Statistical Analysis—Measurement Error

Timo B. Brakenhoff, Maarten van Smeden and Daniel L. Oberski

Abstract

An important aspect of data quality when conducting clinical analyses using real-world data is how variables in the data have been recorded or measured. The discrepancy between an observed value and the *true* value is called measurement error (also known as *noise* in the artificial intelligence and machine learning literature) and can have consequences for your analyses in all kinds of contexts. To properly assess the potential impact of measurement error it is essential to understand the relationship between the true and observed variables as well as the goal of the analysis and how it will be implemented in practice. Commonly, measurement error is distinguished as being classical, Berkson, systematic and/or differential. While it is clear that measurement error can have far-reaching consequences on analyses, the effect can differ depending on whether analyses are descriptive, explanatory or predictive. Validation studies can inform the estimation and characterization of measurement error as well as provide crucial information for correction methods that are available in several statistical programming languages such as SAS, R and Python.

Keywords

Measurement error · Misclassification · Noise · Correction · Bias · Modelling · Estimation

1 Introduction

Before applying an analytical method on data it is important to consider the quality of the data and how that quality might impact the results of the analysis. One important aspect of data quality is how variables in the data have been recorded or measured. There are many different situations in which the variable(s) that are measured or observed are different from what was intended to be measured. This discrepancy between an observed value and the *true* value is called **measurement error** and can have consequences for your analyses in all kinds of contexts (see Box 1 for two examples of the effect of measurement error in practice).

T. B. Brakenhoff (✉)
Julius Clinical, Zeist, The Netherlands
e-mail: timo.brakenhoff@juliusclinical.com

M. van Smeden · D. L. Oberski
Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, The Netherlands

D. L. Oberski
Dept. Methodology & Statistics, Utrecht University, Utrecht, The Netherlands

F. W. Asselbergs et al. (eds.), *Clinical Applications of Artificial Intelligence in Real-World Data*,
https://doi.org/10.1007/978-3-031-36678-9_6

Box 1: Examples of Measurement Error in Practice

- Measuring prevalence using different diagnostic tests
  - In Montreal, Canada a screening and treatment program for intestinal parasite infections was offered to newly arrived Southeast Asian refugees in Canada between July 1982 and February 1983. The 162 Cambodian refugees included in the sample were tested using two different diagnostic tests for the presence of Strongyloides Infection: enzyme-linked immunosorbent assay (immunoglobulin G) *serology* and *stool* examination (see table below for the amount of refugees that tested positive using each diagnostic test) [27, 28]. The observed sample prevalence based solely on serology was 77.2 percent, while it was 24.7 percent using information from stool examinations alone! This absolute difference of over 50 percentage points in prevalence demonstrates how crucial it is to consider the instrument that is being used to measure a quantity of interest, such as the prevalence. Note that these estimates also don't take into account other sources of uncertainty such as sampling variability (only 162 individuals of the whole population of Cambodian refugees were included in this sample) or the performance of the tests themselves (it is likely that several individuals may be false positives or false negatives as neither test has perfect sensitivity or specificity) [34].

	Stool +	Stool −	
Serology +	**38**	**87**	125
Serology −	**2**	**35**	37
	40	122	162

- Computer aided diagnosis of prostate cancer without gold standard outcome labels
  - Nir et al. [51] describe the automatic grading of prostate cancer in digitized histopathology images. They did this using various supervised machine and deep learning methods based on images labeled by pathologists. Just as in many medical image settings, this labeling is not perfect and specialists will not always agree when evaluating the same images. When these images act as important input for machine and deep learning algorithms meant for diagnostic or prognostic settings, this, often unavoidable, measurement error, or noise in the outcome labels can have significant consequences for the performance of the algorithms [35]. In the case of [51] multiple pathologists were asked to rate the same images and different methods were used to best account for the inter-observer variability in prostate cancer grading. While this may not always be possible to apply in practice, there are several other techniques that can help correct for measurement error in the outcome [35].

Where the term "measurement error" is frequently used with regards to errors in the measurement of continuous variables (such as an individual's age or height), the term "misclassification" is often used for discrete variables (such as an individual's preferences of received treatment). In Artificial intelligence and machine learning literature, errors in discrete or non-discrete variables are often called *noise* with noise existing either in the covariates (also known as predictors, features or attributes) or in the outcome(s) (also known as target variables, labels or classes). In this chapter, the term measurement error will be used to describe all these phenomena unless otherwise specified.

Errors in measurement can be caused through various mechanisms including, but not limited to, inaccuracy and imprecision of measurement instruments, errors due to self-reporting, errors in data coding or labeling, lack of data granularity, or when single measurements are taken of naturally fluctuating biological processes such as biomarkers. Common settings where such errors can occur include when measuring smoking [45], blood pressure [2, 53, 75], dietary intake [17, 18, 73], physical activity [16, 41], exposure to air pollutants [22, 69, 78], medical treatments received [5, 65, 71], diagnostic coding [15, 52, 77] and labels for medical images [12, 35, 55, 57].

All of the above mentioned measurement error mechanisms can lead to discrepancies between the sought after, perfectly measured and thus error-free *true* value of a variable and an imperfectly measured *observed* value of that same variable. In most cases we have not observed the former and we are in possession of the latter. This can have severe implications for the results of an analysis. Examples include the following:

- Brakenhoff et al. [7] demonstrate that even when the simplest form of measurement error, random error, is assumed when measuring blood pressure in routine care, this can have very divergent and unexpected consequences on the estimation of the effect of blood pressure on the possible risk of developing cardiovascular disease. The estimated relations can be severely biased positively or negatively depending on the amount of measurement error present in confounders and the relationship of those confounders with the observed blood pressure variable.
- When aiming for the best possible prediction performance using advanced artificial intelligence techniques such as deep learning for medical imaging, multiple authors [12, 35, 57] identify the need for large datasets of trustworthy labelled medical images (which are used as the outcome to be predicted) to train the desired model. The expertise required for this as well as regulations in the medical sector make this a challenging ask which can severely impact the performance of prediction models.

To properly assess the potential impact of measurement error it is essential to understand the relationship between the true and observed variables as well as the goal of the analysis (i.e. is the purpose to *describe*, *explain* or *predict?*) (See Box 3) and how it will be implemented in practice. However, the fact that measurement error may have far-reaching consequences on analyses in the field of statistics, epidemiology or artificial intelligence is nothing new [9, 26, 79]. Yet, despite this understanding and a plethora of recent literature on the subject [8, 36] there is still little attention paid to measurement error consequences and potential solutions in the medical literature [6, 67] and common myths [7, 74] are perpetuated. With the increasing availability of (big) data not collected for research purposes such as medical health records for explanation as well as the application of machine learning and deep learning algorithms for prediction, careful investigation of potential bias due to issues like measurement error is arguably more important than ever [21].

This chapter will provide an overview of the types of measurement error and why it is essential to keep this in consideration when conducting clinical data analysis. Subsequently the consequences of measurement error will be discussed and how this will differ depending on the goal of the analysis and the desired implementation. Lastly, an overview will be given of various tools for the estimation and correction of measurement error.

2 Types of Measurement Error

A common taxonomy to distinguish between types of measurement error differentiates between 4 types: classical, Berkson, systematic and differential. Each of these types can

manifest differently in continuous or discrete data. They represent different ways in which true values and the observed variables relate to each other, which can have different consequences on the analysis being performed.

When considering ***continuous variables***, we can differentiate between multiple *measurement error models*. The simplest of these is called the *classical or random measurement error model* where the observed variable is equal to the true variable plus error, in this case a random variable with mean 0 which is independent of the true variable. This error model can be extended to accommodate *systematic error* or dependencies between the error and the observed variable, the true variable or other auxiliary variables. When the relations between the observed and true variable are non-linear, transformations can be used to make it linear. In specific circumstances it is more appropriate to model the true variable as equal to the observed variable plus a random variable with mean 0 which is independent of the observed variable. This is called *Berkson error.* Lastly, depending on if the error contains information on the outcome variable which you may be interested in or not, the error is referred to as *differential* or *nondifferential* respectively. Box 2 provides technical definitions of these measurement error models.

For ***categorical variables***, discrepancies between the true value of a variable and the observed value is often referred to as misclassification. While misclassification is closely related to measurement error in continuous variables, the categorical nature of the variables means that misclassification is often expressed in terms of mis*classification probabilities*. For example, in the case of a binary observed and true variable, regardless of the type of measurement error assumed, misclassification can best be described in terms of sensitivity, specificity and predictive values (namely positive predictive value and negative predictive value). Note that similar to measurement error models, misclassification can also be (non)differential and have a structure similar to Berkson error (while the latter is not often observed) [36].

Box 2: Technical Definitions of Types of Measurement Error in Continuous Variables

Suppose we are interested in the relationship between an outcome variable $\boldsymbol{Y}$ and a covariate of interest $\mathbf{X}$ given covariates $\mathbf{Z}$. If a variable $\mathbf{X}$ is measured with error, the observed variable is denoted by $\mathbf{X^*}$, with the true value of this variable ($\mathbf{X}$) being unobserved. Note that notation differs across the literature and the notation chosen here is consistent with that of [36 and 68]. The following types of error are most commonly distinguished:

- **Classical measurement error:**
  $X^* = X + U$, where U is a random variable with mean 0 that is independent of X.
- **Linear measurement error**
  $X^* = \alpha_0 + \alpha_X X + U$, where U is a random variable with mean 0 that is independent of X, α_0 is an intercept term and α_X is the coefficient of X. Note that classical measurement error is a special case of linear measurement error where $\alpha_0 = 0$ and $\alpha_X = 1$.
- **Systematic error**
  $X^* = \alpha_0 + \alpha_X X$, where α_0 is an intercept term and α_X is the coefficient of X which each represent systematic error that may be dependent on X.
- **Nondifferential error**
  The distribution of Y given (X, Z, X*) depends only on (X, Z)
- **Berkson measurement error**
  $X = X^* + U$, where U is a random variable with mean 0 that is independent of X*.

3 Consequences of Measurement Error

3.1 Goal of the Analysis

Before discussing the consequences of measurement error it is important to clearly identify the goal of the analysis. A common framework

used to distinguish between the goal of statistical modeling is whether it is used for **description**, **explanation** or **prediction** [70] (See Box 3). Shmueli [70] mostly disregards descriptive modelling as it is frequently used for characterization of the observed data structure and is not often used for theory building. In public health and healthcare research, however, descriptive modelling plays a crucial role, e.g. when estimating incidence rates or prevalences of disease. In the context of measurement error and its impact, this section will mostly focus on the distinction between explanatory and predictive modelling.

Box 3: Definitions of Types of Statistical Modelling

- **Descriptive modelling** is aimed at summarizing or representing the data. E.g. calculating an incidence rate for a disease over a particular time period, or by fitting a regression model to quantify the association between a covariate and an outcome, without causal inference or prediction intentions.
- **Explanatory modelling** is the application of models to data for the purpose of testing and quantifying causal relations. E.g. fitting a regression model to estimate the causal effect of a certain factor (e..g. a medical treatment, registered as a dispensed drug) on the occurrence of a certain outcome (e.g. a health outcome such as (cause-specific) mortality or hospital admission).
- **Predictive modelling** the application of models to data for the main purpose of predicting new or future observations. E.g. fitting a regression model to predict the probability of the occurrence of a certain health outcome (e.g. 5-year mortality) for future individuals taking into account various relevant covariates (e.g. medical history, demographics, laboratory tests, etcetera).

While often not clearly separated in literature, studies with explanation and prediction goals fundamentally differ due to the differences in aims and subsequent diverging choices at every step of the modelling process (designing the study, collecting data, preparing data, exploring data, selecting variables, selecting statistical models, evaluating models and using models in practice). Note that both types of modelling can be used in combination, each achieving a separate specific goal within an overarching analysis that may be of an explanatory or predictive nature. An example of this is the application of prediction models (including machine learning models [44]) to estimate propensity scores [58] that are used to adjust for confounding when estimating causal effects.

The measurement of variables for explanatory modelling generally focuses on obtaining measurements that are as reliable and accurate as possible to appropriately represent the underlying constructs. Conversely, for many predictive modelling studies priority goes towards reliably estimating the outcome/target variable (often called *labeling* [1, 19, 49, 50]), while the measurement quality of the covariates necessary for making predictions should ideally be of a similar quality when the model is constructed as when the model is applied to new patients. So far, however, much of the attention in the measurement error literature [9, 37] has been specifically devoted to explanatory modelling. More recently, attention is being given to the prediction setting, showing the impact of *heterogeneity* in how variables are measured in the training and implementation settings, also referred to as *transportability* [9], and how this impacts the performance of prediction models [42, 43, 54].

The above broad differentiation in modeling goals and the different role of errors in measurement exemplifies the importance of keeping in mind the goal of the analysis, how the results of the analysis will be generalized and in which settings the results will be applied.

3.2 The Impact of Measurement Error in Explanatory Modelling

Much of the health science measurement error literature has been focussed on the consequences of different types of measurement error when engaging in explanatory modelling. Carroll et al. [9], describe how the consequences of measurement error is a "triple whammy": covariate-outcome relationships can be biased, power to detect clinically meaningful relationships is diminished and important features of the data can be masked.

When assuming classical measurement error or misclassification in a single continuous or binary categorical covariate of interest, the estimated univariable covariate-outcome relation will be biased towards the null (also known as *attenuation*). However, when the covariate has more than two categories or when considering a multivariable model (models with more than one covariate) where at least 1 confounder measured with classical error, the estimated covariate-outcome relation can be biased in either direction, even if the covariate of interest is not measured with error [7]. This unpredictability of the magnitude and direction of bias and precision on the estimated effect is compounded if error is systematic or differential. Berkson error on the other hand often does not lead to bias in the estimated covariate-outcome relation, but can diminish precision. Regarding measurement error in the outcome of an explanatory model, classical error will generally not lead to bias in a covariate-outcome relation while other types of error like systematic or differential error can substantially bias estimators [46]. Table 1 of [37] provides a useful overview of the effects of measurement error according to the type of error and target of the analysis for explanatory modelling.

3.3 The Impact of Measurement Error in Predictive Modelling

Attention for the role of measurement error in predictive modelling is relatively recent. In particular, the concept of *measurement heterogeneity,* which means the covariates (predictors) are measured differently (i.e. have different measurement error) between training and external validation settings for prediction models, has been shown to have an important impact on the performance of prediction models. Measurement heterogeneity can, for instance, occur when different measurement protocols or different types of tests are used when developing a clinical prediction model as compared to the setting in which they are externally validated or applied. Various studies [42, 43, 54] have shown how in different measurement scenarios often leads to deteriorated performance of the calibration and discrimination of prediction models.

Regarding the impact of measurement error or noise in the development of machine learning or deep learning models, attribute (i.e. covariate) noise is often considered to have a less severe impact on predictive performance than label (i.e. outcome) noise [25, 66]. Label noise can diminish accuracy of predictions and classification performance as well as increase the amount of training samples required for model development [19, 50]. In addition, error prone outcomes can lead to prediction unfairness if the error differs over subgroups of interest [4]. For an overview of the impact of class and attribute noise, see [79].

Box 4: Five Myths About Measurement Error

van Smeden et al. [74] identifies and debunks 5 common myths about measurement error:

1. Measurement error can be compensated for by large numbers of observations
   a. No, a large number of observations does not resolve the most serious consequences of measurement error in epidemiological data analyses. These remain regardless of the sample size.
2. The effect of a covariate of interest on the outcome is underestimated when variables are measured with error

   a. No, the effect of a covariate of interest can be over- or underestimated in the presence of measurement error depending on which variables are affected, how measurement error is structured and the expression of other biasing and data sampling factors.
3. Covariate measurement error is non-differential if measurements are taken without knowledge of the outcome
   a. No, covariate measurement error can be differential even if the measurement is taken without knowledge of the outcome.
4. Measurement error can be prevented but not mitigated in data analyses
   a. No, statistical methods for measurement error bias corrections can be used in the presence of measurement error provided that data are available on the structure and magnitude of measurement error from an internal or external source. This often requires planning of a measurement error correction approach or quantitative bias analysis, which may require additional data to be collected.
5. Certain types of research are unaffected by measurement error
   a. No, measurement error can affect all types of research.

4 Correction of Measurement Error

Several approaches have been suggested to circumvent (or at least lower) the detrimental consequences of measurement error, in particular to reduce bias (one of the 3 whammies of measurement error). To understand the possible value of correction, the natural first step is in identifying potential error-prone variables. To quantify and correct for measurement error, additional information is required which can often be collected through validation studies.

4.1 Validation Studies

Validation studies (also referred to as ancillary studies) on the error-prone variables can aid the investigation into the structure, type and amount of measurement error present [37]. These studies can also be essential for the application of several correction methods discussed later in this section. Generally speaking, there are four types of validation studies: internal validation studies, calibration studies, replicates studies and external validation studies.

In an **internal validation study**, both the error-prone observed variable as well as (a reliable representation of) the true variable (i.e. gold standard measurement) are observed in a subset of the data. Measurement of a gold standard only in a subset can be motivated by a measurement procedure that is time-consuming, expensive, invasive or even impossible to obtain for the whole study sample. Usually an internal validation study is assumed to contain data from a *random* subset of the study sample, but alternative sampling strategies are available depending on the type of measurement error and the measurement error correction method that can be used [47]. With a suitable internal validation study, the relations between the error-prone observed variable and the true variable can directly be estimated, which can be used for measurement error correction. If the true variable or gold standard measurement is not available, but another variable (reference measurement) unbiased at the individual level is, it is sometimes called a **calibration study**. This type of study can be used as input for the measurement error correction method called regression calibration, if certain assumptions are met.

In a **replicates study,** multiple replicate measurements from the same instrument (e.g. multiple measurements of blood pressure during the same hospital visit) or different instruments that measure the same underlying construct

(e.g. multiple diagnostic tests for the same disease) are collected. When the variable of interest contains random measurement error, having multiple measurements available can provide essential information on the amount and type of measurement error present.

Validation studies can also use data available from external sources such as similar cohorts from another country. For example, for separate individuals not included in the main study, both the error-prone variable as well as the true variable (or gold standard measurement) and necessary covariates might be available. This can then be used to inform measurement error correction methods. Note that for such **external validation studies** it is very important to assess the heterogeneity between the external and internal setting and how transportable the information is. More information on the design and desirable size of validation studies can be found in [37].

4.2 Correction Methods

Characterizing the amount and type of error is an important first step when applying strategies to correct for the measurement error. At the most basic level, common metrics such as the bias and variance or classification probabilities like sensitivity and specificity can be used to characterize how accurate and precise observed variables are compared to the true variables. The next step is to identify the type of measurement error observed (see Sect. 2) and use those models to further quantify various aspects of the error. In general, measurement error correction methods use information obtained through validation studies to take into account measurement error in the analyses by estimating the research results in the counterfactual situation where there was no measurement error.

Many different approaches have been proposed in the literature which characterize the error present as well as correct for the bias that may arise due to this error in the final analyses. Approaches include: regression calibration [11], simulation extrapolation [14, 37], likelihood methods [10], score function methods [3, 72], methods-of-moment correction [20], latent variable analysis [32], structural equation modelling [4, 63], multiple imputation for measurement error correction [13], inverse probability weighting [23], bayesian analyses [26], cluster-based correction [49].

More detailed information on the various types of error and how to correct for them can be found in extensive literature on the topic. Various measurement error text books exist, with [9] focussing on nonlinear models, [26] on Bayesian methods of adjustment and [8]) providing a more broad overview. Similarly, reviews such as the one by Guolo [24] give an overview of robust techniques to correct for measurement error in covariates. More recently, the STRATOS initiative wrote a two-part tutorial on the basic theory of measurement error and simple methods of adjustment [36] as well as on more complex methods of adjustment and advanced topics [68]. Literature focused on the impact of measurement error (referred to as noise) in both covariates and outcomes in the field of machine learning and how to deal with it includes [19, 50, 64, 79].

While several methods can be easily programmed using standard functionality of different software tools, specific packages, macros or procedures are available for more complex measurement error correction in different programming languages. In SAS, for example, macros include *%blinplus* [59], *%relibpls8* [60] and *%rrc* [40] which have been developed for various implementations of regression calibration. Similarly in STATA, procedures include *rcal* and *eivreg* for regression calibration [29], and *simex* and *simexplot* for simulation extrapolation [30]. For the R language, packages include *simex* [39] and *simexaft* [31] for simulation extrapolation approaches, *lavaan* [61] for latent variable analysis and structural equation modelling, as well as *mecor* [48] for measurement error correction in linear regression models. Also in Python, an increasing amount of relevant packages are being developed, such as *pyEMU* [76] for environmental model uncertainty analysis and *snorkel* [56] for rapid training data creation in the face of potential label noise.

An important alternative method to investigate the impact of measurement error on your study results if no suitable additional information is available, is to perform sensitivity analyses. Various amounts of measurement error can be assumed in hypothetical scenarios where the analysis is rerun and the results are compared against the original results. To assess multiple hypothetical scenarios with various amounts of measurement error simultaneously, probabilistic sensitivity analyses can be performed (see Chapter 19 of [62]). A similar technique applied to examine the impact of measurement error (and correct for it) when additional information is lacking in both explanatory and prediction modelling is quantitative bias analysis [33, 38].

References

1. Algan G, Ulusoy I. Label noise types and their effects on deep learning. 2020. ArXiv: https://arxiv.org/abs/2003.10471
2. Bauldry S, Bollen KA, Adair LS. Evaluating measurement error in readings of blood pressure for adolescents and young adults. Blood Press. 2015;24:96–102. https://doi.org/10.3109/08037051.2014.986952.
3. Boeschoten L, Oberski D, De Waal T. Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). J Off Stat. 2017;33:921–62. https://doi.org/10.1515/jos-2017-0044.
4. Boeschoten L, van Kesteren E-J, Bagheri A, Oberski DL. Achieving fair inference using error-prone outcomes. Int J Interact Multimed Artif Intell. 2021;6:9. https://doi.org/10.9781/ijimai.2021.02.007.
5. Boudreau DM, Daling JR, Malone KE, et al. A validation study of patient interview data and pharmacy records for antihypertensive, statin, and antidepressant medication use among older women. Am J Epidemiol. 2004;159:308–17. https://doi.org/10.1093/aje/kwh038.
6. Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol. 2018;98:89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023.
7. Brakenhoff TB, van Smeden M, Visseren FLJ, Groenwold RHH. Random measurement error: why worry? An example of cardiovascular risk factors. PLoS ONE. 2018;13: e0192298. https://doi.org/10.1371/journal.pone.0192298.
8. Buonaccorsi JP. Measurement error: models, methods, and applications. New York: Chapman and Hall/CRC; 2010.
9. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. 2nd ed. New York: Chapman and Hall/CRC; 2006.
10. Carroll RJ, Spiegelman CH, Lan KKG, et al. On errors-in-variables for binary regression models. Biometrika. 1984;71:19–25. https://doi.org/10.1093/biomet/71.1.19.
11. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. J Am Stat Assoc. 1990;85:652–63. https://doi.org/10.1080/01621459.1990.10474925.
12. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15:20170387. https://doi.org/10.1098/rsif.2017.0387.
13. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006;35:1074–81. https://doi.org/10.1093/ije/dyl097.
14. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc. 1994;89:1314–28. https://doi.org/10.1080/01621459.1994.10476871.
15. Delate T, Jones AE, Clark NP, Witt DM. Assessment of the coding accuracy of warfarin-related bleeding events. Thromb Res. 2017;159:86–90. https://doi.org/10.1016/j.thromres.2017.10.004.
16. Ferrari P, Friedenreich C, Matthews CE. The role of measurement error in estimating levels of physical activity. Am J Epidemiol. 2007;166:832–40. https://doi.org/10.1093/aje/kwm148.
17. Freedman LS, Commins JM, Willett W, et al. Evaluation of the 24-hour recall as a reference instrument for calibrating other self-report instruments in nutritional cohort studies: evidence from the validation studies pooling project. Am J Epidemiol. 2017;186:73–82. https://doi.org/10.1093/aje/kwx039.
18. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. JNCI J Natl Cancer Inst. 2011;103:1086–92. https://doi.org/10.1093/jnci/djr189.
19. Frenay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst. 2014;25:845–69. https://doi.org/10.1109/TNNLS.2013.2292894.
20. Fuller WA. Measurement error models. New York: John Wiley & Sons; 1987.
21. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178:1544. https://doi.org/10.1001/jamainternmed.2018.3763.
22. Goldman GT, Mulholland JA, Russell AG, et al. Impact of exposure measurement error in air

pollution epidemiology: effect of error type in time-series studies. Environ Health. 2011;10:61. https://doi.org/10.1186/1476-069X-10-61.
23. Gravel CA, Platt RW. Weighted estimation for confounded binary outcomes subject to misclassification. Stat Med. 2018;37:425–36. https://doi.org/10.1002/sim.7522.
24. Guolo A. Robust techniques for measurement error correction: a review. Stat Methods Med Res. 2008;17:555–80. https://doi.org/10.1177/0962280207081318.
25. Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: a systematic review. Procedia Comput Sci. 2019;161:466–74. https://doi.org/10.1016/j.procs.2019.11.146.
26. Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and bayesian adjustments. CRC Press (2003)
27. Gyorkos TW, Frappier-Davignon L, Dick Maclean J, Viens P. Effect of screening and treatment on imported intestinal parasite infections: results from a randomized, Controlled Trial. Am J Epidemiol. 1989;129:753–61. https://doi.org/10.1093/oxfordjournals.aje.a115190
28. Gyorkos TW, Genta RM, Viens P, Maclean JD. Seroepidemiology of Strongyloides infection in the Southeast Asian refugee population in. Canada. Am. J. Epidemiol. 1990;257–64
29. Hardin JW, Schmiediche H, Carroll RJ. The regression-calibration method for fitting generalized linear models with additive measurement error. Stata J Promot Commun Stat Stata. 2003;3:361–72. https://doi.org/10.1177/1536867X0400300406.
30. Hardin JW, Schmiediche H, Carroll RJ. The simulation extrapolation method for fitting generalized linear models with additive measurement error. Stata J Promot Commun Stat Stata. 2003;3:373–85. https://doi.org/10.1177/1536867X0400300407.
31. He W, Xiong J, Yi GY, SIMEX R package for accelerated failure time models with covariate measurement error. J Stat Softw. 2012;46:1–14. https://doi.org/10.18637/jss.v046.c01
32. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980;36:167–71. https://doi.org/10.2307/2530508.
33. Jiang T, Gradus JL, Lash TL, Fox MP. Addressing measurement error in random forests using quantitative bias analysis. Am J Epidemiol. 2021. https://doi.org/10.1093/aje/kwab010.
34. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141:263–72. https://doi.org/10.1093/oxfordjournals.aje.a117428.
35. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal. 2020;65: 101759. https://doi.org/10.1016/j.media.2020.101759.
36. Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment. Stat Med. 2020;39:2197–231. https://doi.org/10.1002/sim.8532.
37. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8531
38. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43:1969–85. https://doi.org/10.1093/ije/dyu149.
39. Lederer W, Küchenhoff H. A short introduction to the SIMEX and MCSIMEX. Newsl R Proj. 2006;6(4):26–31.
40. Liao X, Zucker DM, Li Y, Spiegelman D. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. Biometrics. 2011;67:50–8. https://doi.org/10.1111/j.1541-0420.2010.01423.x.
41. Lim S, Wyker B, Bartley K, Eisenhower D. Measurement error of self-reported physical activity levels in New York City: assessment and correction. Am J Epidemiol. 2015;181:648–55. https://doi.org/10.1093/aje/kwu470.
42. Luijken K, Groenwold RHH, Calster BV, et al. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med. 2019;38:3444–59. https://doi.org/10.1002/sim.8183.
43. Luijken K, Wynants L, van Smeden M, et al. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. J Clin Epidemiol. 2020;119:7–18. https://doi.org/10.1016/j.jclinepi.2019.11.001.
44. McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013;32:3388–414. https://doi.org/10.1002/sim.5753.
45. Murray RP, Connett JE, Lauger GG, Voelker HT. Error in smoking measures: effects of intervention on relations of cotinine and carbon monoxide to self-reported smoking. The Lung Health Study Research Group. Am J Public Health. 1993;83:1251–7. https://doi.org/10.2105/AJPH.83.9.1251.
46. Nab L, Groenwold RHH., Welsing PMJ, van Smeden M. Measurement error in continuous endpoints in randomised trials: problems and solutions. Stat Med. 2019;38:5182–96. https://doi.org/10.1002/sim.8359.
47. Nab L, van Smeden M, de Mutsert R, et al. Sampling strategies for internal validation samples for exposure measurement error correction: a study of visceral adipose tissue measures replaced by waist circumference measures. Am J Epidemiol Kwab. 2021a;114. https://doi.org/10.1093/aje/kwab114
48. Nab L, van Smeden M, Keogh RH, Groenwold RHH. mecor: An R package for measurement error correction in linear regression models with a continuous outcome. Comput Methods Programs Biomed. 2021b;208:

49. Nicholson B, Sheng VS, Zhang J. Label noise correction and application in crowdsourcing. Expert Syst Appl. 2016;66:149–62. https://doi.org/10.1016/j.eswa.2016.09.003.
50. Nigam N, Dutta T, Gupta HP. Impact of noisy labels in learning techniques: a survey. In: Kolhe ML, Tiwari S, Trivedi MC, Mishra KK, editors. Advances in Data and Information Sciences. Singapore: Springer; 2020. p. 403–11.
51. Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. Med Image Anal. 2018;50:167–80. https://doi.org/10.1016/j.media.2018.09.005.
52. Nissen F, Morales DR, Mullerova H, et al. Validation of asthma recording in the clinical practice research datalink (CPRD). BMJ Open. 2017;7: e017474. https://doi.org/10.1136/bmjopen-2017-017474.
53. Nitzan M, Slotki I, Shavit L. More accurate systolic blood pressure measurement is required for improved hypertension management: a perspective. Med Devices Auckl NZ. 2017;10:157–63. https://doi.org/10.2147/MDER.S141599.
54. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. J Clin Epidemiol. 2019;105:136–41. https://doi.org/10.1016/j.jclinepi.2018.09.001.
55. Pot M, Kieusseyan N, Prainsack B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. Insights Imag. 2021;12:13. https://doi.org/10.1186/s13244-020-00955-7.
56. Ratner A, Bach SH, Ehrenberg H, et al. Snorkel: rapid training data creation with weak supervision. Proc VLDB Endow Int Conf Very Large Data Bases 2017;11:269–282. https://doi.org/10.14778/3157794.3157797
57. Ravì D, Wong C, Deligianni F, et al. Deep learning for health informatics. IEEE J Biomed Health Inform. 2017;21:4–21. https://doi.org/10.1109/JBHI.2016.2636665.
58. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
59. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol. 1990;132:734–45. https://doi.org/10.1093/oxfordjournals.aje.a115715.
60. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol. 1992;136:1400–13. https://doi.org/10.1093/oxfordjournals.aje.a116453.
61. Rosseel, Y. lavaan: an R package for structural equation modeling. J Stat Softw. 2012;48:1–36. https://doi.org/10.18637/jss.v048.i02.
62. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia; 2008.
63. Sánchez BN, Budtz-Jørgensen E, Ryan LM, Hu H. Structural equation models. J Am Stat Assoc. 2005;100:1443–55. https://doi.org/10.1198/016214505000001005.
64. Schnack, H. Bias, noise, and interpretability in machine learning. In: Machine Learning. Elsevier; 2020. p. 307–28
65. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58:323–37. https://doi.org/10.1016/j.jclinepi.2004.10.012.
66. Shanthini A, Vinodhini G, Chandrasekaran RM, Supraja P. A taxonomy on impact of label noise and feature noise using machine learning techniques. Soft Comput. 2019;23:8597–607. https://doi.org/10.1007/s00500-019-03968-7.
67. Shaw PA, Deffner V, Keogh RH, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. Ann Epidemiol. 2018;28:821–8. https://doi.org/10.1016/j.annepidem.2018.09.001.
68. Shaw PA, Gustafson P, Carroll RJ, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics. Stat Med. 2020;39:2232–63. https://doi.org/10.1002/sim.8531.
69. Sheppard L, Burnett RT, Szpiro AA, et al. Confounding and exposure measurement error in air pollution epidemiology. Air Qual Atmosphere Health. 2012;5:203–16. https://doi.org/10.1007/s11869-011-0140-9.
70. Shmueli G. To Explain or to Predict? Stat Sci. 2010;25.https://doi.org/10.1214/10-STS330.
71. Smedt TD, Merrall E, Macina D, et al. Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. PLoS ONE. 2018;13: e0199180. https://doi.org/10.1371/journal.pone.0199180.
72. Stefanski LA. Unbiased estimation of a nonlinear function a normal mean with application to measurement err oorf models. Commun Stat - Theory Methods. 1989;18:4335–58. https://doi.org/10.1080/03610928908830159.
73. Thiébaut ACM, Freedman LS, Carroll RJ, Kipnis V. Is It necessary to correct for measurement error in nutritional epidemiology? Ann Intern Med. 2007;146:65. https://doi.org/10.7326/0003-4819-146-1-200701020-00012.
74. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. Int J Epidemiol. 2020;49:338–47. https://doi.org/10.1093/ije/dyz251.
75. van der Wel MC, Buunk IE, van Weel C, et al. A novel approach to office blood pressure measurement: 30-minute office blood pressure vs

daytime ambulatory blood pressure. Ann Fam Med. 2011;9:128–35. https://doi.org/10.1370/afm.1211.
76. White JT, Fienen MN, Doherty JE. A python framework for environmental model uncertainty analysis. Environ Model Softw. 2016;85:217–28. https://doi.org/10.1016/j.envsoft.2016.08.017.
77. Yu AYX, Quan H, McRae AD, et al. A cohort study on physician documentation and the accuracy of administrative data coding to improve passive surveillance of transient ischaemic attacks. BMJ Open. 2017;7: e015234. https://doi.org/10.1136/bmjopen-2016-015234.
78. Zeger SL, Thomas D, Dominici F, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environ Health Perspect. 2000;108:419–26. https://doi.org/10.1289/ehp.00108419.
79. Zhu X, Wu X. Class noise vs. attribute noise: a quantitative study. Artif Intell Rev. 2004;22:177–210. https://doi.org/10.1007/s10462-004-0751-8.