Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Methods for Adjusting for Covariate Measurement Error in Flexible Modeling of Functional Form: Designing a Blinded, Controlled Neutral Comparison Simulation Study

Anne C. M. Thiébaut[1] | Aris Perperoglou[2] | Mohammed Sedki[3] | Steve Ferreira Guerra[4] | Paul Gustafson[5] | Frank E. Harrell Jr.[6] | Willi Sauerbrei[7] | Michal Abrahamowicz[4] | Laurence S. Freedman[8] | On behalf of Selection of Variables and Functional Forms in Multivariable Analysis Topic Group (TG2) and Measurement Error and Misclassification Topic Group (TG4) of the international STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative

[1]Université Paris-Saclay, UVSQ, Inserm, CESP, Villejuif, France | [2]Statistics and Data Science Innovation Hub, GlaxoSmithKline, London, UK | [3]Université Paris-Saclay, UVSQ, Gustave Roussy, Inserm, CESP, Villejuif, France | [4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada | [5]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada | [6]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA | [7]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany | [8]Information Management Services Inc., Calverton, Maryland, USA

Correspondence: Anne C. M. Thiébaut (anne.thiebaut@inserm.fr)

Received: 9 January 2026 | **Revised:** 19 April 2026 | **Accepted:** 11 June 2026

Keywords: bias | flexible modeling | measurement error | observational studies | STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative

ABSTRACT

This article describes the design of a neutral comparison study in the context of empirical studies where the interest is in learning the functional relationship between a continuous error-prone exposure variable and a binary outcome. The performance of combinations of measurement error correction methods and flexible regression modeling techniques was compared using a simulation study. The project involved four independent teams, one devoted to Data Generation and Evaluation and the other three to specific Methods of measurement error correction (regression-calibration and multiple imputation, simulation-extrapolation, and Bayesian method). The study was conducted in three successive stages. In Stage 1, the first team simulated five datasets differing only by the true exposure–outcome functional form and distribution of true exposure. Furthermore, the implementation of flexible modeling methods (B-splines, P-splines, and fractional polynomials) was standardized. The three Methods teams, blinded to the underlying data generation process, created the codes to implement their methods and provided their results to the first team, who evaluated them. These codes were then used by this team in the next stages of the project. In Stage 2, the team simulated 150 additional datasets where other design parameters varied while using the same five exposure–outcome functions. Stage 3 consisted of simulating independent replications of each of the 150 scenarios considered in Stage 2 to quantify the sampling variance of the estimates. This work emphasizes the relevance of neutral comparison studies to fairly evaluate statistical methods aimed at addressing a complex analytical challenge and demonstrates their feasibility through a large collaborative project.

1 | Introduction

The STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative aims to provide guidance on which statistical methods could be used for various types of observational studies (Sauerbrei et al. 2014, https://stratos-initiative.org). This guidance requires reliable evidence on the comparative advantages of different competing methods. Often this evidence

is provided by simulation studies (Boulesteix et al. 2020). Yet, biostatisticians teach clinicians the importance of unbiased, controlled evaluation of treatments, but rarely subject their own methods to the same standards (Boulesteix et al. 2017; Pawel et al. 2024; Siepe et al. 2024). Neutral comparison studies are those that are designed to provide such an unbiased, controlled comparison of competing statistical methods (Boulesteix et al. 2024). They are especially useful for comparing existing, already described, methods, rather than a new prototype method (Boulesteix et al. 2024; Heinze et al. 2024). When the study design is non-blinded, the researchers involved should be neutral, in the sense that they are reasonably familiar with all compared methods and without strong prior preference for or strong aversion to one of the comparator methods (Boulesteix et al. 2013).

This article addresses a complex issue where a neutral comparison study is desirable. We consider empirical studies where the primary interest is in learning the functional relationship between a continuous predictor, or exposure variable, and an outcome variable when the exposure variable is measured with error. When the shape of the exposure–outcome relationship is nonlinear, it is well known that random error in exposure measurements can mask the features of the relationship so that the observed shape may be distorted from the true shape (Carroll et al. 2006). For example, under specific circumstances (a single error-prone covariate with classical measurement error), the estimated relationship is likely to look more linear than it truly is (Keogh et al. 2012; Ferreira Guerra et al. 2026).

Various statistical methods are available for tackling this problem, with many flexible regression methods available for modeling nonlinear relationships on the one hand (e.g., Perperoglou et al. 2019; Royston et al. 1999) and measurement error correction methods on the other (Keogh et al. 2020; Shaw et al. 2020). However, relatively little is known about their relative performance when these two approaches are combined. To ensure a fair comparison of alternative strategies combining measurement error correction methods with flexible modeling methods, we thought it important to design a neutral comparison simulation study where researchers involved in proposing different analytical strategies would be blinded to the actual shape of the exposure-disease association. This article describes how such a neutral simulation study was designed. This was a joint work of two topic groups (TGs) of the STRATOS initiative: TG2 (selection of variables and functional forms) with expertise in the choice of functional forms for continuous variables in multivariable explanatory models and TG4 (measurement error and misclassification) with expertise in statistical methods to account for measurement error.

2 | Methods

This section is divided into two subsections, one detailing the project organization and how we ensured a neutral comparison design, the other briefly describing the simulation study.

2.1 | Neutral Comparison Design

The project involved four independent teams: one "Data Generation and Evaluation" team and three "Methods" teams (see Appendix for the list of members in each team). Each of the latter considered different measurement error correction methods (regression-calibration and multiple imputation, simulation-extrapolation [SIMEX], and Bayesian). The teams were independent in the sense that (i) membership of the teams was nonoverlapping, and (ii) the methods teams did not confer with each other when implementing their methods. The study was conducted in successive stages (Figure 1):

Stage 1 aimed at implementing the methods under blinded conditions with respect to the data characteristics. The Data Generation and Evaluation team was in charge of simulating the data and standardizing the choice and implementation of alternative flexible modeling methods (B-splines, P-splines, and fractional polynomials [FPs]). Five datasets were generated differing only by the true exposure-disease functional form and distribution of true exposure. The Data Generation and Evaluation team shared the five datasets with the Methods teams without revealing the underlying data generation process. Each of the three Methods teams, in turn, was in charge of creating the codes to implement their methods. After writing their code and applying it to the five datasets, they provided their results to the Data Generation and Evaluation team, who evaluated them. This evaluation was presented to the Methods teams in a blinded manner to give them feedback and motivate their interest in continuing the investigation. The alternative analytical strategies were given a blinded identification code and then ranked according to their relative overall performance, as well as separately for each of the five simulated datasets. When presenting the results, the Data Generation and Evaluation team did not reveal the identities of the analytical strategies nor the true functional form underlying each dataset. After the blinded results were presented, the Methods teams were asked to finalize their codes and submit them to the Data Generation and Evaluation team. Some of the Methods teams created different versions of their methods and were allowed to try them out on the first datasets. At the conclusion of Stage 1, they were asked to decide finally which version they wanted to run with. The codes provided were "generic," that is, allowed implementation of the respective methods for different datasets prepared according to the same principles. These generic codes were then used to analyze additional data simulated in the next stages of the project.

Stage 2 aimed at evaluating the performances of the methods under a much wider range of scenarios. Using the same five functions defining the exposure–outcome relationship as in Stage 1, the Data Generation and Evaluation team simulated 150 additional datasets where the combination of four other design parameters (main study sample size, replication study size, and the measurement error variance and its distribution) was varied. After the data were generated, the Data Generation and Evaluation team ran the codes provided by the Methods team on these new simulated datasets and evaluated the results. They then presented the evaluation results to the Methods team, this time unblinded. Only point estimates were evaluated at this stage.

Stage 3 aimed at assessing sampling variance for the methods' average performance across repeated sampling. It consisted of simulating 10 replications of each of the 150 scenarios considered in Stage 2. A further flexible modeling method, natural splines, was added at this stage in an unblinded fashion, due to concerns

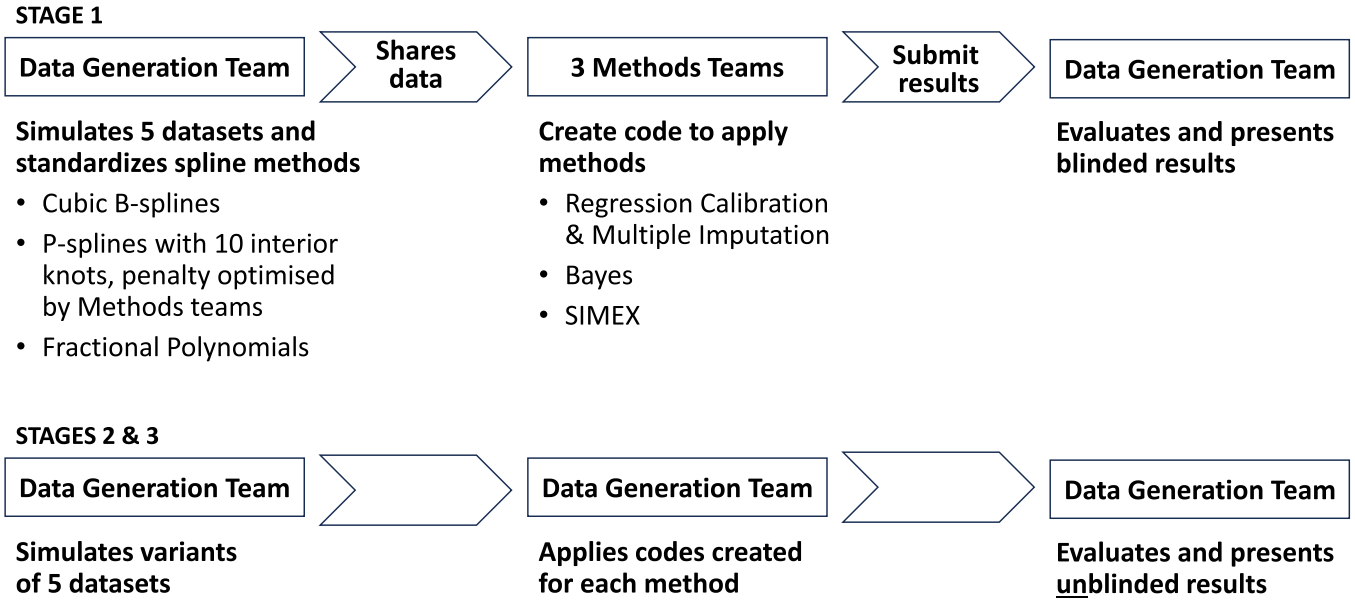

FIGURE 1 | Schematic representation of the project organization.

from members of the Methods teams that the B-spline method chosen initially may not have been optimal.

2.2 | Simulation Study

Following the recommendation of Morris et al. (2019), the description of the simulation study is organized according to the ADEMP structured approach: aims, data-generating mechanisms, estimands, methods, and performance measures. The purpose of this presentation is to give an overview of the simulation study that was conducted to answer the methodological question that motivated a neutral comparison design. More details about the methods and results may be found in Sedki et al. (2026).

2.2.1 | Aims

The aim of the simulation study was to compare the performance of combinations of measurement error correction methods and flexible regression methods in retrieving the true functional relationship between a single continuous exposure variable and a binary outcome variable when the exposure variable is measured with classical error.

2.2.2 | Data Generation

Each dataset mimicked a nested case–control study with a 1:4 case:control ratio. A binary outcome Y was generated from a logistic regression on continuous exposure X, that is, $\text{logit}(P(Y=1|X)) = f(X)$, where $f(X)$ represents a possibly nonlinear exposure–outcome relationship.

We considered five scenarios inspired by examples from real health studies. For each scenario, we searched the literature for data to define plausible parameters for the functional form $f(X)$ and the distribution of true exposure X (Table 1, Figure 2). To ensure that the overall event rate equals 20%, consistent with the pre-specified 1:4 case:control ratio, first a large number of pairs of (Y, X) were generated at random according to the true distribution of X and the relationship $\text{logit}(P(Y = 1)) = f(X)$, as specified in Table 1. Then from this large number of pairs, $n/5$ pairs were drawn at random from the pairs with $Y = 1$, and $4n/5$ pairs were drawn at random from the pairs with $Y = 0$. All simulations considered only $f(X)$ unadjusted for any other covariate.

In the first scenario, $f(X)$ was J-shaped borrowing from the relationship between body mass index (BMI) and all-cause mortality. The parameters of the functional form $f(X)$ were derived from a dose-response meta-analysis of 230 cohort studies (Aune et al. 2016). The distribution of BMI as true exposure X was taken as lognormal on the basis of Silverman and Lipscombe (2022).

The second scenario assumed a linear relationship as reported in the dietary fat and breast cancer analysis from the National Institutes of Health-AARP cohort study (Thiébaut et al. 2007). The exposure was expressed as log percent energy from total fat, that is, the logarithm of calories from usual dietary fat divided by energy intake in calories. Its true distribution was derived using information on the measurement error structure inferred from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study (Kipnis et al. 2003).

The third and fourth simulation scenarios were based on the relationship between air pollution levels and mortality rates (Cakmak et al. 1999). Exposure X was taken as air pollution level measured in $\mu g/m^2$. The authors postulated a threshold model for mortality rate with a linear increase above a certain threshold exposure level, denoted by τ (our third scenario). We alternatively assumed an exponential increase above the threshold exposure (our fourth scenario). Moreover, we changed the changepoint τ, from below (third scenario) to above (fourth scenario) the median of X because previous simulation work had shown that

TABLE 1 | Parameters used in the generation of the five synthetic datasets.

Scenario	Function form $f(X)$	Distribution of X	Values of X to be reported
(1) BMI and all-cause mortality	J-shape: logit $(P(y=1 \mid x)) = -2.202 + 268 \times \exp(-0.383 \times x) + 0.00197 \times \exp(0.139 \times x)$	X is taken as the BMI and is assumed to be lognormal: ln $X \sim N(3.3, 0.25^2)$ The variance of X is 50.5, the mean is 28, and the median is 27	From 10.0 to 50.0 inclusive in steps of 0.2
(2) Dietary fat and breast cancer incidence	Linear: logit $(P(y=1 \mid x)) = -5.78 + 0.545x$	X is taken as log percent energy from fat, that is, logarithm of calories from usual dietary fat intake divided by energy intake in calories and is assumed to be normal: $X \sim N(3.29, 0.24^2)$	From 2.6 to 4.0 inclusive in steps of 0.07
(3) Air pollution and mortality	Linear threshold: logit$(P(y=1 \mid x) = -2.2 + 0.0135\,(x-90)_+$ where $(x-\tau)_+ = 0$ for $x < \tau$ and $= (x-\tau)$ for $x \geq \tau$	X is taken to be air pollution level and is assumed to be normal: $X \sim N(100, 19.4^2)$	From 45.0 to 155.0 inclusive in steps of 0.55
(4) Air pollution and mortality	Exponential threshold: logit$(P(y=1 \mid x) = -3.2 + \exp[0.02\,(x-120)_+]$	X is taken to be air pollution level and is assumed to be lognormal: ln $X \sim N(4.5, 0.23^2)$	From 40.0 to 160.0 inclusive in steps of 0.6
(5) Drug ingestion and side effect	Saturation: logit$(P(y=1 \mid x)) = \ln((3/7)\,x^6 + (1/19)\,50^6) - \ln(50^6 + x^6)$	X represents the dose of medication that may vary across patients due to different practices among prescribing physicians and is assumed to be uniform: $X \sim U(30, 80)$	From 18.0 to 92.0 inclusive in steps of 0.37

Abbreviation: BMI, body mass index.

the position of the threshold may qualitatively impact the biases caused by measurement error (Küchenhoff and Carroll 1997).

The fifth scenario was inspired by the Hill equation that is frequently used in pharmacological modeling to describe a saturation effect (Goutelle et al. 2008). We postulated a side effect whose risk is dependent on the dose of drug X that is taken by the patient, resulting in a monotonic increasing relationship that eventually reaches a plateau. The distribution of the doses was assumed to be uniform, reflecting different practices among prescribing physicians.

For each of the above scenarios, after simulating true exposure values X and corresponding values of outcome Y, an error-prone $X^* = X + e$ was generated, where measurement error e was assumed nondifferential with respect to Y, classical (purely random), with a normal distribution of mean 0 and variance equal to half that of the variance of true exposure X. In Stage 1, the main study data consisted of 15,000 individuals each with an independent realization of (Y, X^*). In addition, a replication sub-study consisting of a random subgroup of 250 participants of the main study, with a second independent realization of X^*, was simulated. Methods teams were not informed of the theoretical measurement error variance, but they were able to estimate it from the replication data provided to them. Relatively large sample sizes were chosen in Stage 1 to ensure convergence of flexible modeling methods. Smaller sample sizes were investigated later in Stages 2 and 3.

2.2.3 | Estimands

The estimands of interest were the values of the adjusted true function $f^*(X)$ at the preselected series of X values. The adjustment was made to reflect the 1:4 nested case–control ratio in the data generated and was given by $f^*(X) = f(X) + \log(0.25(1-p)/p)$, where $p = P(Y=1)$ in the population. For each dataset, the Data Generation and Evaluation team provided the Methods teams with a sequence of 200 X values, for example, for dataset 1, from 10.0 to 50.0 inclusive, in steps of 0.2 (see Table 1). The selected X values covered the range of the erroneous X^* values and thus were much wider than the true range of X. The Methods teams were expected to provide corresponding point estimates of $f^*(X)$ for each of these 200 X values.

2.2.4 | Methods

The general principles of the concurrent methods will be presented here. More details and further references may be found in Keogh et al. (2020) and Shaw et al. (2020).

2.2.4.1 | Flexible Modeling Methods. In Stage 1, three flexible modeling methods were considered. The first one was cubic B-splines with one interior knot at the median of observed X^*, whereas the boundary knots were placed at its minimum and maximum (4 degrees of freedom). The second one was P-splines with 10 interior knots, with penalty to be optimized by the

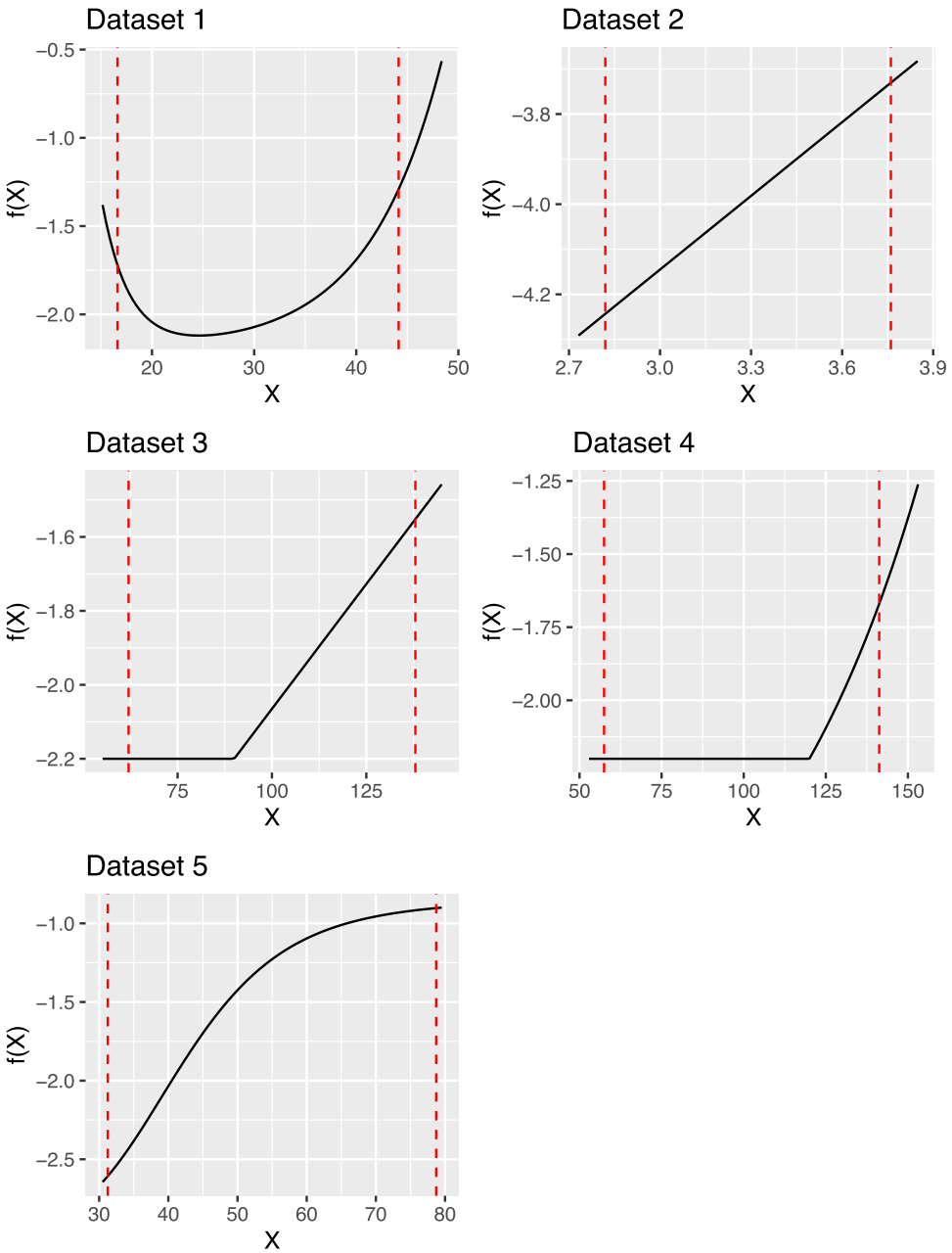

FIGURE 2 | Five forms of $f(X)$ used in the simulations: (1) J-shaped, (2) linear, (3) threshold below the median and linear increase, (4) threshold above the median and exponential increase, and (5) saturation. The red dotted vertical bars correspond to the limit points.

Methods teams. The third one was FPs of second degree (4 degrees of freedom) that permits estimating non-monotone functions, with the same initial set of powers, but leaving it up to the Methods teams to select the optimal two powers.

After disclosure of the true models underlying the simulated data, it was decided to additionally implement natural splines as a fourth flexible modeling method in Stage 3. The natural splines (2 degrees of freedom) were specified as having three knots, one at the median, and two at the boundaries, that is, the 5th and 95th percentiles of X^*.

2.2.4.2 | Measurement Error Correction Methods. After discussion among the project group and with other experts on measurement error adjustment methods, four methods were identified as being accepted, well-researched, and easily programmable: regression-calibration, multiple imputation, SIMEX, and Bayesian. The Methods teams were formed on the basis of the expertise of their members with these methods.

The regression-calibration approach consists of estimating the conditional expectation of the function $f(X)$ given the error-prone covariate X^* and substituting it for the true covariate in the logistic regression (Rosner et al. 1989). In the multiple imputation approach, the imputed $f(X)$ consists of its conditional expectation given X^* and Y plus imputed values of the regression residual (Cole et al. 2006). Imputation was done 10 times using different model parameter values from the corresponding estimated distributions.

The SIMEX approach is a two-step method that has been adapted to various measurement error problems (Cook and Stefanski 1994; Carroll et al. 2006). The general idea is to sequentially simulate new values of error-prone exposures with gradually increasing measurement error, and then, for each of the assumed measurement error levels, estimate the parameter of interest using the generated variables. The estimates being increasingly biased, this establishes a relationship between the estimate (and thus, implicitly, the amount of bias) and amount of added measurement error. The second step consists in extrapolating this relationship back to the case of no error. For this project, the team considered two alternative SIMEX approaches applied either (i) to the individual points on the curve or (ii) to the estimated B-spline or FP coefficients (the latter approach was not feasible for P-splines).

In the Bayesian approach (Gustafson 2003), the team specified three models: an outcome model for Y given X, an exposure model for X, and a measurement error model for X^* given X, along with prior distributions for relevant parameters in each of the three sub-models. This defined a joint posterior distribution of all parameters and latent X values, given all the observed data. The team used a particular Monte Carlo Markov Chain algorithm, Hamiltonian Monte Carlo (Betancourt and Girolami 2013), as implemented in the Stan software (Stan Development Team 2025) to obtain a Monte Carlo approximation to this posterior distribution. Two different summaries for inference were considered: (i) risk, corresponding to the logit of the posterior mean of risk on the $P(Y = 1)$ scale, and (ii) logit, corresponding to the posterior mean on the logit$(P(Y = 1))$ scale.

2.2.5 | Performance Measures

The main evaluation criterion in Stage 1 was the mean squared error (MSE), that is, the mean of the squared difference between the estimated $f^*(X)$, obtained using a given analytical method, and the true $f^*(X)$ over the 200 preselected values of X that lay within limit points. These limit points were chosen by the Data Generation and Evaluation team and kept undisclosed to the Methods teams. They were defined as the 2.5th and 97.5th percentiles of the distribution of true exposure X to avoid an excessive impact of estimates in the tails of the distribution. Values among the 200 estimates of $f^*(X)$ provided by the Methods teams for X outside of these limits were discarded. The overall performance was evaluated as the mean of the MSEs over the scenarios. In Stages 2 and 3, the main criterion was switched to log MSE due to the right-skewed distribution of MSE values.

For Stage 1 results, two sorts of benchmarks were obtained based, respectively, on either using true exposure X (as a lower bound for MSE) or uncorrected B-spline and P-spline estimation applied to observed exposure X^* (expected to mark the upper bound on errors). The latter was not performed for FPs, or with the extended scenarios and replications in Stages 2 and 3.

3 | Results

Here we briefly report on the Stage 1 results to emphasize the interest of the neutral comparison design. Results from the extensive simulation study, including variants of the original five scenarios and the addition of natural splines as a fourth flexible modeling method, are presented in Sedki et al. (2026).

Table 2 displays the unblinded results with benchmarks. It should be noted that these results were initially presented in a blinded fashion to the Methods teams, that is, were displayed in decreasing order of overall performance but with the blinded identification code replacing the labels of specific methods.

Both versions of SIMEX performed best, with a slight advantage of SIMEX applied to individual points over that applied to coefficients. Multiple imputation performed second best. The Bayesian methods and the regression-calibration methods performed, on average, more poorly than the benchmark applied to observed X^*. The ranking across measurement error correction methods did not vary across the flexible modeling methods except for one notable exception: The Bayesian approach combined with B-splines performed much poorer than that combined with P-splines or FPs, and worse than any of the other methods.

For some scenarios, the spline estimates obtained using uncorrected X^* values had MSEs that were lower than several of the correction methods (Table 2). This, rather unexpected, finding may partly reflect the bias-variance tradeoff. Although measurement error correction methods decrease bias, they also increase variance, so that MSEs are not guaranteed to be lower than uncorrected analyses.

4 | Discussion

In this article, we report a multicenter experiment of a neutral comparison study addressing a complex issue often encountered in epidemiological studies: How to retrieve the shape of an exposure–outcome association when the exposure is measured with error? This article is not intended to provide guidance on the most appropriate analytical methods to use, which are addressed in Sedki et al. (2026), but rather to describe the study design for comparing such methods. Designing a neutral comparison study for addressing our statistical question was crucial. Had the Methods teams known the shapes of the relationships, they could have been tempted to adjust tuning, which would have introduced overoptimism and, thus, compromised the practical usefulness of the results.

Neutral comparison studies have been advocated to provide an unbiased, controlled comparison of competing statistical methods (Boulesteix et al. 2024). In the present experiment, neutrality and maintenance of study design were ensured by the blinded design, by the participation of motivated independent international teams, and by the strong commitment of a chair person. Blind analyses have notably been proposed as a correction for confirmation bias, whereby one concurrent method could be favored over the others at the outset (MacCoun and Perlmutter 2017). Among several possible techniques, our choice was to keep the data generation process hidden to those who then implemented the concurrent methods (Dutilh et al. 2019). To prevent peeking, the members of the Methods teams were from different institutions than those of the Data Generation and Evaluation team. To enforce motivation, regular virtual meetings were held

TABLE 2 | Mean Squared Errors (MSE) of analytic methods with benchmarks. The methods are ordered according to the MSE from lowest to highest.

Method	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
Bench-PS X	**0.0035**	**0.00008**	**0.00280**	**0.0029**	**0.0035**	**0.0026**
Bench-BS X	**0.0029**	**0.00160**	**0.00203**	**0.0034**	**0.0040**	**0.0028**
SIMEX-PS	0.0034	0.00149	0.00454	0.0039	0.0103	0.0047
SIMEX-BS	0.0078	0.00264	0.00278	0.0033	0.0156	0.0064
SIMEX-FP	0.0054	0.00159	0.00893	0.0069	0.0137	0.0073
SIMEX-BS-coef	0.0068	0.00236	0.00430	0.0052	0.0223	0.0082
SIMEX-FP-coef	0.0067	0.00098	0.01078	0.0142	0.0131	0.0092
MI-BS	0.0101	0.00080	0.00722	0.0150	0.0200	0.0106
MI-PS	0.0108	0.00040	0.00683	0.0157	0.0209	0.0109
MI-FP	0.0099	0.00073	0.00840	0.0165	0.0207	0.0112
Bench-PS X*	**0.0101**	**0.00418**	**0.00850**	**0.0023**	**0.0314**	**0.0113**
Bench-BS X*	**0.0124**	**0.00449**	**0.00594**	**0.0028**	**0.0311**	**0.0113**
Bayes-PS-risk	0.0064	0.00097	0.00919	0.0188	0.0339	0.0139
Bayes-PS-logit	0.0060	0.00102	0.01012	0.0166	0.0369	0.0141
Bayes-FP-risk	0.0139	0.00135	0.01397	0.0326	0.0161	0.0156
Bayes-FP-logit	0.0137	0.00141	0.01457	0.0322	0.0167	0.0157
RC-BS	0.0234	0.00345	0.01085	0.0447	0.0238	0.0212
RC-PS	0.0318	0.00057	0.00597	0.0545	0.0171	0.0220
RC-FP	0.0320	0.00129	0.01277	0.0543	0.0135	0.0228
Bayes-BS-risk	0.0134	0.00187	0.14832	0.1047	0.0868	0.0710
Bayes-BS-logit	0.0130	0.00210	0.38618	0.1102	0.1093	0.1242

Note: The methods are ordered according to the MSE from lowest to highest. Lines in bold correspond to benchmark results, which are supposed to provide upper and lower bounds (although some methods performed even worse).

Abbreviations: Bench, benchmark; BS, B-splines; FP, fractional polynomials; MI, multiple imputation; PS, P-splines, RC, regression-calibration; SIMEX, simulation-extrapolation.

for the full project group, within each Methods team, as well as within the Data Generation and Evaluation team. Similarly, Geroldinger et al. (2024) built upon an existing consortium of researchers willing "to learn from each other, and to openly discuss advantages and drawbacks of the respective methods." Our experiment is another illustration of an "adversarial collaboration" (Cowan et al. 2020) in the sense that proponents of different measurement error correction methods have agreed and worked together to test their intuition. Combining adversarial collaboration and blind analysis is an innovative feature that has hardly ever been applied in methodological simulation research (Kreutz et al. 2020).

Although Stage 1 results are clearly insufficient to draw conclusions, we should admit that they looked intriguing to the Data Generation and Evaluation team. From theoretical considerations, we had hypothesized that the Bayesian and multiple imputation methods would be expected to perform better than regression-calibration, which, in turn, would perform better than SIMEX. Instead, Stage 1 results suggested that, on average, SIMEX performed best, and multiple imputation was next best. The Bayesian method with FPs or P-splines came third, before regression-calibration, whereas the Bayesian method with B-splines performed poorly. In the face of such surprising results, the blinded design helped to insure against risks of conscious or unconscious biases (MacCoun and Perlmutter 2015).

Beyond the neutral comparison design, another strength of our work is the choice of realistic functional shapes using parameters derived from real datasets (Sauer et al. 2025). The choice of the five simulation scenarios was driven by the need of a variety of plausible functional relationships and was not motivated at all by the idea that some methods may fail or better succeed in specific cases. The only exception was the linear relationship, where we expected all methods to perform best but were also interested to see if measurement error may induce some spurious nonlinearities. Choosing these scenarios independently from the methods implementation and in a blinded fashion ensured prevention of selective reporting and "cherry-picking" of more favorable results, as did evaluation of results by an independent team (Pawel et al. 2024; Siepe et al. 2024). Averaging performances across scenarios in Stage 1 provided a general idea allowing us to discard variants of methods that would perform especially poorly in all instances. In Stage 2, when we considered variations of the five simulation scenarios, some were indeed motivated by

the intuition that one specific method may perform more poorly (Strobl and Leisch 2024). Nonetheless, all methods were subjected to the same scenario variants, so we believe the comparison remains fair and systematic.

A requirement for neutral comparison studies is the choice of the evaluation criteria in a rational way (Boulesteix et al. 2013). The choice of the performance measure was highly debated in the early phases of the project. We considered unweighted MSE and MSE weighted by the density of X and mean absolute error, on either the log odds scale or the risk scale. For Stage 1, after finding that weighted and unweighted MSE were similar, we chose to present unweighted MSE on the log odds scale. For the later stages, however, we finally switched to log MSE. Other criteria would have been possible, as highlighted in Ullmann et al. (2025)'s categorization of performance measures for evaluating estimated associations between continuous predictors and an outcome.

We note some practical disadvantages of the blinded, controlled design. It was more difficult to organize and offered less flexibility to change or introduce new methods. This happened during the present experiment with the later introduction of natural splines that had to be done unblinded. In fact, after seeing unblinded results, members of the Methods teams expressed concerns that the B-spline method was not optimal. For this reason, a fourth method, natural splines, was considered in Stage 3. The fact that the choice was made a posteriori yet weakens any conclusion that can be drawn about the possible superiority of natural splines over other flexible regression methods.

Although efforts have been made to standardize the flexible modeling methods as much as possible, we cannot rule out the possibility that performance discrepancies between two measurement error correction methods using the same flexible modeling approach may result in part from some variations regarding how different Methods teams implemented the same flexible method (e.g., choice of the penalty for P-splines or the algorithm to choose best-fitting second-degree FPs). Indeed, Methods teams sometimes encountered technical difficulties in combining a specific measurement error correction method with a flexible modeling approach. For example, the SIMEX coefficient variant was not feasible for P-splines.

Our study could also be considered limited by the fact that each of the competing methods was implemented by only one of the three nonoverlapping teams. An ideal framework would have involved several teams devoted to the same competing method. It has been recognized that several teams using the same method may indeed provide different results, possibly as a consequence of different levels of expertise (Nieβl et al. 2024).

In conclusion, this work emphasizes the relevance of neutral comparison studies to fairly evaluate statistical methods aimed at addressing a complex analytical challenge and demonstrates their feasibility through a large collaborative project. Blinded, team-based, neutral comparison studies seem most appropriate when knowing the features of the data could influence the tuning of the methods, whereas truth is unobservable in real practice. They are also advisable when one of the methods can currently be conducted only by the team who developed the method.

Acknowledgments

The authors are grateful to all members of Measurement Error and Misclassification Topic Group (TG4) of the international STRengthening analytical thinking for observational studies (STRATOS) initiative for providing valuable comments on this work, in particular (in alphabetical order) Drs. Sabine Hoffmann and Pamela Shaw. They would also like to thank Georg Heinze (STRATOS TG2) and an anonymous referee for constructive and helpful comments that improved the manuscript.

Open access publication funding provided by COUPERIN CY26.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

Aune, D., A. Sen, M. Prasad, et al. 2016. "BMI and All Cause Mortality: Systematic Review and Non-Linear Dose-Response Meta-Analysis of 230 Cohort Studies With 3.74 Million Deaths Among 30.3 Million Participants." *Bmj* 353: i2156.

Betancourt, M., and M. Girolami. 2013. "Hamiltonian Monte Carlo for Hierarchical Models." Preprint, arXiv:1312.0906v1, December 3. http://arxiv.org/abs/1312.0906.

Boulesteix, A. L., M. Baillie, D. Edelmann, L. Held, T. P. Morris, and W. Sauerbrei. 2024. "Editorial for the Special Collection "Towards Neutral Comparison Studies in Methodological Research"." *Biometrical Journal* 66, no. 2: e2400031.

Boulesteix, A. L., R. H. Groenwold, M. Abrahamowicz, et al. 2020. "Introduction to Statistical Simulations in Health Research." *BMJ Open* 10, no. 12: e039921.

Boulesteix, A. L., S. Lauer, and M. J. Eugster. 2013. "A Plea for Neutral Comparison Studies in Computational Sciences." *PLoS ONE* 8, no. 4: e61562.

Boulesteix, A. L., R. Wilson, and A. Hapfelmeier. 2017. "Towards Evidence-Based Computational Statistics: Lessons From Clinical Research on the Role and Design of Real-Data Benchmark Studies." *BMC Medical Research Methodology* 17, no. 1: 138.

Cakmak, S., R. T. Burnett, and D. Krewski. 1999. "Methods for Detecting and Estimating Population Threshold Concentrations for Air Pollution-Related Mortality With Exposure Measurement Error." *Risk Analysis* 19: 487–496.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Chapman and Hall/CRC,.

Cole, S. R., H. Chu, and S. Greenland. 2006. "Multiple-Imputation for Measurement-Error Correction." *International Journal of Epidemiology* 35: 1074–1081.

Cook, J. R., and L. A. Stefanski. 1994. "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association* 89: 1314–1328.

Cowan, N., C. Belletier, J. M. Doherty, et al. 2020. "How Do Scientific Views Change? Notes From an Extended Adversarial Collaboration." *Perspectives on Psychological Science* 15, no. 4: 1011–1025.

Dutilh, G., A. Sarafoglou, and E. J. Wagenmakers. 2019. "Flexible Yet Fair: Blinding Analyses in Experimental Psychology." *Synthese* 198: S5745–S5772.

Ferreira Guerra, S., L. Freedman, P. Gustafson, W. Sauerbrei, F. E. Harrell Jr., and M. Abrahamowicz. 2026. "Assessing the Impact of Measurement Error on Fractional Polynomials and Spline Estimates of Non-Linear Dose-Response Functions: A Simulation Study." Under Review in Biometrical Journal.

Geroldinger, M., J. Verbeeck , K. E. Thiel, et al. 2024. "A Neutral Comparison of Statistical Methods for Analyzing Longitudinally Measured Ordinal Outcomes in Rare Diseases." *Biometrical Journal* 66, no. 1: e2200236.

Goutelle, S., M. Maurin, F. Rougier, et al. 2008. "The Hill Equation: A Review of Its Capabilities in Pharmacological Modelling." *Fundamental & Clinical Pharmacology* 22, no. 6: 633–648.

Gustafson, P. 2003. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman and Hall/CRC.

Heinze, G., A. L. Boulesteix, M. Kammer, T. P. Morris, and I. R. White, the Simulation Panel of the STRATOS Initiative. 2024. "Phases of Methodological Research in Biostatistics—Building the Evidence Base for New Methods." *Biometrical Journal* 66, no. 1: e2200222.

Keogh, R. H., P. A. Shaw, P. Gustafson, et al. 2020. "STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 1—Basic Theory and Simple Methods of Adjustment." *Statistics in Medicine* 39, no. 16: 2197–2231.

Keogh, R. H., A. D. Strawbridge, and I. R. White. 2012. "Effects of Classical Exposure Measurement Error on the Shape of Exposure-Disease Associations." *Epidemiologic Methods* 1, no. 1: 13–32.

Kipnis, V., A. F. Subar, D. Midthune, et al. 2003. "Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study." *American Journal of Epidemiology* 158, no. 1: 14–21.

Kreutz, C., N. S. Can, R. S. Bruening, et al. 2020. "A Blind and Independent Benchmark Study for Detecting Differentially Methylated Regions in Plants." *Bioinformatics* 36, no. 11: 3314–3321. https://doi.org/10.1093/bioinformatics/btaa191.

Küchenhoff, H., and R. J. Carroll. 1997. "Segmented Regression With Errors in Predictors: Semi-Parametric and Parametric Methods." *Statistics in Medicine* 16, no. 1–3: 169–188.

MacCoun, R., and S. Perlmutter. 2015. "Blind Analysis: Hide Results to Seek the Truth." *Nature* 526: 187–189.

MacCoun, R., and S. Perlmutter. 2017. "Blind Analysis as a Correction for Confirmatory Bias in Physics and in Psychology." In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, edited by S. O. Lilienfeld and I. Waldman, 297–322. John Wiley and Sons.

Morris, T. P., I. R. White, and M. J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38, no. 11: 2074–2102.

Nieβl, C., S. Hoffmann, T. Ullmann, and A. L. Boulesteix. 2024. "Explaining the Optimistic Performance Evaluation of Newly Proposed Methods: A Cross-Design Validation Experiment." *Biometrical Journal* 66, no. 1: e2200238.

Pawel, S., L. Kook, and K. Reeve. 2024. "Pitfalls and Potentials in Simulation Studies: Questionable Research Practices in Comparative Simulation Studies Allow for Spurious Claims of Superiority of Any Method." *Biometrical Journal* 66, no. 1: 2200091.

Perperoglou, A., W. Sauerbrei, M. Abrahamowicz, and M. Schmid. 2019. "A Review of Spline Function Procedures in R." *BMC Medical Research Methodology* 19: 46.

Rosner, B., W. C. Willett, and D. Spiegelman. 1989. "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error." *Statistics in Medicine* 8: 1051–1069.

Royston, P., G. Ambler, and W. Sauerbrei. 1999. "The Use of Fractional Polynomials to Model Continuous Risk Variables in Epidemiology." *International Journal of Epidemiology* 28: 964–974.

Sauer, C., F. J. D. Lange, M. Thurow, I. Dormuth, and A. L. Boulesteix. 2025. "Statistical Parametric Simulation Studies Based on Real Data." Preprint, arXiv, April 7. https://doi.org/10.48550/ARXIV.2504.04864.

Sauerbrei, W., M. Abrahamowicz, D. G. Altman, S. le Cessie, and J. Carpenter, On behalf of the STRATOS initiative. 2014. "STRengthening Analytical Thinking for Observational Studies: The STRATOS Initiative." *Statistics in Medicine* 33, no. 30: 5413–5432.

Sedki, M., A. Perperoglou, A. C. M. Thiébaut, et al. 2026. Comparison of methods for adjusting for covariate measurement error in flexible modelling of functional form: results of a blinded, controlled neutral comparison simulation study. arXiv:2606.10123.

Shaw, P. A., P. Gustafson, R. J. Carroll, et al. 2020. "STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 2—More Complex Methods of Adjustment and Advanced Topics." *Statistics in Medicine* 39, no. 16: 2232–2263.

Siepe, B. S., F. Bartoš, T. P. Morris, A. L. Boulesteix, D. W. Heck, and S. Pawel. 2024. "Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting." Psychological Methods.

Silverman, M., and T. Lipscombe. 2022. "Exact Statistical Distribution of the Body Mass Index (BMI): Analysis and Experimental Confirmation." *Open Journal of Statistics* 12: 324–356.

Stan Development Team. 2025. *Stan Reference Manual, Version 2.37*. Stan Development Team. https://mc-stan.org.

Strobl, C., and F. Leisch. 2024. "Against the "One Method Fits all Data Sets" Philosophy for Comparison Studies in Methodological Research." *Biometrical Journal* 26, no. 1: 2200104.

Thiébaut, A. C. M., V. Kipnis, S. C. Chang, et al. 2007. "Dietary Fat and Postmenopausal Invasive Breast Cancer in the National Institutes of Health-AARP Diet and Health Study Cohort." *Journal of the National Cancer Institute* 99, no. 6: 451–462.

Ullmann, T., G. Heinze, M. Abrahamowicz, et al. 2025. "A Systematic Categorization of Performance Measures for Estimated Non-Linear Associations Between an Outcome and Continuous Predictors." *WIREs Computational Statistics* 17, no. 3: e70042.

Appendix

A1 List of members of each team (in alphabetical order)

- Data Generation and Evaluation team

Anne Thiébaut (TG4), Laurence Freedman (TG4), Aris Perperoglou (TG2), and Mohammed Sedki (affiliate)

- Simulation-extrapolation (SIMEX) team

Michal Abrahamowicz (TG2) and Steve Ferreira Guerra (affiliate)

- Regression-calibration and multiple imputation team

Victor Kipnis (TG4), Brian Barrett (affiliate), Matthew Chaloux (affiliate), Kevin Dodd (TG4), Douglas Midthune (TG4), and Amer Moosa (affiliate)

- Bayesian team

Paul Gustafson (TG4), Raymond Carroll (TG4), Frank Harrell (TG2), and Nadja Klein (affiliate)