Published in final edited form as:
Epidemiology. 2022 January 01; 33(1): 105–113. doi:10.1097/EDE.0000000000001432.

On the nature of informative presence bias in analyses of electronic health records

Glen McGee[1], **Sebastien Haneuse**[2], **Brent A. Coull**[2], **Marc G. Weisskopf**[3], **Ran S. Rotem**[3,4]

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[3]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA

[4]Kahn-Sagol-Maccabi Research and Innovation Institute, Maccabi Healthcare Services, Tel Aviv, Israel

Abstract

Electronic health records (EHRs) offer unprecedented opportunities to answer epidemiologic questions. However, unlike in ordinary cohort studies or randomized trials, EHR data are collected somewhat idiosyncratically. In particular, patients who have more contact with the medical system have more opportunities to receive diagnoses, which are then recorded in their EHRs. The goal of this paper is to shed light on the nature and scope of this phenomenon, known as informative presence, which can bias estimates of associations. We show how this can be characterized as an instance of misclassification bias. As a consequence, we show that informative presence bias can occur in a broader range of settings than previously thought, and that simple adjustment for the number of visits as a confounder may not fully correct for bias. Additionally, where previous work has considered only under-diagnosis, investigators are often concerned about over-diagnosis; we show how this changes the settings in which bias manifests. We report on a comprehensive series of simulations to shed light on when to expect informative presence bias, how it can be mitigated in some cases, and cases in which new methods need to be developed.

Keywords

Differential misclassification; Collider bias; Informative visit process; Autism spectrum disorder

Introduction

Electronic health records (EHRs) are a rich data source for epidemiologic research. However, unlike clinical trials or prospective cohort studies, EHRs are not designed to answer a specific research question and are thus subject to somewhat idiosyncratic data

Correspondence: Glen McGee, Department of Statistics and Actuarial Science, University of Waterloo, Mathematics 3, 200 University Ave W, Waterloo, ON N2L 3G1, Canada; glen.mcgee@uwaterloo.ca.

Conflicts of interest: None declared

Software availability: R code to replicate simulations in this paper will be published upon acceptance at https://github.com/glenmcgee/IPbias.

collection procedures [1]. Consider the case of autism spectrum disorder (ASD). Recent epidemiologic work has used EHR data to explore the links between ASD and a wide range of comorbidities [2-11]. However, a patient can only receive a clinical diagnosis of some comorbidity when interacting with the medical system, so the data recorded in an EHR are, in some sense, a function of the number of medical encounters a patient has. This poses a challenge because patients with ASD have substantially more interactions with the medical system compared to individuals without the disorder [2,9,12,13], and therefore have more opportunities to receive diagnoses (correctly or erroneously). In the epidemiologic literature, this difference in contact has been termed informative presence, and can be an important source of bias if unaccounted for [14-17].

Towards resolving informative presence bias, the statistics literature has focused primarily on modeling time-varying outcomes in the presence of informative observation processes (or informative visit processes), which occur when the observation times (visits) are related to the outcomes themselves (for recent reviews, see Neuhaus et al., 2018; Pullenayegum & Lim, 2016; Sisk et al., 2020). Under an informative observation process, standard longitudinal approaches like generalized estimating equations and mixed effects models can lead to biased estimation, but extensions have been proposed for both classes of models. Marginal models can be fit via estimating equations with inverse intensity weights [21-24], among other approaches [25,26]. Mixed effects models have been extended by jointly modeling outcomes and observation times [27-34]. Nevertheless, investigators often analyze univariate outcomes recorded cross-sectionally or by some age cut-off. Examples include a study of the association between diabetes and depression and psychiatric disorders [16,35], and another on the association between ASD and gastrointestinal disorders in childhood [2]. Informative presence in this latter univariate setting is the focus in this paper.

Our goal is to elucidate the nature and scope of informative presence bias. Previous literature has compared it to Berkson's bias, arising when number of medical visits acts as a confounder or as a collider or both [16]. We instead characterize informative presence bias as a form of misclassification, and we show that it manifests more broadly than presently thought. Additionally, this framing clarifies the patterns of bias we observe and explains why this bias depends on whether the association is null or not. Finally, previous work has considered only visit-level sensitivity, such that patients who see the doctor frequently have more opportunities to correctly receive a diagnosis of an underlying condition. Investigators might also be concerned about visit-level specificity, whereby patients who see the doctor frequently may be more likely to be erroneously diagnosed with other comorbidities. We show how different assumptions concerning specificity and sensitivity change the dynamics of informative presence bias in relation to visit frequency.

Motivating Examples

Consider as an example Autism Spectrum Disorder (ASD), a neurodevelopmental disorder characterized by difficulties in social interaction and atypical behavor. The etiology of ASD is largely unclear [36-41], and recent epidemiologic work has reported numerous statistical associations with a wide range of comorbidities, including gastro-intestinal (GI) disorders, epilepsy, certain psychiatric conditions, other developmental delays, sleep disorders, and behavioral problems [2-5,7-11]. However, the underlying biologic mechanisms linking many of these conditions with ASD are often unclear, raising the possibility that some of these associations could be artifacts related to the different ways patients with and without ASD interact with the medical system. In recent analyses of data from a large Israeli health fund [13,42], we noticed children subsequently diagnosed with ASD had nearly twice as many interactions with the medical system in the years preceding diagnosis compared with typically developing children, and similar findings have been reported in other cohorts [2,9,12]. This could be problematic if, for example, mild conditions (like some GI disorders) that are prevalent in the broader population but normally underreported are more likely to be recorded in the EHRs of ASD patients—because more visits means a higher chance that they experience GI problems at the time of visit or in the recent past. Alternatively, ASD patients are screened more frequently for other neuropsychological conditions; if screening is done using diagnostic criteria with relatively low specificity, ASD patients may be more likely to erroneously receive an EHR diagnosis for these comorbidities. In either case, informative presence could lead to a spurious association between ASD and some comorbidity.

This example stems in part from our work on neurodevelopmental outcomes, but informative presence bias is not unique to this setting (nor to EHR data). It could occur in any study that compares groups with different health utilization patterns. For example, the fact that patients suffering from chronic pain have more frequent interactions with the medical system [43] may increase their odds of being diagnosed with fibromyalgia—for which pain is a defining feature [44]—or any other comorbidities whose probability of being recorded depends on visit frequency. The same is true for analyses of asthma [45], anxiety and depression [46], or any chronic condition liable to increase one's healthcare utilization. Informative presence could manifest even more broadly than in EHR based studies, such as in studies using claims data or registries whenever health outcomes are not routinely surveyed.

Conceptualizing Informative Presence Bias

Consider a hypothetical study of the association between some condition, $X \in \{0,1\}$, and outcome, $Y \in \{0,1\}$, which could be chronic or defined as ever/never occurring during the study period. In the first motivating example, X is ASD and Y is a comorbidity such as GI problems (e.g., constipation, diarrhea, gastroesophageal reflux disease) at some point during the study period. Unfortunately, in many EHR-based studies we do not actually observe the underlying X and Y. Instead, at each of N medical visits, a patient has an opportunity to receive a diagnosis of either condition. A phenotyping algorithm would then determine observed X^o and Y^o for analysis purposes; these typically correspond to whether the patient received a diagnosis at any visit [16]. As such, the observed data may be viewed as being subject to misclassification.

Figure 1A provides a graphical representation of the interplay between the true X and Y, the observed X^o and Y^o, and the number of encounters that a patient has in an EHR, N. Based on this conceptualization, one can consider various hypothetical data generating mechanisms. For example, if a person has underlying condition X or Y, they may have more encounters; this is represented by the presence of lines 1 and 2. As another example, the more contact a person has with the medical system, the more opportunities they have to be diagnosed with either condition; this is represented by the presence of lines 3 and 4. Finally we assume that, conditional on N, there is no direct link between X^o and Y^o; only underlying disease and number of encounters affect diagnosis.

Informative Presence as Differential Misclassification

When lines 3 and 4 are present (Figure 1Bi), N has been described as a confounder of the association between X^o and Y^o [16]. As such, when the framing of the analysis is to investigate the association between X^o and Y^o, bias may arise if one fails to adjust for N. In most instances, though, primary interest lies in the association between X and Y. Figure 1 depicts several scenarios in which N could not be described as a confounder. For example, suppose ASD were correctly recorded (i.e. observed and true status are the same) and that the subsequent increase in visits affects GI diagnoses (even if ASD does not affect underlying GI). That is, lines 1 and 4 are present but lines 2 and 3 are not (Figure 1Bii). Here, N does not confound the association between X and Y but, rather, relates to potential misclassification in the outcome.

To that end, recall that misclassification is non-differential if $P(Y^o \mid Y,X) = P(Y^o \mid Y)$, and consider the distribution of the observed outcome given underlying disease:

$$P(Y^o \mid Y, X) = \sum_N P(Y^o \mid Y, X, N)P(N \mid Y, X) = \sum_N P(Y^o \mid Y, N)P(N \mid Y, X),$$

where the second equality follows from the assumption that Y^o depends on X only through the underlying Y and through N. In the absence of line 1 in Figure 1A (i.e., $N \perp X \mid Y$), $P(N \mid Y,X) = P(N \mid Y)$ and hence $P(Y^o \mid Y,X) = \sum_N P(Y^o \mid Y,N)P(N \mid Y) = P(Y^o \mid Y)$. That is, misclassification in Y is non-differential with respect to X. Similarly, in the absence of line 4 in Figure 1A (i.e., $Y^o \perp N \mid Y$), then $P(Y^o \mid Y,X) = \sum_N P(Y^o \mid Y)P(N \mid Y,X) = P(Y^o \mid Y)$, and misclassification is again non-differential. However in the presence of both lines 1 and 4—that is, N depends on X (given Y), and misclassification in Y^o depends on N (given Y)—then, in general, $P(Y^o \mid Y,X) \neq P(Y^o \mid Y)$, indicating differential misclassification. A symmetric argument shows the presence of lines 2 and 3 (Figure 1Biii) induces differential misclassification in X with respect to Y; see supplementary material.

The case previously described as confounding can also be characterized in this way. For example, suppose lines 3 and 4 in Figure 1 are present but lines 1 and 2 are not (Figure 1Bi). Since we necessarily adjust for the misclassified exposure X^o, we write:

$$P(Y^o \mid Y, X, X^o) = \sum_N P(Y^o \mid Y, X, X^o, N)P(N \mid Y, X, X^o) = \sum_N P(Y^o \mid Y, N)P(N \mid Y, X, X^o).$$

Even though N and X are *marginally* independent, they are *conditionally* dependent given X^o, hence $P(N \mid Y,X,X^o) \neq P(N \mid Y,X^o)$ and hence $P(Y^o \mid Y,X,X^o) \neq P(Y^o \mid Y,X^o)$. Re-framing the problem as one of misclassification thus captures both the previously described instances of informative presence as well as new cases not previously considered.

Adjusting for Number of Visits

Treating informative presence as confounding suggested that one might correct for bias by adjusting for N. Viewing the problem instead as one of differential misclassification, there might still be good reason to do so. Returning to the $X \rightarrow N \rightarrow Y^o$ scenario (Figure 1Bii): adjusting for N implies that $P(Y^o | Y,X,N) = P(Y^o | Y,N)$. Intuitively, adjusting for N "breaks" the link between Y^o and X, so that misclassification is no longer differential. Importantly, this does not correct for misclassification outright. Instead it transforms a pernicious problem—differential misclassification—into a more benign one: non-differential misclassification.

In addition to not entirely solving the misclassification problem, adjusting for N is no panacea. Goldstein et al. (2016) rightly highlight that adjusting for N can cause collider bias (or M-bias) when both underlying conditions affect the number of visits (lines 1 & 2; $X,Y \rightarrow N$), whether or not there is misclassification. In the framework presented here, collider bias and differential misclassification are often competing sources of bias. In the simulations that follow we shed light on the tradeoff between the two.

A Practical Note on Visit-Level Sensitivity vs Specificity

Previous work has only considered low-sensitivity settings, where people who have few medical visits may have an underlying condition that goes undiagnosed in their EHR. However, investigators may be equally concerned with low specificity, where a person with frequent medical visits may erroneously receive some diagnosis. A peculiarity of the structure of misclassification here is that sensitivity and specificity interact differently with the number of visits. Take, for example, a simple misclassification model where each visit is independent with some fixed visit-level sensitivity and specificity based on a true underlying condition. For a fixed number of visits, N, the probability of a person erroneously having a diagnosis of $Y = 1$ recorded in their EHR (a false positive) is $P(Y^o = 1 | Y = 0) = 1 - (Spec_y)^N$, and the probability of a false negative is $P(Y^o = 0 | Y = 1) = (1 - Sens_y)^N$. In particular, the probability of a false positive diagnosis recorded in the EHR increases with N, whereas the probability of a false negative decreases with N. Practically speaking, this suggests that in settings where people have frequent contact with the medical system, low sensitivity may not ultimately result in much bias (since a higher N would increase the probability that X^o and Y^o accurately capture underlying X and Y); in settings where people have limited contact with the medical system, low specificity may not result in much bias (since lower N would decrease the probability that X^o and Y^o are erroneously recorded in the EHR in the absence of X and Y).

Simulation Study

Setup

To characterize the impact of informative presence, we simulated data in the different scenarios depicted in Figure 1 (e.g., scenario $X,Y \rightarrow N \rightarrow Y^o$ corresponding to lines 1, 2 & 4; i.e., underlying X,Y affect number of visits, which affects whether a diagnosis of Y is recorded, and X is well observed). For each of 2000 observations, we generated binary (true) X with a prevalence of 50%, then generated binary Y with a baseline prevalence of

25%—first under the null of no association, then with a log-odds ratio (OR) of 0.5 with X. (We set high prevalences in order to characterize the effects of both sensitivity and specificity; see eAppendix for simulations with rare conditions.) Based on the underlying X and Y, we generated a number of visits, N, over a fixed hypothetical timeframe. To do so, we drew $N-1$ from a negative binomial distribution whose mean was permitted to vary by X and Y, depending on the scenario. We investigated both a low visit-frequency setting, motivated by values observed in a large health fund in California for children with and without ASD [12], and a high visit-frequency setting, motivated by visit frequencies observed in a previous analysis of children with ASD in a large Israeli cohort and in an analysis of Medicaid enrolled adults [47]; see the Table for mean number of visits in each scenario. Then at each visit, a person may receive a diagnosis of one or both conditions. We considered two settings to contrast low visit-level sensitivity (sensitivity=0.75, specificity=1) and low visit-level specificity (sensitivity=1, specificity=0.99) for X and Y. Although a visit-level specificity of 0.99 does not appear low, this implies an overall specificity of 0.78 across 25 visits. Ultimately we set univariate summary measure X^o=1 or Y^o=1 if a patient received at least one diagnosis of each condition (corresponding to a typical phenotyping algorithm; Goldstein et al., 2016). See eAppendix 2 for more details on data generation.

Given the simulated sample, we fit an unadjusted logistic regression model using the observed X^o and Y^o, as well as one adjusted for N. We repeat the above for 2000 datasets for each scenario and report average log-OR estimates.

Results

Results regarding the estimated association between X and Y under the null of no association are shown in Figure 2. Focusing first on the infrequent visit setting with low sensitivity, we see that unadjusted estimates are unbiased in scenarios corresponding to non-differential misclassification ($N \rightarrow X^o$; $N \rightarrow Y^o$; $X \rightarrow N \rightarrow X^o$; $Y \rightarrow N \rightarrow Y^o$) as well as in scenario $X, Y \rightarrow N$, with no misclassification. In all other scenarios, unadjusted estimates were biased away from the null to various degrees. It is interesting to note that in the example of informative presence previously described in the literature—the confounding setting ($N \rightarrow X^o, Y^o$)—we observed less bias than in the other informative presence settings (when number of visits depends on X or Y). In the frequent visit setting, by contrast, unadjusted estimates were nearly unbiased across the board.

Adjusting for N fully corrected for bias under the null except in scenarios when N acted as a collider ($X, Y \rightarrow N$; $X, Y \rightarrow N \rightarrow X^o$; $X, Y \rightarrow N \rightarrow Y^o$; $X, Y \rightarrow N \rightarrow X^o, Y^o$). In these cases, collider bias was quite strong relative to the magnitude of informative presence bias. (Note that collider bias outweighs any effects of non-collapsibility here, as the adjusted estimates remain unbiased when N does not act as a collider.)

Results for the low-specificity setting differed in that the impact of visit frequency was reversed: informative presence bias was high when visits were frequent and very low when visits were infrequent. Note that the prevalences of X and Y were both high (0.5 and 0.25, respectively) in our simulations; since specificity applies to observations with $X = 0$ or $Y = 0$, the impact of low specificity would be stronger for rare diseases (see eAppendix).

Results under a non-null association between X and Y are shown in Figure 3. The general pattern of bias among the naïve estimates was similar to the above with one key difference. Here, even scenarios $N \rightarrow X^o$; $N \rightarrow Y^o$; $X \rightarrow N \rightarrow X^o$; $Y \rightarrow N \rightarrow Y^o$, which yielded unbiased estimates above, exhibited bias. This matches our intuition for non-differential misclassification, as estimates were biased towards the null. By the same token, adjusting for N no longer fully corrected for bias, as estimates remained biased towards the null.

Multiple Comorbidities

Thus far we have only considered a single X and Y. More realistic, however, is the case in which more medical visits means more opportunities to be diagnosed with any number of conditions (i.e., multiple Y). As such it is important to characterize the effects of informative presence in studies of a large number of conditions, such as occurs in phenome-wide association studies (PheWAS) in which associations with hundreds of outcomes are investigated [48]—or even in the literature as a whole. To explore the potential for informative presence bias to lead to spurious associations, we conducted a simulation based on the $X \rightarrow N \rightarrow Y^o$ scenario. Here, X had a prevalence of 10%; the mean number of visits was 20 if $X=1$ and 10 otherwise, to reflect the distribution of visits for people with and without the condition. We generated ten outcomes independently, each with prevalence of 30% and none associated with X. In each dataset of size 2000, we again fit models unadjusted and adjusted for N—separately for each comorbidity—and conducted a two-sided Wald test of no association between X and Y. We repeated this for specificities ranging from 0.990 to 0.998; sensitivity was 100%, and X was not misclassified.

Across 2000 datasets, we report the proportion of tests that were rejected, i.e. the observed type I error rate. We then report the family-wise error rate—that is, the proportion of datasets in which we reject at least one of the ten hypotheses—as well as the average number of false positives. To accommodate multiple comparisons, we report results with and without a Bonferroni correction.

Results are shown in Figure 4. As expected, the unadjusted estimator yielded inflated type I error. The effects of the informative presence bias propagated across multiple comparisons, leading to dramatically inflated family-wise error rate as well. This is not solely due to multiple comparisons: Bonferroni-corrected family-wise error rate was still dramatically inflated, whereas adjusting for N yielded nominal type I error and Bonferroni-corrected error rate. The average number of false positives was much higher for the unadjusted estimator; under 99% specificity, even adjusting for multiple comparisons, we would expect a false positive among ten potential comorbidities.

Discussion

In a somewhat different context, Goldstein et al (2019) first suggested that informative presence could be viewed as a misclassification problem. Here, we have formalized this notion and have argued that previous characterizations of informative presence as the result of confounding and/or M-bias are too narrow. By embedding the problem within a misclassification framework, we observe that bias can occur in many more settings—and may be more severe—than previously anticipated. Moreover, recasting the problem in this way explains why adjusting for number of visits as a covariate sometimes fails to fully correct for bias.

We highlight several takeaways. (1) Informative presence mirrors the intuition of differential misclassification: when misclassification in Y depends on X through N (or vice versa), unadjusted estimates are biased even under the null. (2) Adjusting for N reduces the problem to non-differential misclassification: while this corrects for bias under the null, it allows bias towards the null otherwise. (3) When both X and Y affect the number of visits, adjusting for N induces collider bias, which can outweigh informative presence bias when N is strongly correlated with disease. (4) Across multiple comorbidities, informative presence can amplify the expected number of false positives.

We also find that (5) if sensitivity is low, bias is minimal when visits are frequent; if specificity is low, bias is minimal when visits are infrequent. Choosing a phenotyping algorithm requires balancing the two in light of the frequency of visits (see eAppendix for expanded simulations). Although specificity based on ICD-9 or ICD-10 codes is typically high [49], we have shown here that even a specificity of 0.99 could induce bias in frequent visit settings. As such one might mitigate bias by adopting an algorithm that privileges specificity at the expense of sensitivity—e.g., requiring specialist-diagnosed conditions or second opinions.

Informative presence bias can sometimes be mitigated by adjusting for the number of visits—however, this does not fully correct for bias, nor does it always apply. One could investigate bias via a negative control analysis [50] or by expanding approaches for sensitivity analyses under irregular sampling and observation mechanisms [51]. Alternatively, there is a large literature on misclassification (see Carroll et al., 2006) which may provide a path forward for correcting bias. Restricting to diagnoses recorded during prescheduled visits, or using stricter ascertainment definitions (e.g. at least two diagnosis records or other means of diagnosis validation), could also be explored. In any case, strategies for correcting bias would likely require further information about the data generating mechanisms, either via validation data or assumptions.

Informative presence can be viewed as a collapsed version of an informative visit process in a repeated measures analysis. One might thus analyze EHRs by leveraging the statistical literature on longitudinal methods. However, there is no consensus solution to the longitudinal problem either, as some approaches can do more harm than good in some settings [19]. Mixed effects models provide unbiased estimates of some associations but exhibit bias in others [19]. Ultimately, longitudinally resolved data are not always available, and even when they are, longitudinal analyses pose even more complex modeling challenges. Simpler univariate analyses are common in practice, and by highlighting them here we hope to shed light on the consequences of a problem common in many epidemiologic analyses.

Although our simulations spanned a wide array of scenarios, real data may be considerably more complex than those depicted by Figure 1. For example, we have assumed that only underlying disease could impact number of medical visits, but actually receiving a diagnosis may affect number of visits even further. Moreover, visit frequency may not be the only way informative presence could manifest: informative presence bias is closely linked with surveillance bias [53,54] and detection bias [55,56], and even holding visit frequency constant, certain patients may be subject to more scrutiny at each visit, perhaps reflected by length of EHR entries. EHR data is also susceptible to selection bias [57], and the interplay between this and informative presence bias is not yet well-understood. What is more, we have used simplified misspecification models, whereas in reality not all medical visits are equivalent, and sensitivity/specificity may vary over time. These and related scenarios warrant further investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Sources of financial support:

NIH grant ES000002

References

1. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Stat Med. 2020;39(6):773–800. doi:10.1002/sim.8445 [PubMed: 31859414]

2. Alexeeff SE, Yau V, Qian Y, et al. Medical Conditions in the First Years of Life Associated with Future Diagnosis of ASD in Children. J Autism Dev Disord. 2017;47(7):2067–2079. doi:10.1007/s10803-017-3130-4 [PubMed: 28434058]

3. Cawthorpe D. Comprehensive Description of Comorbidity for Autism Spectrum Disorder in a General Population. Perm J. 2017;21:86–90. doi:10.7812/TPP/16-088

4. Croen LA, Zerbo O, Qian Y, et al. The health status of adults on the autism spectrum. Autism. 2015;19(7):814–823. doi:10.1177/1362361315577517 [PubMed: 25911091]

5. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. Pediatrics. 2014;133(1). doi:10.1542/peds.2013-0819

6. Doshi-Velez F, Avillach P, Palmer N, et al. Prevalence of Inflammatory Bowel Disease among Patients with Autism Spectrum Disorders. Inflamm Bowel Dis. 2015;21(10):2281–2288. doi:10.1097/MIB.0000000000000502 [PubMed: 26218138]

7. Kielinen M, Rantala H, Timonen E, et al. Associated medical disorders and disabilities in children with autistic disorder: A population-based study. Autism. 2004;8(1):49–60. doi:10.1177/1362361304040638 [PubMed: 15070547]

8. Kohane IS, McMurry A, Weber G, et al. The co-morbidity burden of children and young adults with autism spectrum disorders. PLoS One. 2012;7(4). doi:10.1371/journal.pone.0033224

9. Peacock G, Amendah D, Ouyang L, et al. Autism spectrum disorders and health care expenditures: The effects of co-occurring conditions. J Dev Behav Pediatr. 2012;33(1):2–8. doi:10.1097/DBP.0b013e31823969de [PubMed: 22157409]

10. Penzol MJ, Salazar De Pablo G, Llorente C, et al. Functional gastrointestinal disease in autism spectrum disorder: A retrospective descriptive study in a clinical sample. Front Psychiatry. 2019;10(4):1–6. doi:10.3389/fpsyt.2019.00179 [PubMed: 30723425]

11. Tye C, Runicles AK, Whitehouse AJO, et al. Characterizing the interplay between autism spectrum disorder and comorbid medical conditions: An integrative review. Front Psychiatry. 2018;9(1):1–21. doi:10.3389/fpsyt.2018.00751 [PubMed: 29410632]

12. Croen LA, Najjar D V., Ray GT, et al. A comparison of health care utilization and costs of children with and without autism spectrum disorders in a large group-model health plan. Pediatrics. 2006;118(4). doi:10.1542/peds.2006-0127

13. Rotem RS, Chodick G, Shalev V, et al. Maternal thyroid disorders and risk of autism spectrum disorder in progeny. Epidemiology. 2020;31(3):409–417. doi:10.1097/EDE.0000000000001174 [PubMed: 32251066]

14. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc. 2013;2013:1472–1477. [PubMed: 24551421]

15. Goldstein BA, Phelan M, Pagidipati NJ, et al. How and when informative visit processes can bias inference when using electronic health records data for clinical research. J Am Med Informatics Assoc. 2019;26(12):1609–1617. doi:10.1093/jamia/ocz148

16. Goldstein BA, Bhavsar NA, Phelan M, et al. Controlling for informed presence bias due to the number of health encounters in an electronic health record. Am J Epidemiol. 2016;184(11):847–855. doi:10.1093/aje/kww112 [PubMed: 27852603]

17. Phelan M, Bhavsar NA, Goldstein A. eGEMs Illustrating Informed Presence Bias in Electronic Health Records Data : How Patient Interactions. eGEMs (Generating Evid Methods to Improv patient outcomes). Published online 2017:1–14.

18. Pullenayegum EM, Lim LSH. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. Stat Methods Med Res. 2016;25(6):2992–3014. doi:10.1177/0962280214536537 [PubMed: 24855119]

19. Neuhaus JM, McCulloch CE, Boylan RD. Analysis of longitudinal data from outcome-dependent visit processes: Failure of proposed methods in realistic settings and potential improvements. Stat Med. 2018;37(29):4457–4471. doi:10.1002/sim.7932 [PubMed: 30112825]

20. Sisk R, Lin L, Sperrin M, et al. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. J Am Med Informatics Assoc. 2020;00(0):1–12. doi:10.1093/jamia/ocaa242

21. Bůžková P, Brown ER, John-Stewart GC. Longitudinal data analysis for generalized linear models under participant-driven informative follow-up: An application in maternal health epidemiology. Am J Epidemiol. 2010;171(2):189–197. doi:10.1093/aje/kwp353 [PubMed: 20007201]

22. Bůžková P, Lumley T. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. Can J Stat. 2007;35(4):485–500. doi:10.1002/cjs.5550350402

23. Lin H, Scharfstein DO, Rosenheck RA. Analysis of longitudinal data with irregular, outcome-dependent follow-up. J R Stat Soc Ser B Stat Methodol. 2004;66(3):791–813. doi:10.1111/j.1467-9868.2004.b5543.x

24. Pullenayegum EM, Feldman BM. Doubly robust estimation, optimally truncated inverse-intensity weighting and increment-based methods for the analysis of irregularly observed longitudinal data. Stat Med. 2013;32(6):1054–1072. doi:10.1002/sim.5640 [PubMed: 23047604]

25. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, et al. Parameter estimation in longitudinal studies with outcome-dependent follow-up. Biometrics. 2002;58(3):621–630. doi:10.1111/j.0006-341X.2002.00621.x [PubMed: 12229997]

26. Fitzmaurice GM, Lipsitz SR, Ibrahim JG, et al. Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. Biostatistics. 2006;7(3):469–485. doi:10.1093/biostatistics/kxj019 [PubMed: 16428260]

27. Liang Y, Lu W, Ying Z. Joint modeling and analysis of longitudinal data with informative observation times. Biometrics. 2009;65(2):377–384. doi:10.1111/j.1541-0420.2008.01104.x [PubMed: 18759841]

28. Liu L, Huang X, O'Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. Biometrics. 2008;64(3):950–958. doi:10.1111/j.1541-0420.2007.00954.x [PubMed: 18162110]

29. Gasparini A, Abrams KR, Barrett JK, et al. Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study. Stat Neerl. 2020;74(1):5–23. doi:10.1111/stan.12188 [PubMed: 31894164]

30. Ryu D, Sinha D, Mallick B, et al. Longitudinal studies with outcome-dependent follow-up: Models and bayesian regression. J Am Stat Assoc. 2007;102(479):952–961. doi:10.1198/016214507000000248 [PubMed: 18392118]

31. Sun J, Park DH, Sun L, et al. Semiparametric regression analysis of longitudinal data with informative observation times. J Am Stat Assoc. 2005;100(471):882–889. doi:10.1198/016214505000000060

32. Sun J, Sun L, Liu D. Regression analysis of longitudinal data in the presence of informative observation and censoring times. J Am Stat Assoc. 2007;102(480):1397–1406. doi:10.1198/016214507000000851

33. Sun L, Song X, Zhou J. Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times. J Stat Plan Inference. 2011;141(8):2902–2919. doi:10.1016/j.jspi.2011.03.013

34. Zhou J, Zhao X, Sun L. A new inference approach for joint models of longitudinal data with informative observation and censoring times. Stat Sin. 2013;23(2):571–593. doi:10.5705/ss.2011.285

35. Wu LT, Ghitza UE, Batch BC, et al. Substance use and mental diagnoses among adults with and without type 2 diabetes: Results from electronic health records data. Drug Alcohol Depend. 2015;156:162–169. doi:10.1016/j.drugalcdep.2015.09.003 [PubMed: 26392231]

36. Hansen SN, Schendel DE, Parner ET. Explaining the increase in the prevalence of autism spectrum disorders: The proportion attributable to changes in reporting practices. JAMA Pediatr. 2015;169(1):56–62. doi:10.1001/jamapediatrics.2014.1893 [PubMed: 25365033]

37. Maenner MJ, Rice CE, Arneson CL, et al. Potential impact of dsm-5 criteria on autism spectrum disorder prevalence estimates. JAMA Psychiatry. 2014;71(3):292–300. doi:10.1001/jamapsychiatry.2013.3893 [PubMed: 24452504]

38. Baio J, Wiggins L, Christensen DL, et al. Prevalence of autism spectrum disorder among children aged 8 Years - Autism and developmental disabilities monitoring network, 11 Sites, United States, 2014. MMWR Surveill Summ. 2018;67(6). doi:10.15585/mmwr.ss6706a1

39. Levy SE, Giarelli E, Lee L, et al. Psychiatric , and Medical Conditions Among Children in Multiple Populations of the United States. J Dev Behav Pediatr. 2010;31(4):267–275. [PubMed: 20431403]

40. Russell G, Rodgers L, Ukoumunne O, et al. Prevalence of parent-reported ASD and ADHD in the UK: findings from the Millennium Cohort Study. Published online 2016.

41. Green H, McGinnity Á, Meltzer H, et al. Mental Health of Children and Young People in Great Britain, 2004.; 2005.

42. Davidovitch M, Slobodin O, Weisskopf MG, et al. Age-Specific Time Trends in Incidence Rates of Autism Spectrum Disorder Following Adaptation of DSM-5 and Other ASD-Related Regulatory Changes in Israel. Autism Res. Published online 2020.

43. Mann EG, Johnson A, VanDenKerkhof EG. Frequency and characteristics of healthcare visits associated with chronic pain: results from a population-based Canadian study. Can J Anesth. 2016;63(4):411–441. doi:10.1007/s12630-015-0578-6 [PubMed: 26846618]

44. Goldenberg DL, Bradley LA, Arnold LM, et al. Understanding fibromyalgia and its related disorders. Prim Care Companion J Clin Psychiatry. 2008;10(2):133–144. doi:10.4088/pcc.v10n0208 [PubMed: 18458727]

45. Inoue H, Kozawa M, Milligan KL, et al. A retrospective cohort study evaluating healthcare resource utilization in patients with asthma in Japan. npj Prim Care Respir Med. 2019;29(1). doi:10.1038/s41533-019-0128-8

46. Knox SA, Britt H. The contribution of demographic and morbidity factors to self-reported visit frequency of patients: A cross-sectional study of general practice patients in Australia. BMC Fam Pract. 2004;5:1–7. doi:10.1186/1471-2296-5-17 [PubMed: 15053839]

47. Vohra R, Madhavan S, Sambamoorthi U. Comorbidity prevalence, healthcare utilization, and expenditures of Medicaid enrolled adults with autism spectrum disorders. Autism. 2017;21(8):995–1009. doi:10.1177/1362361316665222 [PubMed: 27875247]

48. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205–1210. doi:10.1093/bioinformatics/btq126 [PubMed: 20335276]

49. Quan H, Li B, Duncan Saunders L, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008;43(4):1424–1441. doi:10.1111/j.1475-6773.2007.00822.x [PubMed: 18756617]

50. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls: A tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383–388. doi:10.1097/EDE.0b013e3181d61eeb [PubMed: 20335814]

51. Beesley LJ, Fritsche LG, Mukherjee B. An analytic framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. Stat Med. 2020;39(14):1965–1979. doi:10.1002/sim.8524 [PubMed: 32198773]

52. Carroll RJ, Ruppert D, Stefanski LA, et al. Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition. Chapman and Hall/CRC; 2006.

53. Haut ER, Pronovost PJ. Surveillance bias in outcomes reporting. JAMA - J Am Med Assoc. 2011;305(23):2462–2463. doi:10.1001/jama.2011.822

54. Chiolero A, Santschi V, Paccaud F. Public health surveillance with electronic medical records: At risk of surveillance bias and overdiagnosis. Eur J Public Health. 2013;23(3):350–351. doi:10.1093/eurpub/ckt044 [PubMed: 23599219]

55. Sackett DL. Bias in analytic research. J Chronic Dis. 1979;32(1-2):51–63. doi:10.1016/0021-9681(79)90012-2 [PubMed: 447779]

56. Arfè A, Corrao G. Tutorial: Strategies addressing detection bias were reviewed and implemented for investigating the statins-diabetes association. J Clin Epidemiol. 2015;68(5):480–488. doi:10.1016/j.jclinepi.2014.12.001 [PubMed: 25554519]

57. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why? eGEMs (Generating Evid Methods to Improv patient outcomes). 2016;4(1):16. doi:10.13063/2327-9214.1203

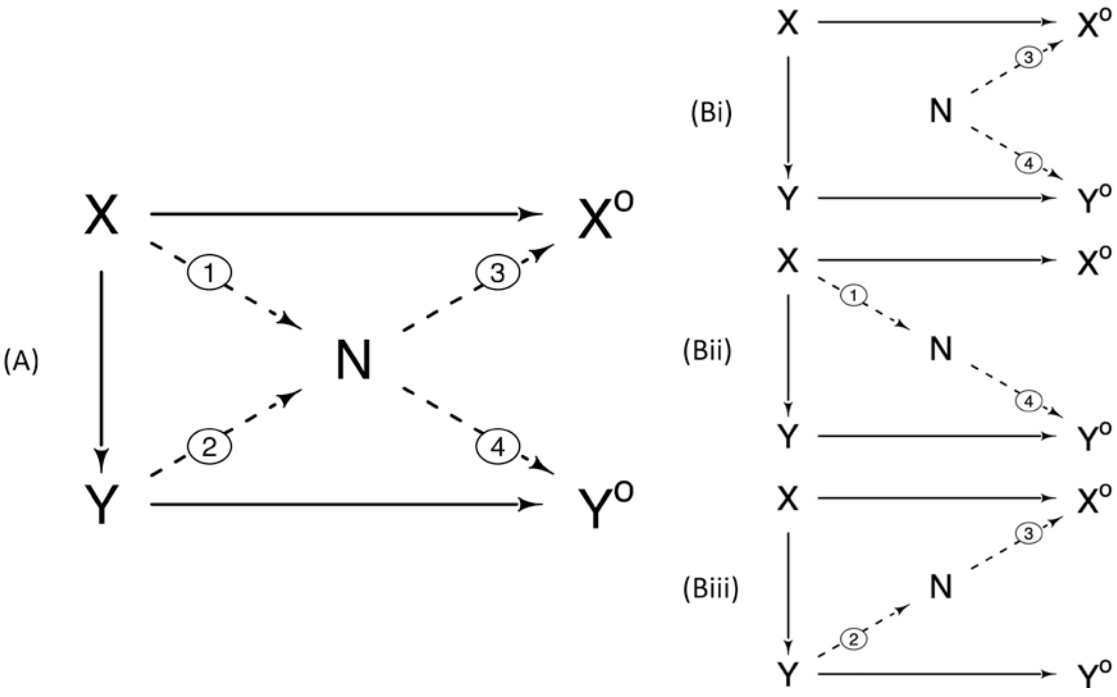

Figure 1.

(A) Potential data generating mechanisms. Dashed lines represent possible scenarios. (Bi)—(Biii) represent specific scenarios. In (Bi), N can be viewed as a confounder of the X^o~Y^(o)association as in Goldstein et al. (2016).

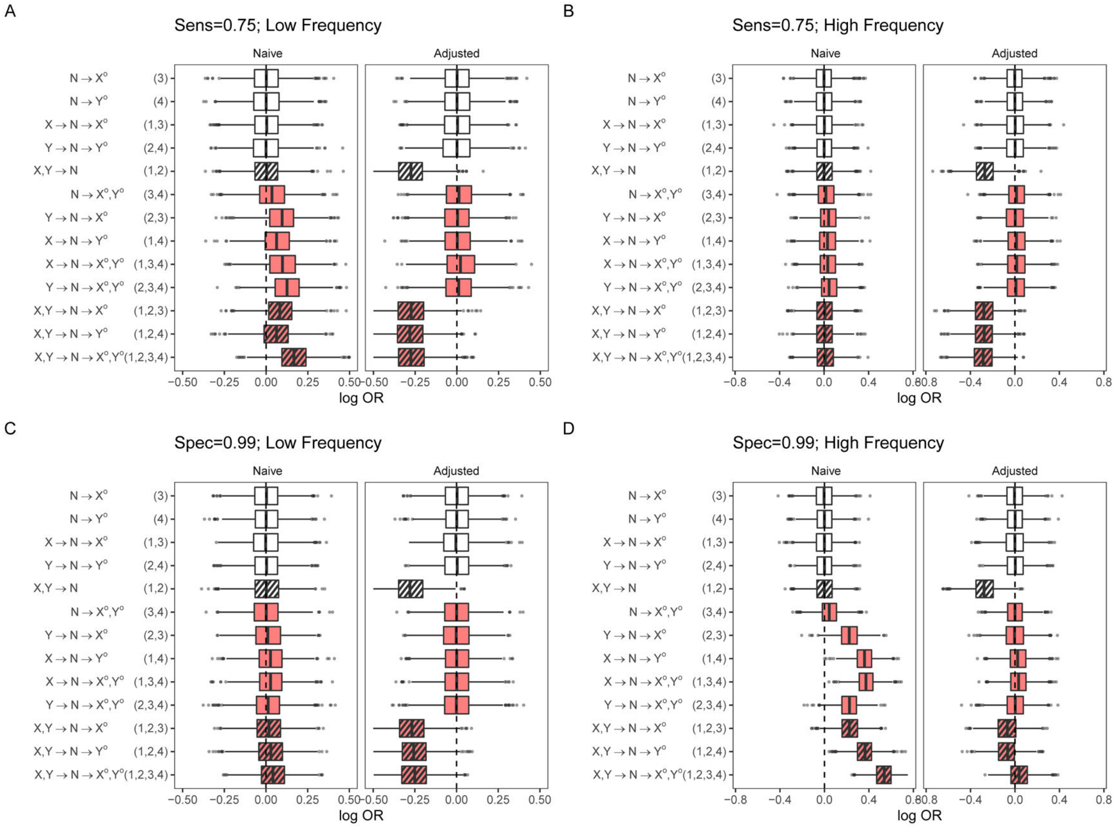

Figure 2.

Simulation results under the null (no association between X & Y) in four settings (low visit sensitivity vs low visit specificity; low vs. high visit frequency). Shaded bars correspond to differential misclassification settings; Stripes correspond to collider bias settings. “Adjusted” models adjust for N; Naïve models do not.

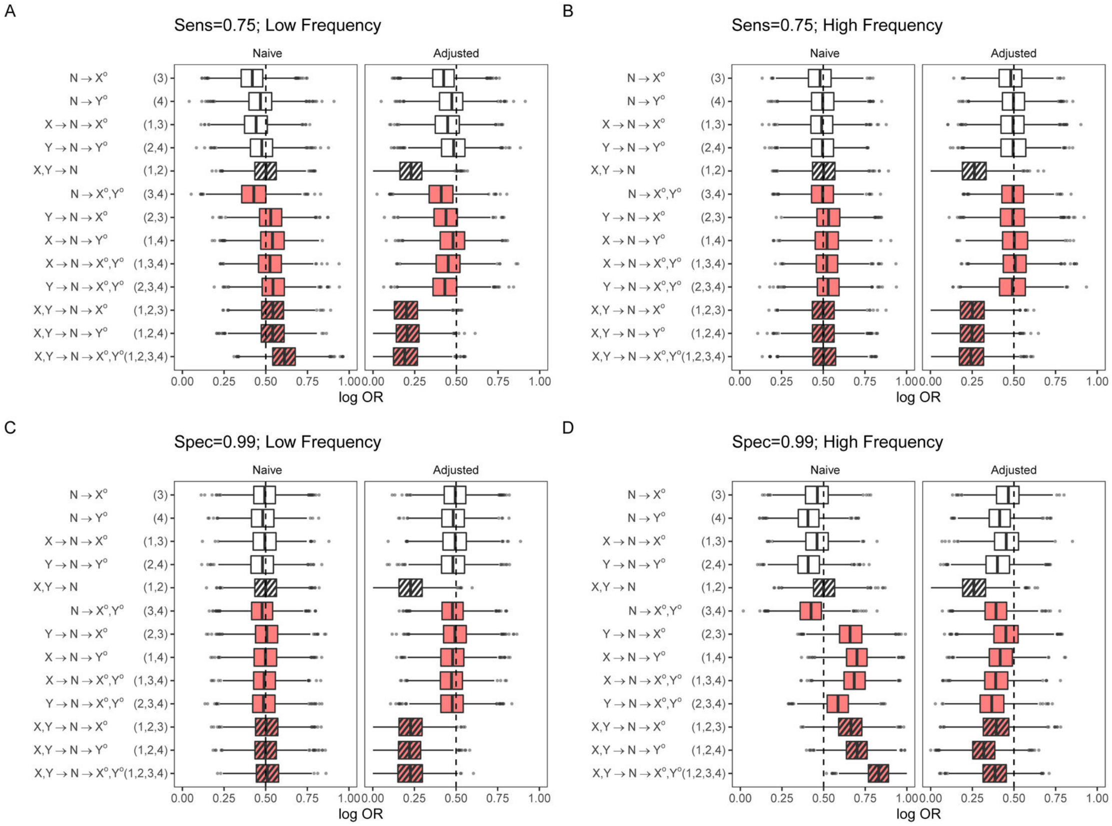

Figure 3.

Simulation results under a non-null association between X and Y (logOR=0.5) in four settings (low visit sensitivity vs low visit specificity; low vs. high visit frequency). Shaded bars correspond to differential misclassification settings; Stripes correspond to collider bias settings. “Adjusted” models adjust for N; Naïve models do not.

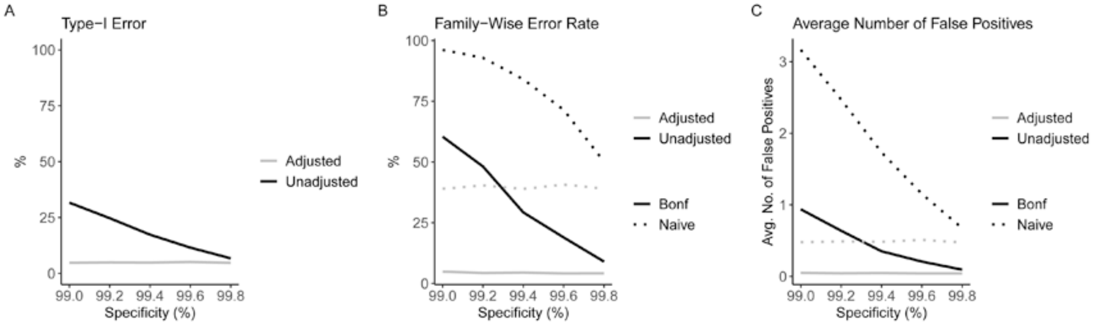

Figure 4.

Simulation results: observe (A) type-I error rate for a single condition, (B) family-wise error rate and (C) average number of false positives across 2000 datasets. Naïve refers to no correction for multiple hypothesis testing, Bonf refers to Bonferroni-corrected testing. Adjusted model adjusts for N; unadjusted does not.

Table:

Mean number of visits in study period in each simulation scenario. Scenarios correspond to data-generating mechanisms as depicted in Figure 1A: · → N indicates neither X nor Y affect N (neither line 1 or 2 in Figure 1A); X → N indicates only X affects N (Figure 1A line 1); Y → N indicates only Y affects N (Figure 1A line 2); X,Y → N indicates both X and Y affect N (Figure 1A lines 1 and 2).

Scenario (see Figure 1)	**Low Visit Frequency**					**High Visit Frequency**			
	X=0, Y=0	**X=0, Y=1**	**X=1, Y=0**	**X=1, Y=1**		**X=0, Y=0**	**X=0, Y=1**	**X=1, Y=0**	**X=1, Y=1**
$\cdot \rightarrow N$ (neither lines 1 or 2)	1.6	1.6	1.6	1.6		10	10	10	10
$X \rightarrow N$ (line 1)	1.6	1.6	2.3	2.3		10	10	25	25
$Y \rightarrow N$ (line 2)	1.6	2.3	1.6	2.3		10	25	10	25
$X, Y \rightarrow N$ (lines 1 and 2)	1.6	2.3	2.3	3.6		10	25	25	30