HHS Public Access

Author manuscript

Glob Epidemiol. Author manuscript; available in PMC 2022 July 15.

Published in final edited form as:

Glob Epidemiol. 2021 November ; 3: . doi:10.1016/j.gloepi.2021.100066.

Practical data considerations for the modern epidemiology student

Nguyen K. Tran[a], **Timothy L. Lash**[b], **Neal D. Goldstein**[a,*]

[a]Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, Philadelphia, PA, USA

[b]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA

Abstract

As an inherent part of epidemiologic research, practical decisions made during data collection and analysis have the potential to impact the measurement of disease occurrence as well as statistical and causal inference from the results. However, the computational skills needed to collect, manipulate, and evaluate data have not always been a focus of educational programs, and the increasing interest in "data science" suggest that data literacy has become paramount to ensure valid estimation. In this article, we first motivate such practical concerns for the modern epidemiology student, particularly as it relates to challenges in causal inference; second, we discuss how such concerns may be manifested in typical epidemiological analyses and identify the potential for bias; third, we present a case study that exemplifies the entire process; and finally, we draw attention to resources that can help epidemiology students connect the theoretical underpinning of the science to the practical considerations as described herein.

Keywords

Data science; Epidemiology; Biostatistics; Causal inference; Education and training

Introduction

We are taught that epidemiologic research often proceeds under a continuum [1]. A research question is conceived, a study is designed and implemented, the analysis is conducted, and interpretation offered. Many epidemiologists receive rigorous training in the theoretical and methodological underpinnings to answer research questions. For example, in observational etiologic research, we learn of six mechanisms under which an exposure, X, may be related to an outcome, Y: (1) chance, (2) uncontrolled confounding, (3) selection bias, (4)

*Corresponding author at: Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, 3215 Market St., Philadelphia, PA 19104, USA. ng338@drexel.edu (N.D. Goldstein).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

information bias, (5) reverse causality, and (6) true causality [2]. And we learn study designs and analysis strategies to limit mechanisms 1 to 5, so that mechanism 6 sends the clearest signal. Epidemiologists who engage in descriptive, experimental, and quasi-experimental work [3] have similar issues to contend with, as do applied epidemiologists. In short, our training emphasizes a rigorous science.

For the modern epidemiology student, regardless of sub-discipline or application, inherent in epidemiology is the collection and analysis of electronic data. As such, data literacy is crucial to the success of this field, yet the computational skills necessary to collect, clean, and analyse data are often taught separate from traditional training in epidemiology. A 2019 review of graduate curricula among 20 master's level public health programs in the U.S. noted that training in "data science", which includes methods of data management and manipulation, was rarely required as a standalone course in contemporary epidemiology programs, and there was a clear delineation between coursework in epidemiology and biostatistics [4]. While this review was unable to evaluate whether such data literacy skills were integrated within existing epidemiology or biostatistics courses, the growing interest in and use of "big data" sources necessitate greater emphasis on pragmatic considerations when conducting epidemiologic studies. Even a well-conceived, elegant theoretical model studying a pressing public health problem could be derailed by a single poorly conceived, haphazardly measured variable (especially true if this variable were X or Y). We contend that these more practical data decisions are as important for the student to learn as the theoretical and methodological concerns of the practicing epidemiologist, and use this article to outline these more pragmatic considerations and how they interact with the underpinnings of modern-day epidemiology [2,5–8]. Our intention is to demonstrate how decisions about data impact causal inference in observational research, but our observations are germane to all sub-disciplines and applications of epidemiology including descriptive studies, experimental, quasi-experimental studies, and fieldwork. Our audience is trainees in epidemiology - clinical or nonclinical - or anyone who will be undertaking an epidemiological inquiry. To illustrate our points, we provide six prototypical examples, their potential impact on the interpretation of results, and possible solutions, summarized in Table 1, and follow with a use case further demonstrating these issues in a real-world study.

We note at the outset several defining features of our commentary. First, we deliberately use the terms *data management* and *analysis* broadly. By data management we refer to the process of collecting data, including strategies on accessing, harmonizing, cleaning, storing, and preparing data for analysis. By analysis we refer to the process of producing the $P(Y|X)$ estimation in statistical software; in other words, quantifying the exposure to outcome relation. We refer to an "analytic dataset" as the basis for all statistical modelling and the product of data management. Second, although we treat each of the strategies within the research continuum distinctly for didactic presentation, we recognize that they are not mutually exclusive. For example, misspecification of a variable's definition during the study design can impact decisions about data management and analysis, inducing spurious associations between X and Y. Third, we do not offer new definitions, frameworks, or theories of epidemiology, and the generalizations made herein may not be true among all epidemiology training programs. Fourth and specific to the use case, it is not our intention

to critique the authors or findings of the cited studies, but rather to demonstrate the potential for invalid inference based on assumptions made during data management and analysis.

Data management

Broadly speaking, study data in an analytic dataset can fall under two categorizations. Data that describe the "rows" in a dataset, i.e., the *observations*, and data that describe the "columns" of a dataset, i.e., the *variables*. Note that the scenarios described below may occur without the awareness of the researcher. Practical issues may not represent fatal flaws in the data whereby software would flag an error drawing the researcher's attention to them. Rather these seemingly innocuous issues can slip through undetected and wreak havoc in the final analysis.

The theoretical underpinnings of missing data and its influence on causal inference have been well described [9,10]. Missing data may occur for both the observations and variables, where the reason may (or may not) be related to the observed, apparent data [11]. Analysis decisions about missing data may induce selection or information bias, such as when the analyst conducts a complete case analysis where observations (or variables) were discarded from the analytic sample because data were incomplete. Such analytic procedure assumes non-informative missingness, a common assumption of statistics models. As a result, $P(Y|X)$ in the analytic sample loses internal validity and is not reflective of the source population [12]. Simple descriptive statistics [13] and causal diagrams [14] may reveal the patterns of missingness and determine the most appropriate remediation [15,16].

As opposed to missing data that represent a lack of information, duplicate observations represent excessive information, and may have resulted from an incorrect data merge (append) operation, and thus represent invalid rows. Duplicate observations are often a by-product of many reporting tools integrated within electronic health records and registries, and as a result, deduplication algorithms are commonplace. However, for data privacy and security purposes, data linkage may be performed by a third party, which makes it difficult to determine the quality of linkage. Depending upon the proportion of invalid duplicate observations and the strength of $P(Y|X)$, the causal estimand may be biased from the influence of the extraneous data. One way of thinking about this systematic error is through selection bias, in that the probability of inclusion in the analytic sample for any one person is conditional on the variables used to merge the two datasets together [17]. Thus, in the case of duplicate observations, there are unequal selection probabilities for those individuals. One recommended approach to evaluate the degree of duplication is to calculate the percentage of observed versus potential number of record linkage [18].

The use of inconsistent variable definitions, incorrect constructs, and other problems that arise during data management and cleaning may also impact causal inference. For instance, when multiple datasets are linked together, as is the case when separate instruments were used and the data were recorded in separate files, the operationalization of the variables may have differed. As a simple example, this could be a coding of 0 = male and 1 = female in the first file, and 0 = female and 1 = male in the second file. Failure to recognize this inconsistency can induce an information bias in the final analysis without the researcher ever

being aware [17]. This could also affect continuous variables, if, for example, one dataset defined weight in pounds and the other in kilograms. Exploratory data analysis should reveal a potential problem, but this may be subtle if the scales substantially overlap. Furthermore, several recommended practices have been proposed for evaluating the impact of data linkage error such as comparing the linked data to a training dataset or gold standard, comparing linked and unlinked data, and sensitivity analysis to evaluate how robust results are to different linkage procedures [19]. See Doidge and Harron for a summary of strengths and limitations of these methods [17].

Analysis

Analysis of epidemiological data can include straightforward univariable descriptive statistics to complex simulation or regression-based approaches. Regardless of the complexity of the analysis, again there are practical considerations that can influence causal inference. Practical analytic decisions that have the potential to induce P($Y|X$), or lack thereof, include implications from the study design, model specification and assumptions, and variable selection.

Aside from the well-known challenges of estimating causal relations from epidemiologic studies [20], there are practical considerations in the study design that can impact inference. For example, a study analysing perinatal outcomes may have multiple rows of data representing multiple gestations. Unlike the earlier case where the duplicate observations were an artifact of data management decisions, here the repeated nature of the data is intentional and inherent to the study design. In this case, failure to correctly account for the correlated observations, especially if there was a relatively large proportion of multiple births as could be expected in a study of fertility treatments [21], may artificially inflate the statistical model error terms and thereby obscure an otherwise apparent association in the data [22–24].

The choice of which statistical model to employ brings about a host of practical considerations as all statistical procedures carry assumptions. One must first consider the functional form of their model (i.e., model specification) [25]. When given a continuous outcome, it is perhaps plausible to assume that the average risk of Y varies linearly as a function of X. However, such assumptions without any descriptive assessment of the functional relation between X and Y may lead to a poorly fitted model and erroneous inferences. It is possible that other model forms such as quadratic, exponential, or spline may better characterize the functional relation between X and Y, and failure to capture this functional relation may bias one’s estimates. In addition, common regression methods such as ordinary least squares regression, for example, includes assumptions for independence of observations, linearity, homoskedasticity, and non-informative missingness. From a practical lens, violation of these assumptions can result in incorrect point estimates or error terms. This violation may in turn lead to a biased or chance association [26]. In such instances, the use of model diagnostic procedures is vital to detect both systematic and isolated departures from the data [27]. Outliers are an obvious example, and many diagnostic techniques such as residual plots, Cook’s distance, DFBETAS, and goodness-of-fit tests have been developed to evaluate the robustness of these model assumptions [28]. Furthermore, failure to evaluate

whether missing data are informative will default to the model's assumption and treatment of missing data, which is typically a complete cases analysis. Non-informative missingness often does not hold, and in such cases, methods of multiple imputation by chained equations [16] and weighting approaches [29] have been developed to address this concern.

Relatedly, in most observational studies in epidemiology, one must also consider the variables to be included in a model, especially in research seeking to estimate the P($Y|X$). This is because with observational data, we have no expectation of exchangeability, thus, attempts to control for a sufficient set of confounders to produce valid estimates of P($Y|X$) are necessary. Although causal diagrams are helpful to depict the relational structure [30,31], this requires substantive background knowledge of the data-generating mechanism to appropriately identify a set of covariates, often using the backdoor path criterion [30], that are sufficient to control for confounding. Once these variables are identified, many techniques are currently available such as matching, stratification, multivariable adjustment [2], and propensity score methods [32] to control for confounding. However, knowledge of the entire causal structure is never available, which is why research needs to reflect this uncertainty given that residual confounding and undiagnosed measurement error can induce spurious associations [33]. In such instances, bias analysis techniques have been developed to quantify the impact of potential uncontrolled confounding and measurement error [34–40]. For the epidemiology student, practical approaches that are relatively simple to implement and impose fewer assumptions such as estimation of the E-value (the minimum strength of association of an unmeasured confounder to fully explain away an exposure-outcome relation as measured on the risk ratio scale) [41] or incorporation of negative controls (replication of the proposed experiment under conditions that are expected to produced null results) [42] will strengthen the student's ability to assess the quality of evidence from observational data.

In addition, various data-driven strategies have been developed for variable selection, including the significance criteria, information criteria, penalized likelihood (e.g., LASSO), change-in-estimate criterion, and variable selection algorithms [43,44]. Overreliance of data-driven methods for variable selection, however, can lead to the inclusion of too few or too many confounders, resulting in residual confounding. For example, given the rise in machine learning approaches to variable selection in high-dimensional datasets [4], there is the possibility of including inappropriate covariates in the statistical model. Except in the case of mediation analysis, one would not want to adjust for a mediator, yet because it is correlated with the exposure and outcome, a naive algorithmic approach would nevertheless include this variable, introducing the risk of biased estimation [45]. This further highlights the importance of accounting for the timing of covariates when making decisions on variable selection and the use of methods such as inverse probability weighting of exposure adjust for time-varying confounding [46]. The tension between a sound theoretical approach and a pragmatic data-driven approach is demonstrated in automated variable selection algorithms, which have largely been discouraged in epidemiological circles [26]. Thus, decisions to include or exclude variables in a model should be supported by background knowledge about the strength of evidence for their association with the exposure and outcome. In instances where variable selection algorithms are employed, one must consider

the uncertainty resulting from the selection process and its impact on inference through sensitivity analysis. For a more in-depth discussion of these strategies, see Heinze et al. [43]

Case study

The following scenario is adopted from Goldstein et al. [47] in which the authors undertook a replication study of the association between a certain type of physical activity and all-cause mortality using exposure and covariate data from the National Health and Nutrition Examination Survey (NHANES) with mortality outcome data linked from the National Center for Health Statistics [48]. The motivation was to test the feasibility of reproducing a study's findings based solely on the methods disclosed in the manuscript. As such, many implicit assumptions about the data were necessary during data management and analysis, and although these likely would not be disclosed in a typical original research article, herein we detail each consideration from Table 1 and how it may have affected the results.

1. *Missing data*. Goldstein et al. noted how decisions about the treatment of missing data in NHANES could cut the analytic sample in half.[49] Specifically, when operationalizing a single, latent variable based on the results of multiple survey questions, if the answers to one of the questions is missing, a data decision is needed: should the entire latent variable be set to missing and respondents without the latent variable excluded from analyses, or should the question with missing data simply be discarded from the construct of the latent variable? The authors therefore needed to balance the potential for selection bias, if the latent variable was omitted from a large number of respondents, versus an information bias, if only some aspect of the latent variable was ignored. On the other hand, a more prudent approach in the original work may have been to impute the missing data using one of the techniques discussed earlier.

2. *Duplicate observations*. In most cases, each row in an NHANES data file corresponds to a unique participant. However, this is not always true. For example, the repeated measure of physical activity in this study was based on accelerometer data that were captured on a minute-by-minute basis [48]. Thus, each participant who wore an accelerometer had a one-to-many relationship to these data and failure to correctly perform a merge operation to create the appropriate person-level data may result in an extreme number of duplicate observations. This could induce several types of bias from selection to overly precise errors. As such, the authors could benefit from calculating the percentage of observed to potential record linkage to better under the degree of false or missed linkage.

3. *Inconsistent variable definition*. Data linkage is commonplace when working with NHANES data. In fact, for the NHANES 2003–2004 and 2005–2006 survey cycles, the authors noted over 300 unique raw data files available [49]; this was in addition to the need to link to external mortality outcome data. As the practice of NHANES is to separate the various domains and instruments into separate raw data files, this necessitates data linkage. Not only does this present a problem for duplicate or missing data, but this also presents a problem

for inconsistent variable definitions. For example, when using NHANES data one may consider measures of self-reported hepatitis C diagnosis from the questionnaire data [50] and laboratory-based measure of hepatitis C antibody and viral load [51]. The treatment of these two variables as interchangeable would be inappropriate as the questionnaire-based data are self-reported and subject to greater misclassification. Therefore, combining these two measures may induce an information bias. It is incumbent upon the researcher to recognize the conceptual differences of similar data collected through different methods, in this case, a self-reported versus laboratory-based diagnosis of hepatitis C infection.

4. *Study design*. The original study's research aim was to evaluate the association between a certain type of physical activity and all-cause mortality. Participants were sampled from NHANES, which employed a complex survey design to ensure representativeness of the U.S. civilian noninstitutionalized resident population [52]. Failure to consider this study design may affect estimates of variance, and consequently biased test statistics and confidence intervals. As the data used in this example were from two survey cycles, namely 2003–2004 and 2005–2006, NHANES guidance documents stipulated several considerations before aggregating data [52]. Specifically, the authors needed to ensure the proper weighting variable was used before combining the datasets, as there are multiple survey weights given.

5. *Model assumptions*. The authors applied a Cox proportional hazard model to estimate how physical activity was associated with all-cause mortality in NHANES. This type of statistical model carries several assumptions including those common to all regression models, such as testing for the presence of influential observations, non-linearity, and non-informative missingness, as well as Cox-specific assumptions, namely proportional hazards [53]. While the authors of the original article stated that "the assumption of proportional hazards was tested and held true for our [physical activity] exposure," the details are not provided nor are the other assumptions described, particularly procedures for handling missing data [48]. This is not unusual; statistical modelling assumptions are rarely discussed in original research articles [54]. Sharing of data and analytic code can facilitate reproducibility when adequate details cannot be provided in the methods due to word limits [55].

6. *Variable selection*. Goldstein et al. noted eight separate questionnaire items in NHANES pertaining to alcohol consumption [47]. In order to create a single measure of consumption one may simplistically check for a "positive" response to any of these eight items; however, this risks an information bias as these individual items may represent unique constructs and not be internally reliable. Relatedly, inclusion of this combined measure may result in residual confounding as it may be a poor variable to control for the underlying differences in alcohol consumption between respondents of lower physical activity levels to those of higher physical activity levels. The use of quantitative bias analysis,

estimation of the *E*-value, or incorporation of negative controls may help the investigator evaluate the extent of residual confounding and measurement error.

Discussion

In the modern era of epidemiology, much thought and consideration of complex topics has yielded a rich theoretical and methodological foundation for researchers [20,56]. Our comments underscore their connection to practical considerations as the basis for valid epidemiology. In short, our work depends upon sound data, and we should not necessarily take for granted that current epidemiology training programs impart the knowledge and skills needed to manipulate complex study data prior to—and in some cases—post analysis [4]. Thus, we emphasize the importance of data literacy for students in epidemiology training programs. In fact, modern epidemiology may be more reflective of a computer and data scientist's skillset than a physician's mastery of medicine, contrary to the origin of our field.

The six practical considerations in Table 1 are not intended to be an exhaustive list. There are practical issues encountered at all stages of epidemiology: from asking an addressable public health question, to collecting data, to disseminating findings in an appropriate matter. The idealized research continuum many are taught may also conflict with real-world epidemiology, such as emergency public health responses during an outbreak, which may invoke some of the practical considerations in Table 1. We have focused our comments on the data aspect of epidemiologic research as opposed to the other, albeit equally important, facets such as the description of disease frequency that may provide salient information for understanding health disparities and risk factors. Additionally, epidemiologic data will always be imperfect, but not every data issue will lead to invalid inference. For example, in a study of thousands of individuals, a single duplicate record will have a negligible impact on the standard errors. In the case where selection or information bias is expected to substantially diminish the validity of results, analyses should reflect this uncertainty and investigators need to consider methods in bias analysis to compute bias-adjusted estimates to address systematic errors [2,36,37,57]. Through the case study, we illustrated how the analyst must think about such data concerns systematically, which, as a side benefit, can aid in the reproducibility of study findings [47]. Transparency in data and computing codes is one mechanism whereby others can vet the more practical issues of analysing data, for example, through a peer review process specific to research materials [49].

In summary, epidemiology is built upon sound theoretical reasoning, appropriate methodology, and valid data. The first two are well known; the third should not be taken for granted.

Funding:

This work was supported in part by the National Library of Medicine [R01LM013049, PI T. L. Lash] and by the National Institute of Allergy and Infectious Diseases [K01AI143356, PI N. D. Goldstein] of the US National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Petersen ML, van der Laan MJ. Causal models and learning from data. Epidemiology. 2014;25(3):418–26. 10.1097/EDE.0000000000000078. [PubMed: 24713881]

[2]. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Lippincott Williams & Wilkins; 2008.

[3]. KJ F, LT F, KRB J. Threats to the internal validity of experimental and quasi-experimental research in healthcare. J Health Care Chaplain 2018;24(3):107–30. 10.1080/08854726.2017.1421019. [PubMed: 29364793]

[4]. Goldstein ND, LeVasseur M, McClure LA. On the convergence of epidemiology, biostatistics, and data science. Harv Data Sci Rev April 30, 2020. 10.1162/99608f92.9f0215e6. Published online.

[5]. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int. J. Epidemiol 1986; 15(3):413–9. 10.1093/ije/15.3.413. [PubMed: 3771081]

[6]. Maldonado G, Sander G. Estimating causal effects. Int. J. Epidemiol 2002;31(2): 431–2. 10.1093/ije/31.2.422. [PubMed: 11980809]

[7]. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am. J. Public Health 2005;95(S1):S144–50. 10.2105/AJPH.2004.059204. [PubMed: 16030331]

[8]. Hernán MA. A definition of causal effect for epidemiological research. J. Epidemiol. Community Health 2004;58(4):265–71. 10.1136/jech.2002.006361. [PubMed: 15026432]

[9]. Howe CJ, Cain LE, Hogan JW. Are all biases missing data problems? Curr Epidemiol Rep 2015;2(3):162–71. 10.1007/s40471-015-0050-8. [PubMed: 26576336]

[10]. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. Int. J. Epidemiol 2015;44(4):1452–9. 10.1093/ije/dyu272. [PubMed: 25921223]

[11]. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581. 10.2307/2335739.

[12]. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection Bias. Epidemiology. 2004;15(5):615–25. 10.1097/01.ede.0000135174.63482.43. [PubMed: 15308962]

[13]. Wærsted M, Børnick TS, Twisk JWR, Veiersted KB. Simple descriptive missing data indicators in longitudinal studies with attrition, intermittent missing data and a high number of follow-ups. BMC Res Notes 2018;11(1):123. 10.1186/S13104-018-3228-6. [PubMed: 29433533]

[14]. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. In: Statistical methods in medical research. 21. London, England: SAGE Publications Sage UK; 2012. p. 243–56. 10.1177/0962280210394469. [PubMed: 21389091]

[15]. van Smeden M, de Vries BBLP, Nab L, Groenwold RHH. Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies. J. Clin. Epidemiol 2021;131:89–100. 10.1016/J.JCLINEPI.2020.11.006. [PubMed: 33176189]

[16]. IR W, P R, AM W. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med 2011;30(4):377–99. 10.1002/SIM.4067. [PubMed: 21225900]

[17]. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. Int. J. Epidemiol 2019;48(6):2050–60. 10.1093/ije/dyz203. [PubMed: 31633184]

[18]. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int. J. Epidemiol 2002;31(6):1246–52. 10.1093/IJE/31.6.1246. [PubMed: 12540730]

[19]. Harron KL, Doidge JC, Knight HE, et al. A guide to evaluating linkage quality for the analysis of linked data. Int. J. Epidemiol 2017;46(5): 1699–710. 10.1093/IJE/DYX177. [PubMed: 29025131]

[20]. Hernan MA, Robins JM. Causal inference: What if. Chapman & Hall/CRC; 2020.

[21]. Luke B, Gopal D, Cabral H, Stern JE, Diop H. Adverse pregnancy, birth, and infant outcomes in twins: effects of maternal fertility status and infant gender combinations; the Massachusetts outcomes study of assisted reproductive technology. Am. J. Obstet. Gynecol 2017;217(3):330.e1–330.e15. 10.1016/j.ajog.2017.04.025. [PubMed: 28455086]

[22]. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988;44(4):1049. 10.2307/2531734. [PubMed: 3233245]

[23]. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models73; 1986.

[24]. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38(4):963. 10.2307/2529876. [PubMed: 7168798]

[25]. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Regression analysis part I: Model specification. In: Modern epidemiology. 4th ed. Wolters Kluwer; 2021. p. 473–503 Accessed October 12, 2021 https://shop.lww.com/Modern-Epidemiology/p/9781451193282.

[26]. Greenland S Modeling and variable selection in epidemiologic analysis. Am. J. Public Health 1989;79(3):340–9. 10.2105/AJPH.79.3.340. [PubMed: 2916724]

[27]. Davison AC, Tsai C-L. Regression model diagnostics. Int Stat Rev Revue Int Stat 1992;60(3):337. 10.2307/1403682.

[28]. Atkinson A Plots, transformations, and regression. Oxford: Clarendon Press; 1985.

[29]. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. Stat. Methods Med. Res 2013;22(1):14–30. 10.1177/0962280211403597. [PubMed: 21705435]

[30]. Pearl J Causal diagrams for empirical research. Biometrika. 1995;82(4):669–88. 10.1093/BIOMET/82.4.669.

[31]. VanderWeele TJ. Principles of confounder selection. Eur. J. Epidemiol 2019;34(3): 211–9. 10.1007/S10654-019-00494-6. [PubMed: 30840181]

[32]. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46(3): 399–424. 10.1080/00273171.2011.568786. [PubMed: 21818162]

[33]. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am. J. Epidemiol 2007; 166(6):646–55. 10.1093/AJE/KWM165. [PubMed: 17615092]

[34]. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics. 1998;54(3):948. 10.2307/2533848. [PubMed: 9750244]

[35]. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J. R. Stat. Soc 1983;45(2):212–8. 10.1111/J.2517-6161.1983.TB01242.X.

[36]. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. 2009. 10.1007/978-0-387-87959-8.

[37]. Greenland S, Copas J, Jones DR, et al. Multiple-bias modelling for analysis of observational data. J Royal Stat Soc Ser A: Stat Soc 2005;168(2):267–306. 10.1111/j.1467-985X.2004.00349.x.

[38]. TJ V, OA A. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology 2011;22(1):42–52. 10.1097/EDE.0B013E3181F74493. [PubMed: 21052008]

[39]. Gustafson P Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments. Chapman & Hall/CRC; 2004.

[40]. P G, LC M. Probabilistic approaches to better quantifying the results of epidemiologic studies. Int. J. Environ. Res. Public Health 2010;7(4):1520–39. 10.3390/IJERPH7041520. [PubMed: 20617044]

[41]. TJ V, P D. Sensitivity analysis in observational research: introducing the E-value. Ann. Intern. Med 2017;167(4):268–74. 10.7326/M16-2607. [PubMed: 28693043]

[42]. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and Bias in observational studies. Epidemiology 2010;21(3):383. 10.1097/EDE.0B013E3181D61EEB. [PubMed: 20335814]

[43]. Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. Biom. J 2018;60(3):431–49. 10.1002/bimj.201700067. [PubMed: 29292533]

[44]. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. Int. J. Epidemiol. 2016;45(2):565–75. 10.1093/IJE/DYW040. [PubMed: 27097747]

[45]. Efron B. Prediction, estimation, and attribution. J. Am. Stat. Assoc 2020;115(530): 636–55. 10.1080/01621459.2020.1762613.

[46]. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JAC. Methods for dealing with time-dependent confounding. Stat. Med 2013;32(9):1584–618. 10.1002/sim.5686. [PubMed: 23208861]

[47]. Goldstein ND, Hamra GB, Harper S. Are descriptions of methods alone sufficient for study reproducibility? An example from the cardiovascular literature. Epidemiology. 2020;31(2):184–8. 10.1097/EDE.0000000000001149. [PubMed: 31809339]

[48]. Saint-Maurice PF, Troiano RP, Matthews CE, Kraus WE. Moderate-to-vigorous physical activity and all-cause mortality: do bouts matter? J. Am. Heart Assoc 2018;7(6). 10.1161/JAHA.117.007678.

[49]. Goldstein ND, Hamra GB, Harper S. Are descriptions of methods alone sufficient for study reproducibility? An example from the cardiovascular literature. Epidemiology. 2020;31(2):184–8. 10.1097/EDE.0000000000001149. [PubMed: 31809339]

[50]. Centers for Disease Control and Prevention (CDC). National health and nutrition examination survey questionnaire. National Center for Health Statistics (NCHS); 2020. Published online.

[51]. Centers for Disease Control and Prevention (CDC). National health and nutrition examination survey laboratory protocol. Published online. 2020.

[52]. Mirel LB, Mohadjer LK, Dohrmann SM, et al. National health and nutrition examination survey: estimation procedures, 2007–2010. In: Vital and health statistics Series 2, Data evaluation and methods research. 159; 2013. p. 1–17.

[53]. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 3rd. Springer; 2012. p. 161–91.

[54]. Ernst AF, Albers CJ. Regression assumptions in clinical psychology research practice-a systematic review of common misconceptions. PeerJ. 2017;2017(5). 10.7717/peerj.3323.

[55]. Hamra GB, Goldstein ND, Harper S. Resource sharing to improve research quality. J. Am. Heart Assoc 2019;8(15). 10.1161/JAHA.119.012292.

[56]. Savitz DA, Wellenius GA. Interpreting epidemiologic evidence: Connecting research to applications. 2nd ed. Oxford University Press; 2016.

[57]. Lash TL, Fox MP, Maclehose RF, Maldonado G, Mccandless LC, Greenland S. Good practices for quantitative bias analysis. Int. J. Epidemiol 2014;43(6): 1969–85. 10.1093/ije/dyu149. [PubMed: 25080530]

Table 1

Summary of the potential impact on causal inference given various practical considerations of epidemiological data.

Research stage	Practical consideration	Hypothetical example	Potential impact on inference	Possible technical solutions
Data Management	1. Missing data	Complete case analysis omits important data	Selection or information bias	Multiple imputations using chained equations [16] or inverse probability weights of censoring [15,29]
	2. Duplicate observations	Data reported from a registry without de-duplication	Selection bias	Calculate predicted values of record linkage [18]
	3. Inconsistent variable definition	Data linkage with resulting inconsistent operationalization	Information bias	Comparison of linked and unlinked data, sensitivity analysis of linkage procedure [19]
Analysis	4. Study design	Failure to consider an appropriate model for the survey design	Biased error	Evaluate research questions and hypotheses to implement appropriate model
	5. Model specification and assumption	Unresolved heteroskedasticity, relationship not linear, or correlated observations	Biased error	Evaluate distributions of data and use of model diagnostics
	6. Variable selection	Inclusion or omission of covariates in the statistical model, mismeasurement of key variables	Uncontrolled confounding or information bias	Causal diagrams [31], bias analysis [39], *E*-values [41], or negative controls [42]