Published in final edited form as:

Ann Epidemiol. 2018 November ; 28(11): 821–828. doi:10.1016/j.annepidem.2018.09.001.

Epidemiologic analyses with error-prone exposures: review of current practice and recommendations

Pamela A. Shaw, PhD^{a,*}, Veronika Deffner, PhD^b, Ruth H. Keogh, DPhil^c, Janet A. Tooze, PhD^d, Kevin W. Dodd, PhD^e, Helmut Küchenhoff, PhD^b, Victor Kipnis, PhD^e, Laurence S. Freedman, PhD^{f,g} Measurement Error and Misclassification Topic Group (TG4) of the STRATOS Initiative

^aDepartment of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia

^bDepartment of Statistics, Statistical Consulting Unit StaBLab, Ludwig-Maximilians- Universität, Munich, Germany

^cDepartment of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

^dDepartment of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Blvd, Winston–Salem, NC

^eDivision of Cancer Prevention, Biometry Research Group, National Cancer Institute, Bethesda, MD

^fInformation Management Services, Inc, Rockville, MD

^gBiostatistics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel

Abstract

Purpose: Variables in observational studies are commonly subject to measurement error, but the impact of such errors is frequently ignored. As part of the STRENGTHENING Analytical Thinking for Observational Studies Initiative, a task group on measurement error and misclassification seeks to describe the current practice for acknowledging and addressing measurement error.

Methods: Task group on measurement error and misclassification conducted a literature survey of four types of research studies that are typically impacted by exposure measurement error: (1) dietary intake cohort studies, (2) dietary intake population surveys, (3) physical activity cohort studies, and (4) air pollution cohort studies.

Results: The survey revealed that while researchers were generally aware that measurement error affected their studies, very few adjusted their analysis for the error. Most articles provided incomplete discussion of the potential effects of measurement error on their results. Regression calibration was the most widely used method of adjustment.

*Corresponding author. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, 606 Blockley Hall, Philadelphia, PA 19104. Tel.: 215-898-9599. shawp@upenn.edu (P.A. Shaw).

Supplementary data

Supplementary data associated with this article can be found in the online version at <https://doi.org/10.1016/j.annepidem.2018.09.001>.

Conclusions: Methods to correct for measurement error are available but require additional data regarding the error structure. There is a great need to incorporate such data collection within study designs and improve the analytical approach. Increased efforts by investigators, editors, and reviewers are needed to improve presentation of research when data are subject to error.

Keywords

Air pollution; Cohort studies; Measurement error; Misclassification; Nutritional epidemiology; Physical activity

Introduction

Measurement error is a challenge in many settings in epidemiology. Exposures such as dietary intakes, environmental contaminants, and physical activity are difficult to measure because patterns of exposure are complex and because accurate (unbiased) and precise (with minimal variability) ways to measure many such exposures of interest are either unavailable or are too impractical to use in a large study. Practical measures for these exposures are error prone, that is, they will contain sizable random deviations from a target exposure, such as a short-term mean dietary intake due to biological variability or assay error, and also potentially systematic bias, for example, from inaccurate self-reported exposures. Here, we refer to both of these kinds of deviations as measurement error, with random error defined as mean zero and independent error and systematic error defined as covariate-dependent bias. In some cases, through intensive monitoring and/or better instruments, more accurate and precise measurements are available in a subset of subjects or from an independent validation study. These reference, or gold standard, measures can provide the data necessary to apply a method that corrects for the instrument error [1–6] or a quantitative bias analysis [7–11]. It is well-established in the statistical and epidemiological literature that if the measurement error in an exposure variable is ignored, analyses can be subject to biased estimation and incorrect inference [1–11].

Analysis techniques to address exposure measurement error have been the focus of methodologic research for several decades. These efforts have produced many methodological advances in both analysis techniques and study designs. Statistical [1–4] and epidemiological texts [5,6] have summarized a number of these methods. Several review articles describe existing methods or compare methods in specific settings [12–16]. In addition, many articles in epidemiologic journals have advocated that quantitative bias analyses be provided for any analysis that involves error-prone exposure measures [7–11]. Despite these efforts, a surprising number of articles have been published in the biomedical literature with no adjustment in the data analysis and little to no discussion of how measurement error potentially impacted the study results [17]. This has been true even in research areas, such as nutritional epidemiology, where there is a well-established literature in topic matter journals, instrument-specific software [18], and webinars to make these methods more accessible [19].

The international STREngthening Analytical Thinking for Observational Studies (STRATOS) Initiative is a large collaboration of experts in many different areas of

biostatistical research that was formed in response to an observation that many methodological advances in statistics are not put to practice and that the design and analysis of observational studies commonly exhibit serious weaknesses [20]. The objective of STRATOS is to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies. The STRATOS Initiative to date formed working groups in nine topics: missing data, selection of variables and functional forms in multivariate analysis, initial data analysis, measurement error and misclassification, study design, evaluating diagnostic tests and prediction modeling, causal inference, survival analysis, and high-dimensional data.

In this article, we present the results of a literature survey done by the STRATOS Measurement Error and Misclassification Topic Group (TG4) [21] to assess current practice for handling measurement error in the biomedical literature. We performed literature surveys in four areas of epidemiology where measurement error is a well-known concern: (1) dietary intake cohort studies, (2) dietary population surveys, (3) physical activity cohort studies, and (4) air pollution cohort studies. In the cohort studies, we were specifically interested in analyses of the association between an error-prone exposure and outcome, and in the dietary population surveys, we were specifically interested in analyses used to estimate the distribution of intake of a dietary component (a nutrient or food). The purpose of the survey was to examine whether investigators are using appropriate statistical methods to adjust for or assess the potential effects of measurement error on study results, and to what degree authors do or do not discuss the impact of measurement error on their results. We also describe which methods are used by those who address measurement error in their analysis. We present the survey results, highlighting common problems relevant to anyone performing or interpreting statistical analyses with error-prone exposures. We conclude with recommendations for how to overcome the short comings in current practice for statistical analyses, and consequently in the resulting scientific conclusions, in fields where measurement error remains a challenge.

Materials and methods

Overview

STRATOS TG4 conducted a literature survey of four research areas: (1) dietary intake cohort studies, (2) dietary intake population surveys, (3) physical activity cohort studies, and (4) air pollution cohort studies. For the three cohort study literature surveys, articles for review were identified by two types of search: (A) a search with general search terms related to the topic area and (B) a similar search with additional required terms related to measurement error or misclassification (Table 1). The purpose of search A was to identify articles for a general review of the topic areas to understand the current practice for how error-prone exposures are handled. Search B, performed only for the cohort studies, was done in expectation that few articles from search A would have a measurement error adjusted analysis. The purpose of search B was to identify articles that in some way did address measurement error or misclassification in the analysis, to be able to summarize which methods are currently in use. For dietary cohort studies, using a method to adjust for the mismeasured exposure was required to be eligible for search B; due to a lack of such

articles, measurement error terms were only included as search terms for the other cohort studies. For the dietary intake survey, only search A was performed because, while issues of variability around usual intake are appreciated in this setting, the terms misclassification or measurement error are typically not used.

A general protocol that specified the questions in the literature survey across the topic areas was developed in advance, to assess aspects of the study design or analysis that would be informative regarding how measurement error may impact results, how it was addressed, and whether any limitations regarding bias or reliability was directly addressed (Supplemental Materials). Questions examined the statistic of primary interest, whether measurement error in the exposure variable was mentioned as a potential problem, whether a reliability (repeated measures) or calibration (comparison with a reference instrument) substudy was included, and whether any methods were used to address measurement error in the data analysis. We performed separate literature surveys by topic area, tailored to increase relevance for each specific setting. For the cohort studies, the error-prone measure of interest was required to be an exposure and not an outcome in the presented analyses. For the dietary intake and physical activity cohort surveys, we included questions regarding whether exposures were analyzed as categorical or continuous variables. In those surveys, we also collected information on whether multiple error-prone exposures were included in the regression analysis, focusing on the common examples of (1) physical activity and dietary intake and (2) dietary intake and smoking. Data extraction instruments were reviewed by the TG4 working group before initiation of the literature search. We performed the literature search in three stages: (1) identified research articles using PubMed/web of science, (2) reviewed titles and abstracts to select articles in scope, and (3) conducted detailed review of selected articles for data abstraction, making further exclusions if necessary. Meta-analyses, review articles, interventional studies, and retrospective case-control studies were excluded. Further details by topic area are provided in the following.

Reviews of search A and search B articles were done by one or more primary reviewers, generally one reviewer per article. For purposes of quality control (QC), a 20% subsample, stratified by reviewer and search type, was randomly chosen for review by a second reviewer. Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) flow diagrams are provided in the Supplemental Tables and Figures [22]. At the time of finalizing this article (May 2018), a limited survey (30% the size of the original review) was performed to update the study results with the most recent literature (starting with Jan 2017).

Dietary intake cohort study

Search A was restricted to the most recent 12 months and a few disease areas for which diet is a well-known exposure of interest, namely one of “cancer,” “cardiovascular disease,” and “diabetes.” Search A also included the term “risk” to try to separate studies of associations between dietary exposures and health outcomes from the many articles considering adequacy of intake for a certain population. The ePub date June 2, 2014—June 2, 2015 captured 51 articles. Search terms are provided in Table 1.

Search B required the terms “measurement error” or “misclassification” and was expanded to three different searches and the previous 15 years to identify a target of 30 qualifying

articles (criteria in Table 1). Ultimately 31 search B articles with ePub dates January 1, 2001—July 15 2015 were selected for full review. The randomly selected QC subsample included 10 search A and six search B articles, split equally by two primary reviewers (R.H.K., P.A.S.) and reviewed independently by a third reviewer (V.K.).

Dietary intake population survey

Because we anticipated that many studies would assess an individuals' intake with multiple 24-hour recalls, questions were added to assess whether the within-person variability was mentioned or accounted for in the analysis. Articles were identified in one query in PubMed with date range: 01/01/2012–08/18/2015 and search terms that included a variant of the word “nutrition” and terms related to typical dietary intake survey instruments (Table 1). The query returned 2801 articles; title and abstract review was performed by a single primary reviewer (K.W.D.) on the most recent 717 identified (ePub date March 2014–August 2015). The review was restricted to surveys whose aim was to describe a population of some geographic region. The first 67 articles identified as within scope were given detailed review; all 67 were confirmed eligible and information was extracted. A QC sample of 13 articles was randomly selected and reviewed by a second, independent reviewer (L.S.F.).

Physical activity cohort study

The construct of physical activity has been captured using multiple different measures (e.g. self-reported minutes of moderate and vigorous activity, sedentary behavior, or total activity counts from an accelerometer), which may lead to different types of measurement error. Thus, additional questions assessing the physical activity measures were added to this survey. For search A, a query in PubMed was performed with date range 07/01/2012–06/30/2015 and required terms that aimed to narrow the search to prospective cohort studies with physical activity exposures (see Table 1). For search B, terms relating to measurement error and misclassification were added to the query required terms (Table 1). There were 8760 articles returned from search A and 610 from search B. We selected a random subset of search A articles equal to the number of search B articles identified. After abstract review, there were 50 articles from search A and 87 from search B determined to be eligible for data abstraction. Fifty articles from each search (a random subset for search B) were selected for review. Upon detailed review, there were 30 articles abstracted from search A and 39 articles from search B. Reviews were done by a single primary reviewer (J.A.T.) and a QC sample of 10 randomly selected search A and 10 search B articles were reviewed by an independent reviewer (L.S.F.).

Air pollution cohort

For these articles, our survey additionally assessed whether the exposure was measured by fixed-site monitors or by devices meant to directly capture personal exposure, as this choice can lead to very different sources of measurement error (e.g. Berkson or classical error). We also assessed temporal resolution. For search A, a query in web of science was performed with date range 01/01/2012–12/31/2014 and the search terms shown in Table 1. The search was restricted to prospective cohort studies with health outcomes (e.g. cardiovascular disease, cancer mortality, hospital admissions, and respiratory disease) that are often subjects of research on the impact of air pollution on human health. For search B,

measurement error terms and “health” were added. In addition, a general search with the terms “measurement error” and “air pollution” was conducted.

Search A returned 4682 articles; search B returned 386 articles. After title/abstract review, there were 451 eligible articles for search A and 32 for search B. For search A, randomly selected articles were read in detail (Matthias Assenmacher, Veronika Deffner, Andreas Hueck, Helmut Küchenhoff, Thomas Maierhofer), and data were extracted until 50 eligible articles were identified. For search B, all 32 articles were reviewed (Veronika Deffner) and 25 found eligible upon detailed review. Independent reviewers (Nathan Huey, Matthias Assenmacher) extracted data for a randomly selected QC sample of 10 articles for search A and five articles for search B.

Results

We describe the results of each literature survey separately. Tables 2 and 3 summarize the main results for searches A and B, respectively. Supplemental eTables 1 and 2 present results for the survey update and are similar to those for the original survey presented in this section.

Dietary intake cohort study

In the general literature survey (search A), 46 (90%) of articles analyzed the nutritional intake exposures with no adjustment for measurement error in the analysis. Of the five that performed some adjustment, two (4%) used regression calibration [1,28] and three (6%) used a cumulative average of multiple assessments over time. Most authors (48, 94%) did acknowledge in some way that their exposure was prone to measurement error. Only 31 (61%) mentioned that their reported association with the error-prone exposure was subject to bias, whereas the remaining articles either made no or a vague reference to errors in the exposure measurement, such as that their instrument had been validated and had an acceptable reliability coefficient, which may or may not have been provided. When measurement error was mentioned, frequently an incomplete or erroneous claim was made about its impact on the presented analysis. A common incomplete claim was authors stated the effect estimates were subject to attenuation bias, but did not mention they were also subject to inflation bias, which can occur in many settings such as if there are multiple error-prone variables, categorized error-prone exposures or differential measurement error. Erroneous claims also included that the study was not subject to bias because it was prospective or because their instrument had good reliability. None described adjusting the study design to accommodate the error in the study measurement, that is, increasing sample size to offset loss of power. Nearly all articles (50, 98%) categorized their error-prone exposure and 27 (53%) exclusively analyzed the exposure as categorical.

For search B, four of 31 articles upon detailed review were found not to have addressed exposure error in the analysis and were excluded; three reported no method and one reported using only an energy-adjustment method. This latter article was discarded from the eligible search B articles as energy-adjustment methods are generally not considered a method to address error in the exposure. Thus, 27 articles were included in the reported search B analysis. All but one article (96%) used regression calibration [1,28] to address the error;

one of these articles [31] additionally used SiMulation Extrapolation [29]. The remaining article considered both an average of multiple measures and a latent variable technique as alternative analysis approaches. Among the 26 articles that used regression calibration, six were based on a reliability substudy (repeated measures of the same instrument), 14 calibrated to a different self-report instrument or food record (FR), four calibrated to an objective recovery biomarker, one performed calibration separately by three different instruments (biomarker, food frequency questionnaire [FFQ], FR) and one did not report the calibration instrument. Twenty-four (89%) search B articles reported using a method that adjusted the standard error estimation for the extra uncertainty induced by the measurement error adjustment. None discussed considering the error in the study design.

For search B articles, there were no explicitly erroneous claims about error, but 44% had an incomplete discussion. Incomplete discussions included failing to mention the limitations of using a reference instrument with errors correlated with those of the main study instrument and failing to acknowledge that calibrating only for within-person variability may have been inadequate due to systematic errors in the main study instrument (such as a FFQ).

Several articles had more than one error-prone exposure in their association analysis. Seventy-six percent of search A and 81% of search B articles reported having considered self-reported physical activity as an exposure in a multivariate regression; none of the search A and 2 (7%) of the search B articles explicitly reported adjusting their analysis for errors in both physical activity and nutritional intake. There were 92% (89%) of the search A (B) articles that additionally adjusted for self-reported smoking, generally also considered to be error prone. None adjusted for errors in the smoking variable.

Dietary intake population survey

The selected dietary intake survey articles investigated measures of central tendency (mean/median) of intake or fraction consuming specific dietary components, ranking food/food group contributors to overall diet, and examining the distribution of intake (i.e., percentiles, fractions consuming less or more than specified limits (such as a recommended daily intake), and quantities such as a dietary pattern score).

Most (53, 79%) of the articles assessed dietary intake using one or more 24-hour recalls. Of the remaining 14 articles, three analyzed diaries/FRs, and 11 analyzed FFQs. Of these 14, all but one (13, 93%) recognized that the self-report instrument used could be subject to underreporting or bias, but generally (10, 77%) such recognition was presented only in the Discussion of the article as a limitation. Of the 53 articles concerned with 24-hour recalls, 32 (60%) presented analyses of a single recall (even if multiple recalls were available), and 21 (40%) presented analysis including multiple recalls on at least some respondents. Both single-recall and multirecall articles were likely (69%/86%) to note that 24-hour recalls could be subject to underreporting/systematic bias due to measurement error. Complex modeling methods are available to estimate the full distribution of usual intake from 24-hour recall data, considering features such as within-subject variability and episodic consumption [23,24,26,32–34]. These methods can also be used to estimate fractions with usual intake below fixed cut points or distributions of scores ostensibly computed on usual intake. However, these methods generally cannot be applied when only one 24-hour recall per

respondent is available. Nevertheless, 10 (31%) of the 32 articles with a single recall did present statistics other than the mean, which were subject to bias. By comparison, 15 of the 21 articles with multiple recalls (71%) also presented these types of statistics (although they did not necessarily use the complex methods). Half (16, 50%) of the single-recall articles made no mention of within-person variation or usual intake, compared to five (24%) of the multirecall articles. When the single-recall articles mentioned within-person variation or usual intake, it was often in the context of justifying their analysis on the basis that a single recall can be used to estimate the mean of the usual intake distribution under the classical (independent and unbiased) measurement error assumption. Overall, only 19 (28%) used a method to adjust their analyses for measurement error (Tables 2 and 3).

Physical activity cohort study

Most studies examined in searches A (24, 80%) and B (37, 92%) were prospective cohort studies. A number of different constructs of physical activity were measured in both searches, with the most common being minutes of moderate or vigorous activity (search A: 11, 37%; search B: 9, 22%), sedentary activity (A: 6, 20%; B: 5, 12%), and adherence to guidelines (A: 3, 10%; B: 7, 18%). Other constructs included metabolic equivalents minutes, activity energy expenditure, total energy expenditure, and various scale-based or ad hoc measures of summarizing activity. Most authors used only subjective measures of activity (A: 25, 83%; B: 34, 85%); a small number (A: 4, 13%; B: 5, 12%) used only objective measures, such as those from an accelerometer and one from each search used both. In search A, 21 (70%) categorized the physical activity variable, in search B, 31 (78%) did.

None of the physical activity articles identified in search A analyzed physical activity with adjustment for measurement error, although over half of the articles (18, 60%) mentioned measurement error or misclassification as a limitation. Of these 18 articles, 13 mentioned bias due to self-report: of these, six mentioned attenuation, five mentioned that physical activity may be over-reported, and four hypothesized that the error was most likely nondifferential, that is, not associated with the outcome of interest, and therefore likely to lead to attenuation, although in fact this is not necessarily the case. Only two articles mentioned designing the study to account for measurement error. None of the studies examined had a calibration substudy. In search A, three studies (10%) included repeated measures, but only to assess change over time, not to address repeatability; 11 (37%) of the primary analysis regression models also adjusted for nutritional self-report exposures.

In search B, five articles (12%) mentioned that measurement error was considered in some way in the study design; four of these (10%) had a calibration substudy, and two (5%) had an adjustment for measurement error. Both studies used a form of regression calibration. Overall, 28 articles (70%) mentioned measurement error as a limitation, and four articles specifically addressed reliability or validity of measures. Of these 28 articles, 21 (75%) mentioned bias due to self-report; seven mentioned measurement error may have attenuated the estimated relationship with the outcome; two mentioned that physical activity may be overreported; two mentioned power loss, and two hypothesized that the error was most likely non-differential. Interestingly, one article acknowledged their use of questionnaires could result in error and residual confounding, with an unknown magnitude and direction of

bias and that categorization could reduce power, and introduce differential error [35]. Three studies (8%) used repeated measures to assess within-person variability. In search B, 23 (58%) of the primary analysis models also included adjustment for nutritional self-report data.

Air pollution cohort study

Table 4 describes characteristics of the search A articles. In the search A articles, an individual's exposure was predominantly measured at fixed-site monitoring stations (35, 70%) that recorded data on hourly or daily temporal resolution (30, 60%). In general, measurement error in air pollution cohort studies arises by temporal and/or spatial aggregation of the exposure data from a fixed monitor that is then applied to an individual, limited availability of exposure measurements with temporal and/or spatial variability, and error in the instruments themselves. Fewer than half the articles (23, 46%) mentioned measurement error as a potential problem, and those that did mention it did not describe its sources or its impact on the study results in detail. Four (8%) search A articles used a method to address the measurement error in their analysis. One study used a type of regression calibration to account for systematic measurement error in the older of two instruments, but the method was not denoted as regression calibration. Three studies conducted sensitivity analyses. Two (4%) of all the studies included a subsample in which a reference measure was available enabling examination of measurement error (one for regression calibration and one to measure reproducibility).

In the search B articles identified, the measurement error problem was predominantly only mentioned or discussed, but not formally analyzed. Only five (20%) studies applied methods to address the measurement error; all five studies used aggregated exposures (fixed-site exposure measurements or estimated exposure values). A single study applied an instrumental variable approach to deal with potential measurement error; the authors of the other four studies performed sensitivity analyses to describe the robustness of their results under different assumptions concerning the amount of measurement error present.

Discussion

Dietary intake and physical activity exposures were typically collected using self-reported data, which can be subject to systematic errors dependent on the true value and classical (random) errors independent of the true value. Most air pollution studies used fixed-site monitoring stations, which can have complex error structures including systematic and classical errors, as above, but also Berkson error (i.e., random error independent of the measured value, but arises due to measuring aggregate rather than individual exposure) [1]. Yet, for most articles reviewed in the cohort studies, there was an inadequate discussion of the impact of measurement error on study results. Several authors had no or only vague discussion of the measurement error, stating only that it was present in the exposure measurement but not being clear on its origin, size, structure, or potential impact on estimated associations. Consideration was also not given to the impact of categorizing a continuous error-prone variable, which can be unpredictable in terms of direction and size of bias of the regression coefficients [36]. Among authors who did discuss measurement error's

impact on their association analysis, several incorrectly claimed that attenuation was the only possible direction of bias induced by error. Although for dietary intake exposures it has been noted that the predominant direction of bias has been observed to be attenuation, attenuation is not generally guaranteed in multivariable models [37]. Furthermore, adjusting for multiple error-prone exposures was common and the fact that this can influence the direction of the bias for both error-prone and precise predictors was generally not mentioned. Authors ignoring errors were also not adjusting their analysis for the uncertainty in the exposures, thus overstating the precision of the target association. A few articles that appropriately discussed the impact of measurement error are highlighted in the Supplemental Materials.

In nearly all studies with multiple error-prone exposures, error in at most one exposure was directly addressed. It was also clear that authors were not fully taking advantage of available information in their study regarding the structure of the measurement error, such as repeat measurements in settings where they could have been used to assess the impact of within-person variation on study estimates. Many also cited validation studies of the exposure instruments, which may have included simultaneous assessment with a reference instrument, and therefore may have been used for calibration.

There were a few topic-specific themes that also arose. In many dietary surveys, interest is in “usual intake,” defined as long-term average daily intake of a dietary component. Researchers in that niche commonly use 24-hour recalls or FRs to assess dietary intake and seem willing to make the working assumption that their instruments are unbiased, although many concede in their discussion that this assumption is probably violated. Even those applying complex methods to derive usual intake, to adjust study estimates for error in the dietary intake measurements, generally attempt to adjust for only within-person variation and not systematic bias.

Our review of the physical activity literature found that multiple different constructs are used to describe physical activity. Some constructs use time and not intensity, some focus only on leisure activities; some provide continuous measures, while most are categorized. The choice of construct is important as different measures are subject to different types of bias; however, regardless of construct, there was very little attention in the analysis or interpretation of results regarding the impact of measurement error.

The measurement error problem in air pollution cohort studies arises from a complex exposure, and the error structure can vary considerably by study design. Berkson error is prevalent [1]. Only a few of these studies discuss the underlying assumptions for validity of estimated parameters for exposures containing Berkson error. Comments about Berkson error are mostly confined to discussion of the spatial variability of the actual exposure; potential biases due to Berkson error in complex settings like air pollution studies are usually not discussed. Exposure measurements with classical measurement error occur in studies with personal exposure measurements; none of the five identified articles with personal exposure measurements applied any adjustment for measurement error.

Conclusion

The presented literature survey reveals that articles with inadequate treatment of exposure measurement error and misclassification in the analysis and discussion of their study results remain commonplace in the literature. We focused on covariates prone to mismeasurement, as this setting has been the dominant focus of existing methods to address measurement error; however, we expect similar problems exist in published analyses that also involve error-prone outcomes, for which the naive analysis is also prone to bias. Investigators, reviewers, editors, and consumers of the literature all have a role to play in improving the quality of observational studies that rely on measures that are subject to systematic and random errors. More attention to these issues needs to be paid at the peer review stage. It is important that reviewers and editors be alert to the problems of measurement error and demand authors give some consideration of its impact in their research article. Several authors present the case that quantitative bias analyses, which assess the robustness to plausible assumptions about the nature of the measurement error, should be part any analysis that involves error-prone exposure measures [7–11]. As consumers of these studies, we need to take care to not cite or use the results of these studies without some acknowledgment of their limitations. It is perhaps only when an incomplete treatment of measurement error threatens the success of publication, that authors will be willing to invest the necessary effort into more fully addressing this limitation of their studies. With professional outreach, such as the activities of STRATOS that include preparation of a guidance article on measurement error and statistical methods to mitigate their bias, hopefully more investigators will have a better understanding of the impact of instrument bias and measurement error and the possible ways to address them. Available data will determine which approaches of addressing measurement error are feasible for a given study. At a minimum, authors should state very explicitly the assumptions they are making about the structure of measurement error and the possible impact of those errors on their results. Oftentimes, there is at least some information about measurement properties of an instrument, such as from previous validation or reliability study, that can be used as a starting point for sensitivity analyses. In short, the current practice for presentation of results from studies with appreciable measurement error in the principle exposure measurement(s) needs to improve and in many studies this could be achieved using readily available resources and methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

The authors would like to thank Matthias Assenmacher, Andreas Hueck, Thomas Maierhofer, Nathan Huey at Ludwig-Maximilians-Universität for their assistance with the article reviews, Julie Herbinger for initial article screening, and Mingh Anh Le for a pilot study for the air pollution cohort study literature survey. R.H.K. is supported by a Medical Research Council Fellowship (MR/M014827/1). J.A.T. is supported in part by National Cancer Institute Cancer center Support Grant P30 CA012197.

References

- [1]. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models: A Modern Perspective. Boca Raton: CRC press; 2006.
- [2]. Gustafson P Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Boca Raton: CRC Press; 2003.
- [3]. Fuller WA. Measurement Error Models. New York: John Wiley & Sons; 2009.
- [4]. Buonaccorsi JP. Measurement Error: Models, Methods, and Applications. Boca Raton: CRC Press; 2010.
- [5]. White E, Armstrong BK, Saracci R. Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating and Improving Measures of Disease Risk Factors. Oxford: University Press; 2008.
- [6]. Willett W Nutritional Epidemiology. Oxford: Oxford University Press; 2012.
- [7]. Fox MP. Creating a demand for bias analysis in epidemiological research. J Epidemiol Community Health 2009;63(2):91. [PubMed: 19141660]
- [8]. 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. [PubMed: 25080530]
- [9]. Greenland S Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996;25(6):1107–16. [PubMed: 9027513]
- [10]. Fox MP, Lash TL. On the Need for Quantitative Bias Analysis in the Peer-Review Process. Am J Epidemiol 2017;185(10):865–8. [PubMed: 28430833]
- [11]. MacLehose RF, Olshan AF, Herring AH, Honein MA, Shaw GM, Romitti PA. Bayesian methods for correcting misclassification: an example from birth defects epidemiology. Epidemiology 2009;20(1):27–35. [PubMed: 19234399]
- [12]. Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. Stat Med 2008;27(25):5195–216. [PubMed: 18680172]
- [13]. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. Stat Med 2008;27(30):6332–50. [PubMed: 18937275]
- [14]. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol 2006;35(4):1074–81. [PubMed: 16709616]
- [15]. White IR. Commentary: Dealing with measurement error: multiple imputation or regression calibration? Int J Epidemiol 2006;35(4):1081–2. [PubMed: 16847023]
- [16]. Bang H, Chiu YL, Kaufman JS, Patel MD, Heiss G, Rose KM. Bias Correction Methods for Misclassified Covariates in the Cox Model: comparison of five correction methods by simulation and data analysis. J Stat Theory Pract 2013;7(2):381–400. [PubMed: 24072991]
- [17]. Jurek AM, Maldonado G, Greenland S, Church TR. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. Eur J Epidemiol 2006;21(12):871–6. [PubMed: 17186399]
- [18]. <https://epi.grants.cancer.gov/diet/usualintakes/macros.html>. [Accessed 19 July 2018].
- [19]. <https://epi.grants.cancer.gov/events/measurement-error/>. [Accessed 19 July 2018].
- [20]. Sauerbrei W, Abrahamowicz M, Altman DG, Cessie S, Carpenter J. Strengthening analytical thinking for observational studies: The STRATOS initiative. Stat Med 2014;33(30):5413–32. [PubMed: 25074480]
- [21]. The STRATOS Initiative Topic Group 4: Measurement error and Misclassification. http://stratos-initiative.org/group_4. [Accessed 19 July 2018].
- [22]. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. J Clin Epidemiol 2009;62(10):1006–12. [PubMed: 19631508]
- [23]. Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semiparametric transformation approach to estimating usual daily intake distributions. J Am Stat Assoc 1996;91:1440–9.

- [24]. Haubrock J, Nöthlings U, Volatier JL, Dekkers A, Ocké M, Harttig U, et al. Estimating usual food intake distributions by using the Multiple Source Method in the EPIC-Potsdam Calibration Study. *J Nutr* 2011;141:914–20. [PubMed: 21430241]
- [25]. Harttig U, Haubrock J, Knüppel S, Boeing H. The MSM program: web-based statistics package for estimating usual dietary intake using the Multiple Source Method. *Eur J Clin Nutr* 2011;65(S1):S87–9. [PubMed: 21731011]
- [26]. Tooze JA, Midthune D, Dodd KW, Freedman LS, Krebs-Smith SM, Subar AF, et al. A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application to Their Distribution. *J Am Diet Assoc* 2006;106:1575–87. [PubMed: 17000190]
- [27]. Tooze JA, Kipnis V, Buckman DW, Carroll RJ, Freedman LS, Guenther PM, et al. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Stat Med* 2010;29(27):2857–68. [PubMed: 20862656]
- [28]. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982;69(2):331–42.
- [29]. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *J Am Stat Assoc* 1994;89(428):1314–28.
- [30]. Wientzek A, Tormo Díaz MJ, Castaño JM, Amiano P, Arriola L, Overvad K, et al. Cross-sectional associations of objectively measured physical activity, cardiorespiratory fitness and anthropometry in European adults. *Obesity* 2014;22(5):E127–34. [PubMed: 23804303]
- [31]. Beydoun MA, Kaufman JS, Sloane PD, Heiss G, Ibrahim J. n-3 Fatty acids, hypertension and risk of cognitive decline among older adults in the Atherosclerosis Risk in Communities (ARIC) study. *Public Health Nutr* 2008;11(1):17–29. [PubMed: 17625029]
- [32]. Nusser SM, Fuller WA, Guenther PM. Estimation of usual dietary intake distributions: Adjusting for measurement error and non-normality in 24-hour food intake data In: Trewin D, editor. *Survey Measurement and Process Quality*. New York: Wiley; 1996 p. 689–709.
- [33]. Dekkers AL, Verkaik-Kloosterman J, van Rossum CT, Ocké MC. SPADE, a New Statistical Program to Estimate Habitual Dietary Intake from Multiple Food Sources and Dietary Supplements. *J Nutr* 2014;144:2083–91. [PubMed: 25320187]
- [34]. Zhang S, Midthune D, Guenther PM, Krebs-Smith SM, Kipnis V, Dodd KW, et al. A New Multivariate Measurement Error Model with Zero-Inflated Dietary Data, and its Application to Dietary Assessment. *Ann Appl Stat* 2011;5:1456–87. [PubMed: 21804910]
- [35]. Mansournia MA, Danaei G, Forouzanfar MH, Mahmoodi M, Jamali M, Mansournia N, et al. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis: analysis with marginal structural models. *Epidemiology* 2012;23(4):631–40. [PubMed: 22415107]
- [36]. Gustafson P, Le ND. Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics* 2002;58(4):878–87. [PubMed: 12495142]
- [37]. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. *J Natl Cancer Inst* 2011;103:1086–92. [PubMed: 21653922]

Table 1

Literature search terms for the four topic areas: dietary intake cohort studies, dietary intake population survey, physical activity cohort study, and air pollution cohort studies

Search	Search terms
Dietary intake cohort	
Search A	(((((cardiovascular disease[Title]) OR cancer[Title]) OR diabetes[Title])) AND (((risk[Title]) OR risks[Title]) OR association[Title])) AND (((diet[Title]) OR consumption[Title]) OR intake[Title]) OR dietary[Title])) AND ("2012/01/01"[Date-Publication]: "3000"[Date-Publication])) NOT case-control[Title/Abstract]) NOT review[Title/Abstract]) NOT meta-analysis[Title/Abstract]) NOT cross-sectional[Title/Abstract]) AND cohort))
Search B1	[Search A terms with date range extended to 2001/01/01] AND measurement error
Search B2	(measurement error OR misclassification) AND nutritional epidemiology
Search B3	(((((measurement error[Title/Abstract] OR misclassification[Title/Abstract] OR reliability[Title/Abstract])) AND ((cardiovascular disease OR cancer OR diabetes))) AND ((risk OR association OR risks))) AND (diet[Title/Abstract] OR dietary[Title/Abstract] OR consumption[Title/Abstract] OR intake[Title/Abstract] OR intakes[Title/Abstract])) AND (cohort[Title/Abstract] OR men[Title/Abstract] OR women[Title/Abstract])) NOT meta-analysis) NOT case-control) NOT review
Dietary intake population survey	
Search A	survey AND (nutrition OR nutritional OR diet OR dietary OR food OR nutrient) AND (FFQ OR FPQ OR record OR recall OR diary OR semi-quantitative OR semiquantitative).
Search B	N/A
Physical activity cohort	
Search A	survey AND (nutrition OR nutritional OR diet OR dietary OR food OR nutrient) AND (FFQ OR FPQ OR record OR recall OR diary OR semi-quantitative OR semiquantitative)
Search A	(((((“Cohort Studies”[Mesh:noexp]) OR “Follow-Up Studies”[Mesh]) OR “Longitudinal Studies”[Mesh]) OR “Prospective Studies”[Mesh] OR cohort OR prospective study)) AND (exercise OR recreation OR physical activity OR sedentary OR energy expenditure))
Search B	[Search A terms] AND (“measurement error” OR misreport* OR misclassif* OR bias OR attenuat* OR calibrat*)
Air pollution cohort	
Search A	((“cardiovascular disease” OR cancer OR mortality OR “hospital

Search	Search terms
	admissions" OR "respiratory disease" OR diabetes OR biomarker OR physiol* OR myocard*) AND (risk* OR association* OR effect* OR impact) AND ("air pollution" OR particulate* OR "environmental quality" OR "air quality")) NOT (case-control OR review OR meta-analysis OR cross-sectional).
Search B1	((health OR "cardiovascular disease" OR cancer OR mortality OR "hospital admissions" OR "respiratory disease" OR diabetes OR biomarker OR physiol* OR myocard*) AND (risk* OR association* OR effect* OR impact) AND <("air pollution" OR particulate* OR "environmental quality" OR "air quality")) NOT (case-control OR review OR meta-analysis OR cross-sectional) AND ("measurement error" OR misreport* OR misclassif* OR bias OR attenuat*)
Search B2	"measurement error" AND "air pollution"

Table 2

Study summary from general literature search without measurement error added to search terms (search A)

Question	Dietary intake cohort N = 51	Physical activity cohort N = 30	Dietary intake population survey N = 67	Air pollution cohort N = 50
Mention ME as potential problem N (%)	48 (94%)	17 (57%)	53/67 (79%)	23 (46%)
Included reliability substudy	1	0 (0%)		1 (2%)
Included calibration substudy	2	0 (0%)		1 (2%)
Used a method to adjust for ME N (%)	5 (10%)	0 (0%)	19/67 (28%)	4 (8%)
Measurement error method [*]	Reg. Cal 2/5 (40%) Cum. Avg 3/5 (60%)	—	NCI 10/19 (53%) Means 7/19 (37%) ISU 1/19(5%) MSM 1/19(5%)	Reg. Cal 1/4 (25%) Sens. Anal 3/4 (75%)
Categorized exposure [†]	Any 50/51 (98%) Exclusively 27/51 (53%)	Primary Exposure 21/30 (70%)		
Statistic(s) of main interest [‡] N (%)	HR 46 (90%) OR 3 (6%) RR 2 (4%) Slope 4 (8%)	HR 11 (37%) OR/RR 9 (30%) LM 5 (17%) Other 5 (17%)	Mean 51 (76%) Median 28 (42%) Percentiles 21 (31%) Quality 31 (46%)	

Some questions were not assessed by all surveys.

^{*} Cum. Avg: cumulative average; ISU: Iowa State University Method [23]; MSM: Multiple Source Method [24,25]; NCI: National Cancer Institute Method [26,27]; Reg Cal: Regression calibration [1,28]; Sens Anal: Sensitivity analysis.

[†] Articles were categorized as to whether they had categorized at least one dietary intake exposure in the statistical analysis (any) and whether all analyses were done with categorized intakes (exclusively).

[‡] LM: Linear model or linear mixed model; HR: hazard ratio; ME: measurement error; OR: odds ratio; RR: relative risk.

Table 3

Study summary from the literature search with measurement error added to search terms (search B) *

Question	Dietary intake cohort N = 27	Physical activity cohort N = 40	Air pollution cohort N = 25
Mention ME as potential problem N (%)	27 (100%)	28 (70%)	21 (84%)
Included reliability substudy	6 (22%)	3 (8%)	0 (0%)
Included calibration substudy	21 (78%)	4 (10%)	1 (4%)
Used a method to adjust for ME N (%)	27 (100%)	2 (5%)	5 (20%)
Measurement error method	Reg Cal 26 (96%) [‡] SIMEX 1 (4%) Other 1 (4%)	Reg Cal 2 [‡] (100%)	Sens Anal 4 (80%) Instr Var 1 (20%)
Categorized exposure [§]	Any 19/27 Exclusively 2/27	31 (78%)	
Statistic(s) of main interest	HR 19 (70%)	HR 15 (38%)	
N (%)	OR 6 (22%) RR 0 (0%) Slope 3 (11%)	OR/RR 9 (22%) LM 10 (25%) Other 6 (15%)	

Some questions were not assessed by all surveys.

* Instr Var: Instrumental variables; ISU: Iowa State University Method [23]; MSM: Multiple Source Method [24,25]; NCI: National Cancer Institute Method [26,27]; Reg Cal: Regression calibration [1,28]; Sens Anal: Sensitivity analysis; SIMEX: Simulation Extrapolation Method [1,29].

[‡] One article for the dietary intake cohort used both SIMEX and regression calibration so percentages do not add up to 100.

[‡] One study did not use the term regression calibration but applied an equivalent method (i.e., beta coefficient adjustment for the intraclass correlation coefficient) [30].

[§] Articles were categorized as to whether they had categorized at least one dietary intake exposure in the statistical analysis (any) and whether all analyses were done with categorized intakes (exclusively).

^{||} LM: Linear model or linear mixed model; HR: hazard ratio; ME: measurement error; OR: odds ratio; RR: relative risk.

Table 4

Characteristics of articles reviewed for the air pollution search A survey (N = 50)

Main outcome	N (%)	Temporal resolution	N (%)	Type of measurement	N (%)
Mortality	17 (34%)	Minutely	1 (2%)		
Hospital admissions	12 (24%)	Hourly	9 (18%)		
Cardiovascular disease	2 (4%)	Between daily and hourly	3 (6%)		
Cancer	1 (2%)	Daily	21 (42%)	Fixed site	35 (70%)
Respiratory disease	6 (12%)	Weekly	1 (2%)	Personal	5 (10%)
Diabetes	1 (2%)	Yearly	3 (6%)	Estimated exposure	12 (24%)
Physiological parameters	5 (10%)	Study period	4 (8%)		
Biomarker	2 (4%)	Other	7 (14%)		
Others	6 (12%)				