among the 6,570 patients actively enrolled at 24 months, 1,604 (24.4 percent) initiated at least one encounter in the 24-month ±7 days window; 2,485 patients (37.4 percent) initiated at least one encounter in the 24-month ±14 days window; and, 3,688 patients (56.1 percent) initiated at least one encounter in the 24-month ±30 days window. Focusing on the latter group, the third set of columns in Table 1 indicate that, in contrast to sub-mechanism 1, gender is strongly associated with initiation of an encounter: female patients are estimated to have 24 percent higher odds than males. Furthermore, while age is again significantly associated with initiation of an encounter, the magnitude of the association is substantially smaller than for sub-mechanism 1 (i.e., OR 1.10 compared to 1.41). As with sub-mechanism 1, treatment choice appears to be significantly associated with initiation of an encounter; the magnitudes of the associations for SARIs and tricyclics are stronger than they were for sub-mechanism 1, and buproprion appears to be marginally positively associated with an increased odds of initiating an encounter compared to fluoxetine (OR 1.14; 95 percent CI: (0.98, 1.34)).

Sub-mechanism 3: Measurement of Weight at 24 Months

Finally, even if a patient is enrolled at 24 months and initiates a clinical encounter, a weight measurement may nevertheless not have been recorded in the EHR. The first patients in the second and third rows of Figure 1, for example, were enrolled and had a clinical encounter at 24 months, yet neither had a weight measurement recorded. In practice, although standards of care at Group Health indicate that weight should be measured during all primary care visits, it may be that these patients refused to be weighed or that their health care providers decided not to weigh them because of the specific focus of the visit (e.g., an acute illness) and because of timing considerations.

Among all 3,688 patients who initiated at least one encounter in the 24-month ±30 days window, 2,408 (65.3 percent) have at least one weight measurement recorded in the EHR during the same window. From Table 1, we find that neither gender nor age is associated with a patient having at least one weight measurement in the EHR given that they are enrolled and have an encounter. Treatment choice is, overall, significantly associated with having at least one weight measurement, with patients treated with a tricylic having higher estimated odds compared to those treated with fluoxetine (OR 1.28; 95 percent CI (1.01, 1.61)) and patients treated with a serotonin-specific reuptake inhibitor (SSRI) estimated to have lower odds (OR 0.80; 95 percent CI: (0.66, 0.96)). Finally, in contrast to the moderate association for sub-mechanism 2, baseline weight is strongly associated with whether or not a weight measurement is recorded given that an encounter was initiated (OR 1.05 for a 20-lb. increase in weight; 95 percent CI: (1.03, 1.08)).

Discussion

As researchers make use of EHR data for CER, the unique challenges posed by the complexity and heterogeneity of the observed data are well recognized. Since standard methods (i.e., those developed outside the EHR-based setting) have been found to be inadequate, the recent literature has seen a number of important advances to address these challenges, including methods that facilitate the coding and classification of text-based notes,[37,38] methods for record linkage in the absence of unique patient identifiers,[39,40] and methods for the control of confounding bias.[11-21] Common throughout this recent literature is the general philosophy that one should make use of as much of the available information in the EHR as possible. This clearly has appeal in the sense that information is not thrown away and, presumably, statistical efficiency and power are maximized. As researchers grapple with selection bias, however, application of this philosophy has two important drawbacks, both of which are exemplified by Figure 1. First, because EHR systems are typically designed to support clinical and/or billing activities, not with any specific research agenda in mind, the standard notions of "complete" or "missing" data do not have well-defined meanings; these notions only have meaning with respect to some data structure that is (typically) pre-specified by the study design. Second, given the complexity and heterogeneity of EHR data, making use of all of the available information will likely require the development and fit of large, complex models—the components of which may be poorly identified. Consider, for example, the challenging task of accurately modeling the underlying weight trajectories of all 8,631 patients in the antidepressants study who have complete baseline weight values. To resolve these challenges, we have proposed a new general framework for addressing selection bias in EHR-based settings. Central to the framework are two key principles that explicitly address the drawbacks of the standard philosophy: (1) the analysis is grounded in some pre-specified ideal study, and (2) the data provenance, which is the process that gives rise to the available EHR data, is decomposed into a series of manageable components. This, we believe, represents a fundamental shift in how selection bias is addressed in EHR-based studies.

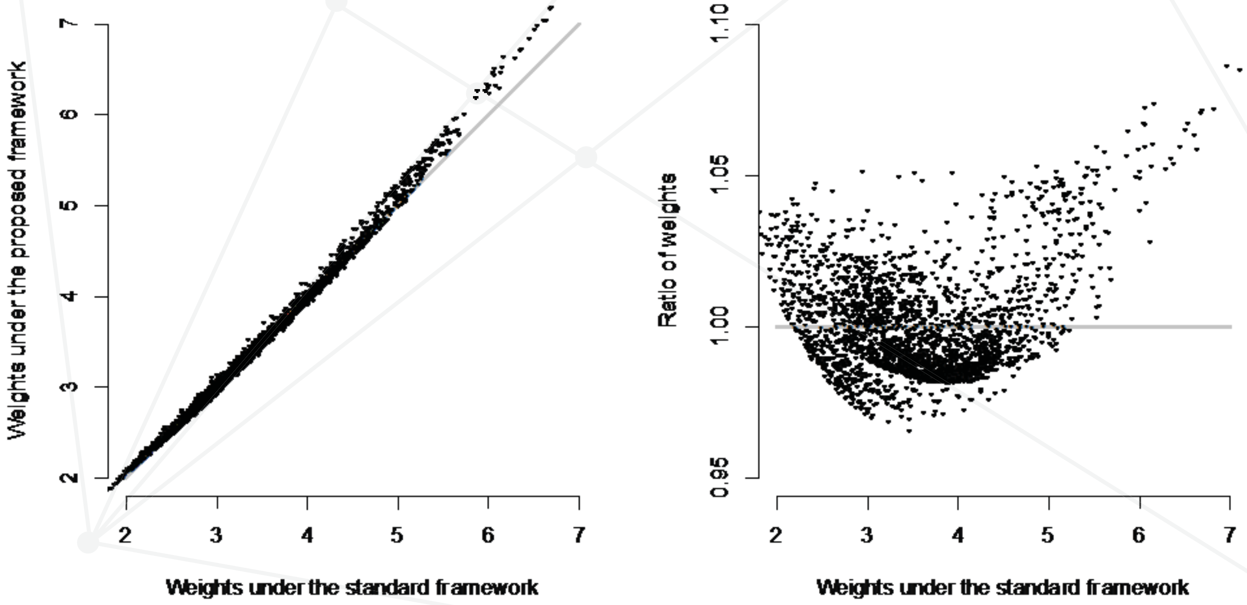
Practically, the proposed framework enjoys numerous important benefits. First, it provides focus in the elicitation process during which researchers consult with subject-matter experts on reasons and determinants of completeness. This may be particularly useful if the sub-mechanisms interact in such a way that if a particular event has not occurred then whether or not a subsequent event occurs is deterministic (e.g., a patient cannot have a weight measurement recorded if there was no encounter with the clinical system). Second, it provides flexibility in that the various sub-mechanisms may not be driven by the same set of covariates. In Table 1, for example, there is strong evidence that all four covariates are associated with sub-mechanism 2 but not necessarily with sub-mechanisms 1 and 3. Third, it provides flexibility in that any given covariate may have differential effects across sub-mechanisms. In Table 1, patients treated with an SSRI are estimated to have lower odds of being actively enrolled at 24 months compared to those treated with fluoxetine (OR 0.87) and having a weight measurement at 24 months (OR 0.79) but higher odds of initiating an encounter at 24 months (OR 1.08). Fourth, the decomposition of observance into a series of sub-mechanisms provides a clearer framing for consideration of critical assumptions. Specifically, after consulting with subject-matter experts, one may find that the MAR assumption is reasonable for some sub-mechanisms but not others. This, in turn, provides researchers with the ability to target sensitivity analyses specifically to those sub-mechanisms for which MNAR is suspected.[22,41]

Notwithstanding these benefits, implementing the proposed framework in any given CER study is not without challenges. Specifically, as mentioned, the framework is not prescriptive, in the sense that no single implementation will be adequate for all EHR-based studies. While Figure 2(b) is, arguably, a reasonable step forward from Figure 2(a), it could not be used as a general template. In this sense, the proposed framework requires researchers to make a series of potentially challenging decisions including the specification of the ideal study, the specification of potential sub-mechanisms relevant to the EHR system, and the specification of covariates that may influence the collection of sub-mechanisms. These tasks will typically be nontrivial, although the use of flow diagrams analogous to those in Figure 2 may be useful during the elicitation process as well as during modeling and sensitivity analyses. To further aid these tasks we are developing a suite of data-driven strategies, analogous to those recently developed for confounding bias,[13,20] that combine clinical knowledge with model selection methods[42-44] to identify relevant sub-mechanisms and their determinants.

Several features of the antidepressants application, as presented, are also worth noting. First, we chose to illustrate the framework in the context of a scientific question for which the appropriate analysis is an intent-to-treat analysis. Such questions clearly have clinical value, although they do not address important aspects of the relationship between treatment choice and 24-month weight change, including the potential impact of stopping treatment or treatment switching for which an appropriately adjusted as-treated analysis would be more appropriate. Furthermore, the intent-to-treat analysis does not consider the impact of intermediate events such as the resolution of the initial treatment episode. For each of these alternative scientific questions, however, the proposed framework could readily be applied. Second, to simply the development we restricted attention to patients with at least two years of potential follow-up (i.e., we only included patients with a new treatment episode prior to November 2007). An alternative would have been to include all patients who initiated a treatment episode prior to November 2009 and used survival analysis as a means to exploring enrollment status at 24 months while accommodating censoring. Finally, we considered only select baseline covariates for inclusion in the models in Table 1. In reality, it is likely that each of the sub-mechanisms will depend on patient characteristics that evolve over time. In principle, as with treatment changes over time, one could readily fold consideration of these covariates and the relevant timing of their measurement into the ideal study and sub-mechanism specification of the proposed framework.

Finally, while understanding mechanisms and consideration of assumptions is a necessary first task in any analysis involving missing data, ultimately the most important question is whether or not the additional effort required by the proposed framework makes a difference in the overall study results and conclusions. To answer this, statistical analysis methods must be aligned with the proposed framework. One approach to doing so could be to adapt existing methods based on inverse-probability weighting. In the context of Figure 2, rather than reweighting the main analyses by the inverse of P(S=1) (i.e., the fitted values from a model based on a single mechanism) one could reweight by the inverse of P(S1=1,S2=1,S3=1)=P(S1=1)xP(S2=1|S1=1) xP(S3=1|S1=1,S2=1), where each of the latter three components are taken as the fitted values from sub-mechanism-specific regression models (i.e., those in Table 1). Interestingly, for the antidepressants study these two sets of sampling weights do not differ substantially (see Figure 5). As such, although details are not shown, primary results for the main analyses investigating the association between choice of treatment and body weight change at 24 months do not differ substantively. Clearly this will not be the case for all EHR-based studies, although the fact that it is the case in the antidepressants study raises an important question regarding whether a penalty is paid for using an unnecessarily complex analysis. That is, is there a loss of efficiency when the observance mechanism is overspecified? When coupled with the potential for bias when the observance mechanism is underspecified, a potential bias-variance trade-off arises. Understanding this trade-off and providing practical guidance is an avenue of research that we are actively pursuing.

Figure 5. Fitted Sampling Weights Obtained from the Standard Single Missingness Mechanism Framework Compared to Those Obtained from Impementation of the Proposed Framework with Three Sub-mechanisms: Active Enrollment, Initiation of an Encounter, and Recording of a Body Weight Measurement



Acknowledgements

The work presented in this paper was funded, in part, by National Institutes of Health grants R01-MH083671 and R01-183854.

References

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*. Apr 2005;58(4):323-337.
2. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Annals of Internal Medicine*. Sep 1 2009;151(5):359-360.
3. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. Jan 1 2013;20(1):144-151.
4. Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. *J Comp Eff Res*. Nov 2013;2(6):529-532.
5. Institute of Medicine (U.S.). Committee on Comparative Effectiveness Research Prioritization. *Initial national priorities for comparative effectiveness research*. Washington, D.C.: National Academies Press; 2009.
6. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Medical Care*. Aug 2013;51(8 Suppl 3):S80-86.
7. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. Aug 2013;51(8 Suppl 3):S30-37.
8. Overhage J OL. Sensible use of observational clinical data. *Statistical Methods in Medical Research*. 2013;22(1):7-13.
9. Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part III. *Value in Health*. Nov-Dec 2009;12(8):1062-1073.
10. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. Sep 2004;15(5):615-625.
11. Pearl J. *Causality : models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press; 2000.
12. Morgan SL, Winship C. *Counterfactuals and causal inference*: Cambridge University Press; 2014.
13. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. Jul 2009;20(4):512-522.
14. Toh S, Garcia Rodriguez LA, Hernan MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiology and Drug Safety*. Aug 2011;20(8):849-857.
15. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*. Aug 1 2005;162(3):279-289.
16. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*. May 2006;15(5):291-303.
17. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration--a simulation study. *American Journal of Epidemiology*. May 15 2007;165(10):1110-1118.
18. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*. May 2006;17(3):268-275.
19. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety*. Jun 2010;19(6):537-554.
20. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*. Sep 2012;68(3):661-671.
21. Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F. Model feedback in Bayesian propensity score estimation. *Biometrics*. Mar 2013;69(1):263-273.
22. Daniels MJ, Hogan JW. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis:* CRC Press; 2008.
23. Little RJ, Rubin DB. *Statistical analysis with missing data*: John Wiley & Sons; 2014.
24. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *eGEMs*. 2013;1(3).
25. Deshmukh R, Franco K. Managing weight gain as a side effect of antidepressant therapy. *Cleveland Clinic Journal of Medicine*. 2003;70(7):614-623.
26. Schwartz T, Nihalani N, Jindal S, Virk S, Jones N. Psychiatric medication-induced obesity: a review. *Obesity Reviews*. 2004;5(2):115-121.
27. Serretti A, Mandelli L. Antidepressants and body weight: a comprehensive review and meta-analysis. *The Journal of Clinical Psychiatry*. 2010;71(10):1259-1272.
28. Berkowitz RI, Fabricatore AN. Obesity, psychiatric status, and psychiatric medications. *Psychiatric Clinics of North America*. 2011;34(4):747-764.
29. Blumenthal SR, Castro VM, Clements CC, et al. An electronic health records study of long-term weight gain following antidepressant use. *JAMA Psychiatry*. 2014;71(8):889-896.
30. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. June 1996;91(434):473-489.

31. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
32. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. Sep 1994;89(427):846-866.
33. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics*. Mar 2012;68(1):129-137.
34. Hernan MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. Nov 2008;19(6):766-779.
35. Danaei G, Garcia Rodriguez LA, Cantero OF, Logan R, Hernan MA. Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical methods in medical research*. Oct 19 2011.
36. Schneeweiss S. Understanding secondary databases: a commentary on “Sources of bias for health state characteristics in secondary databases”. *Journal of Clinical Epidemiology*. Jul 2007;60(7):648-650.
37. Denny JC. Mining electronic health records in the genomics era. *PLoS Computational Biology*. 2012;8(12):e1002823.
38. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*. 2010;17(6):646-651.
39. Larsen MD, Rubin DB. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*. 2001;96(453):32-41.
40. Li X, Shen C. Linkage of patient records from disparate sources. *Statistical Methods in Medical Research*. 2013;22(1):31-38.
41. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*. 1998;93(444):1321-1339.
42. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine*. Feb 28 1997;16(4):385-395.
43. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006;101(476):1418-1429.
44. Meier L, Van De Geer S, Buhlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*. 2008;70(1):53-71.