

ISCB-GMDS conference – Mini-Symposium of the STRATOS initiative

Methods for adjusting for covariate measurement error
in flexible modelling of functional form:
Results of a blinded, controlled neutral comparison simulation study

Anne Thiébaud, Mohammed Sedki, Aris Perperoglou, Laurence Freedman
on behalf of Selection of Variables and Functional Forms in Multivariable Analysis Topic Group (TG2)
and Measurement Error and Misclassification Topic Group (TG4)

October 1st, 2026



Outline

1 Background and Research Objectives

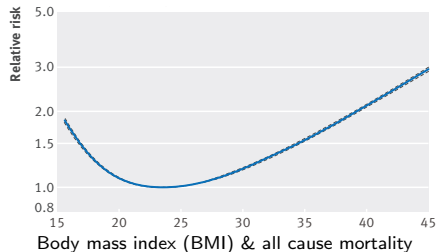
2 Material and Methods

- Study design and project organization
- Data generation
- Estimands
- Methods
- Performance measures

3 Results

4 Discussion

Background: Functional form of a continuous covariate



Aune et al., BMJ 2016

- Mathematical relationship between the independent variable (X) and the dependent variable (Y)
- $f(X)$ can be linear, nonlinear, or a combination of both
- The choice of functional form can have a significant impact on the results of statistical analysis

TG2: Perperoglou et al., BMC Med Res Methodol 2019
stratostg2.github.io/glossary.html

Background: Measurement error (ME) in a continuous covariate

- X cannot be observed directly, we only have an error-prone measurement of it, X^*
 - ▶ Classical measurement error: $X^* = X + U$ where $\mathbb{E}(U|X) = 0$ and $U \perp\!\!\!\perp X$
 - ▶ Nondifferential with respect to the outcome: $f(Y|X, X^*) \equiv f(Y|X)$
- Well-documented impacts on the regression relationship
 - ▶ When X is linearly related to Y
 - ✓ Bias (attenuation) in the estimated association
 - ✓ Loss of statistical power
 - ▶ When X is not linearly related to Y
 - ✓ Distorsion of the functional form towards linearity

TG4: Keogh et al., Shaw et al., Stat Med 2020

TG2-TG4: Ferreira Guerra et al., Biom J 2026

Popular flexible modeling methods

When X is measured exactly

- ✓ B-splines
- ✓ Natural (cubic) splines
- ✓ P-splines
- ✓ Fractional polynomials

TG2: Perperoglou et al., BMC Med Res Methodol 2019

Sauerbrei et al., Diagn Progn Res 2020

Popular flexible modeling and ME correction methods

When X is measured exactly

- ✓ B-splines
- ✓ Natural (cubic) splines
- ✓ P-splines
- ✓ Fractional polynomials

When X is measured with error and $f(X)$ is linear

- ✓ Simulation-extrapolation (SIMEX)
- ✓ Regression-calibration
- ✓ Multiple imputation
- ✓ Bayesian estimation

All these remove bias but do not recover lost precision

TG2: Perperoglou et al., BMC Med Res Methodol 2019

Sauerbrei et al., Diagn Progn Res 2020

TG4: Keogh et al., Shaw et al., Stat Med 2020

Research objectives

To compare the performance of methods combining

- a ME correction method with • one flexible modeling approach

in retrieving the true functional form $f(X)$ when X is measured with error

Rationale for a neutral comparison study

- Existing methods already described, no new prototype method
- In real settings, the form of the relationship is unknown

⇒ Experimental design (simulations)

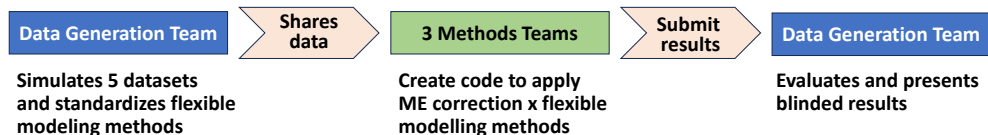
However,

- Knowing the form of the relationship that is simulated, one could be tempted to tune ME correction methods
- This could introduce overoptimism and, thus, compromise the practical usefulness of the results

⇒ Implementation of competing methods should be carried out **blind to data generation**

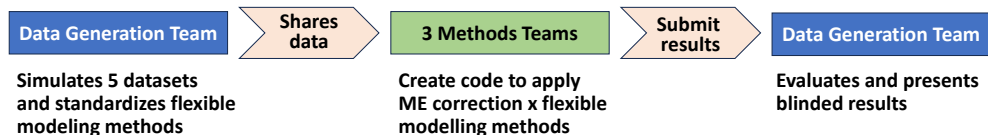
Project organization: 3 stages, 4 independent teams

- **Stage 1:** Implement the methods under blinded conditions with respect to data characteristics



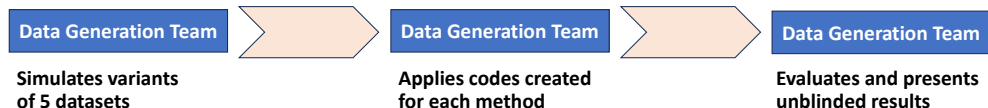
Project organization: 3 stages, 4 independent teams

- **Stage 1:** Implement the methods under blinded conditions with respect to data characteristics



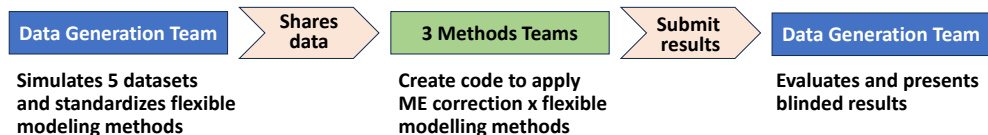
- **Stage 2:** Evaluate methods' performances under wider range of scenarios

- ▶ Variations of design parameter, same 5 functional relationships



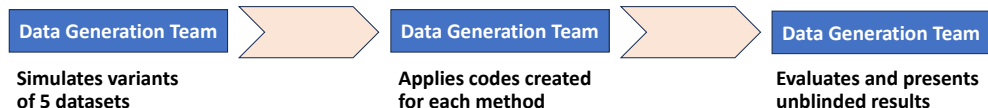
Project organization: 3 stages, 4 independent teams

- **Stage 1:** Implement the methods under blinded conditions with respect to data characteristics



- **Stage 2:** Evaluate methods' performances under wider range of scenarios

- ▶ Variations of design parameter, same 5 functional relationships



- **Stage 3:** Assess sampling variance

- ▶ Replications of each of Stage 2 scenarios

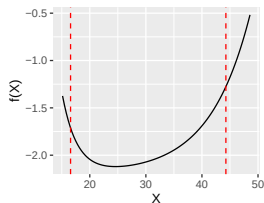
Data generation: General principles

- Binary outcome Y linked to continuous X by logistic regression

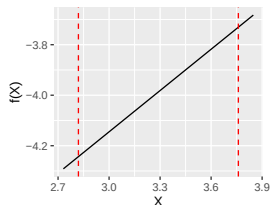
$$\text{logit}(P(Y = 1|X)) = f(X)$$

- Case-control ratio 1:4
- Undisclosed values or distribution of X and undisclosed form of $f(X)$
- Values of X^* (X perturbed by nondifferential, classical ME) provided in place of X
- Variance and distribution of measurement error undisclosed, subset of replicated values of X^* provided instead

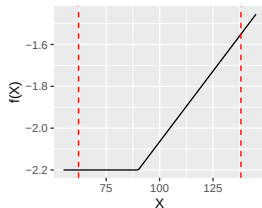
Functional forms of $f(X)$ inspired by real health studies



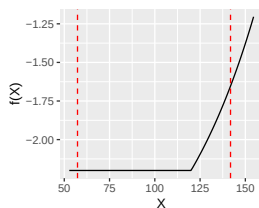
(1) J-shape: BMI & mortality



(2) Linear: Dietary fat & breast cancer

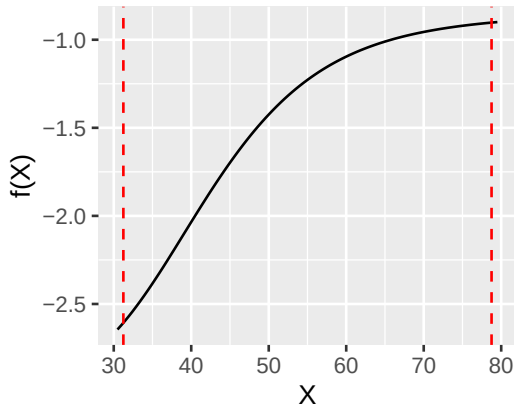


(3) Threshold with change below median and linear increase: Air pollution & mortality



(4) Threshold with change above median and exponential increase: Air pollution & mortality

Functional forms of $f(X)$ inspired by real health studies



(5) Saturation: Dose of drug & pharmacological effect (Hill equation)

Data generation: Stage 1

- 5 datasets
- Main study: $N = 15,000$ independent realizations of
 - ▶ Binary outcome Y (3,000 individuals having $Y = 1$)
 - ▶ Continuous covariate measured with error X^*
 $\sim$ log-normal (datasets 1 & 4), normal (datasets 2 & 3) or uniform (dataset 5) distribution
- Measurement error
 - ▶ Normal distribution
 - ▶ Variance = $\text{Var}(X)/2$
- Replication substudy: $n = 250$ individuals
 - ▶ Duplicate values for X^*

Data generation: Stages 2 & 3

- Stage 2: 30 design parameter combinations not considered in Stage 1
 - ▶ Main sample size $N = (15,000, 30,000)$ later expanded to smaller (5,000, 2,000)
 - ▶ Replication substudy size $n = (250, 750)$
 - ▶ ME variance = $\text{Var}(X)/2$ or $\text{Var}(X)$
 - ▶ ME distribution = normal or shifted Gamma with shape parameter 3, scaled to target variance
- Stage 3: each of Stage 2 scenario $\times$ parameter combinations simulated 10 times
 $\Rightarrow$ 1,500 datasets

Estimands

- Values of adjusted true function

$$f^*(X) = f(X) + \log(0.25(1 - p)/p) \quad \text{where } p = P(Y = 1)$$

reflecting the 1:4 nested case-control ratio

- Data Generation and Evaluation team provided a sequence of 200 X values covering the range of the error-prone X^*
- Methods teams were expected to provide corresponding 200 point estimates of $f^*(X)$

Flexible modeling methods

- Cubic B-splines

- ▶ One interior knot at the median of observed X^*
- ▶ Boundary knots at its minimum and maximum
- ▶ 4 degrees of freedom

- Penalized (P-)splines

- ▶ 10 interior knots at quantiles of X^*
- ▶ Penalty to be optimized by Methods teams

- Second degree fractional polynomials (FP2)

- ▶ Same initial set of powers
- ▶ Methods teams to select the optimal 2 powers

- Natural cubic splines (in Stage 3)

- ▶ 3 knots at median, 5th and 95th percentiles of X^*

TG2: Perperoglou et al., BMC Med Res Methodol 2019
Sauerbrei et al., Diagn Progn Res 2020

Measurement error correction methods: Imputation methods

- Regression-calibration

- ① Regression: estimate conditional expectation of X given X^*
- ② Calibration: substitute it for X in logistic regression

Rosner et al., Stat Med 1989

- Multiple imputation

- ▶ Conditional expectation of X given X^* and Y + imputed value of the regression residual
- ▶ 10 imputations

Cole et al., Int J Epidemiol 2006

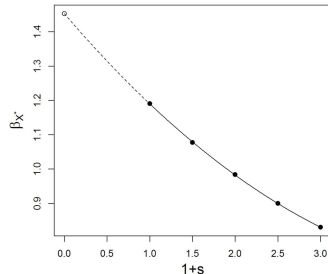
Measurement error correction methods: SIMEX

- Simulation-extrapolation

- 1 **Simulation**: new values of X^* with increasing measurement error $\rightarrow$ estimate parameter of interest
- 2 **Extrapolation** back to the case of no error

Cook & Stefanski, JASA 1994

- ▶ 2 variants: apply SIMEX to
 - ✓ “Pointwise”: Individuals points on the curve
 - ✓ “Coefficient-based”: Estimated B-spline or FP coefficients (P-splines not feasible)



Keogh et al., Stat Med 2020

Measurement error correction methods: Bayes

- Bayesian approach

- ▶ 3 models

- ✓ Outcome model for Y given X
 - ✓ Exposure model for X
 - ✓ Measurement error model for X^* given X
 - + Prior distributions for relevant parameters

⇒ Joint distribution

$$f(Y, X, X^*) = f(Y|X)f(X^*|X)f(X)$$

under non-differential ME

- ▶ Sampling using Hamiltonian Monte Carlo algorithm in Stan (Betancourt & Girolami, arXiv 2013)
 - ▶ 2 posterior inferences
 - ✓ “Risk”: Logit of posterior mean of risk on $P(Y = 1)$ scale
 - ✓ “Logit”: Posterior mean on $\text{logit}(P(Y = 1))$ scale

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

Performance measures

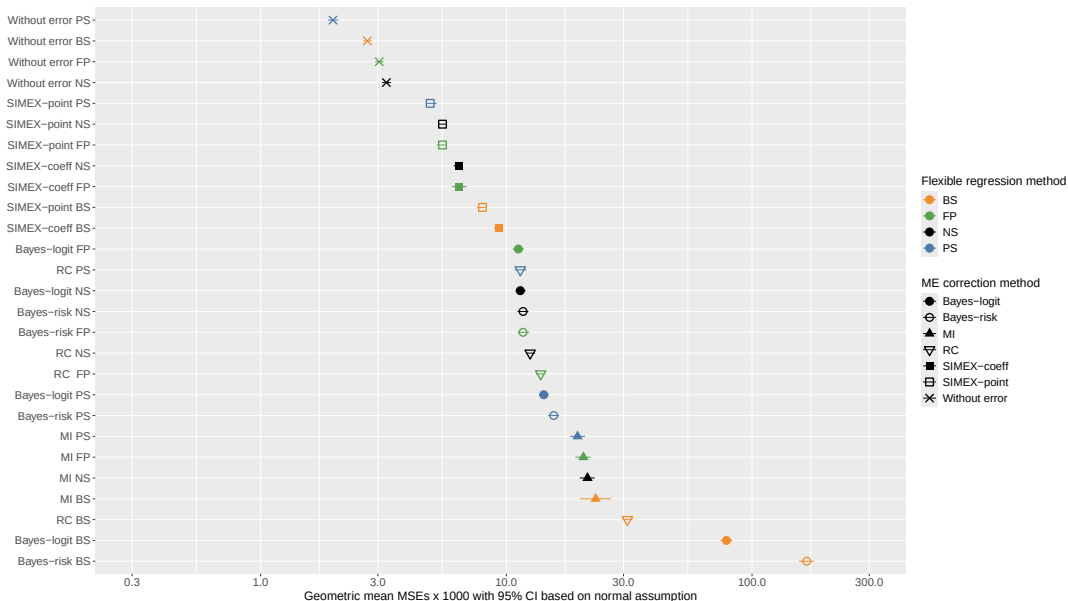
- 23 analytic methods within 6 families of ME correction methods
- For each method $\times$ dataset $\times$ scenario, log Mean Squared Error calculated as

$$\log \text{MSE} = \log \left\{ \frac{\sum_{j=1}^m \left(\widehat{f^*}(x_j) - f^*(x_j) \right)^2 \mathbf{1}_{\{x_{\text{low}} \leq x_j \leq x_{\text{high}}\}}}{\sum_{j=1}^m \mathbf{1}_{\{x_{\text{low}} \leq x_j \leq x_{\text{high}}\}}} \right\}$$

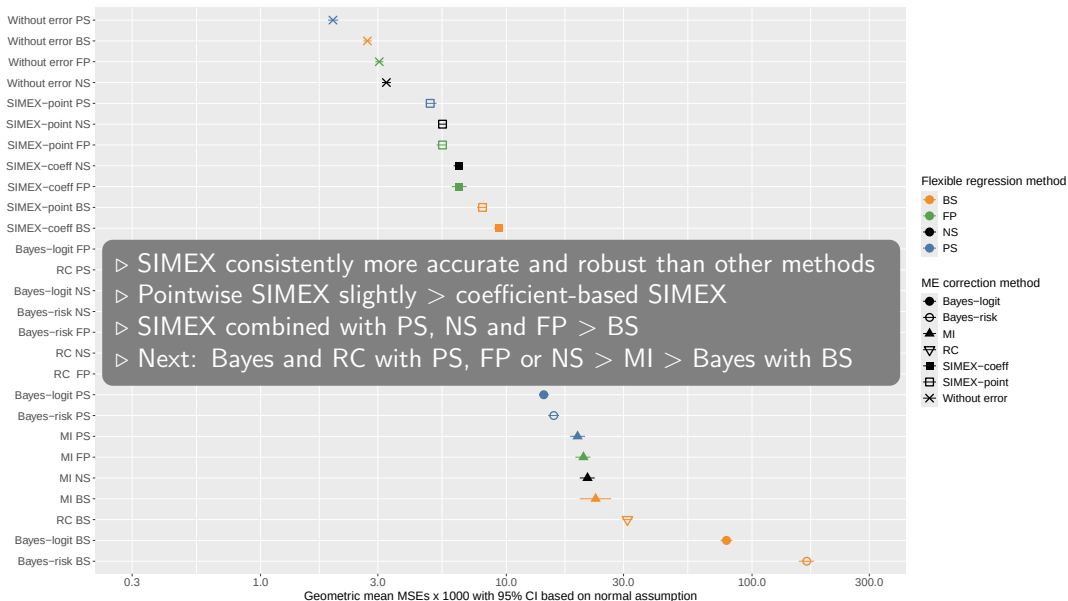
where $\mathbf{1}_{\{x_{\text{low}} \leq x_j \leq x_{\text{high}}\}}$ restricts calculation within 2.5th and 97.5th percentiles of the distribution of true X (limit points undisclosed to Methods teams)

- Overall performance as geometric mean over scenarios
- Secondary analyses: linear models of log MSEs to investigate interactions between methods and design parameters

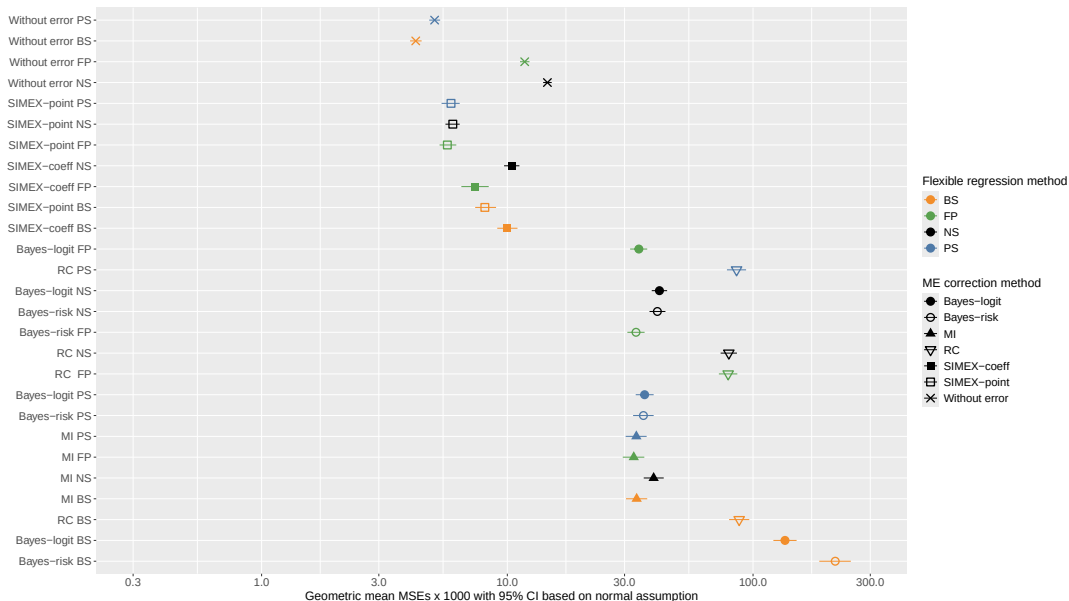
Results: All Stage 3 scenarios combined



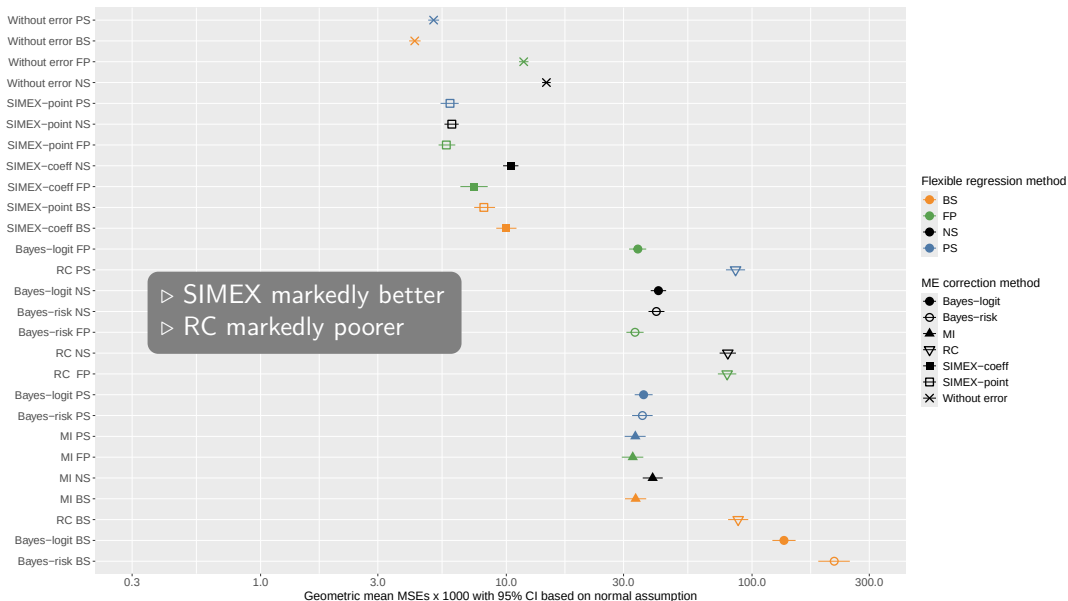
Results: All Stage 3 scenarios combined



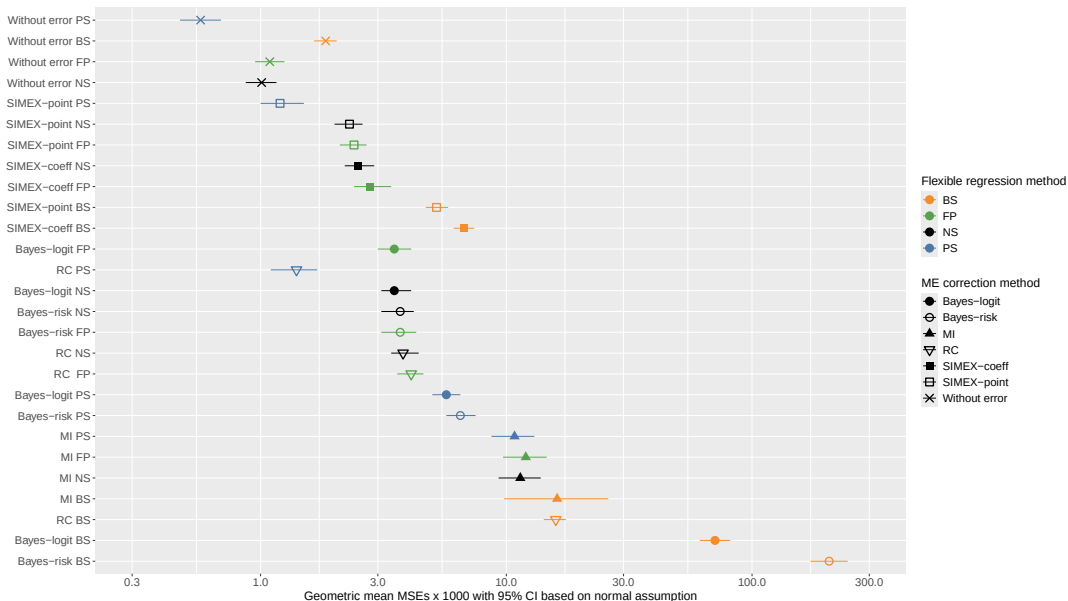
Results: J-shape scenarios



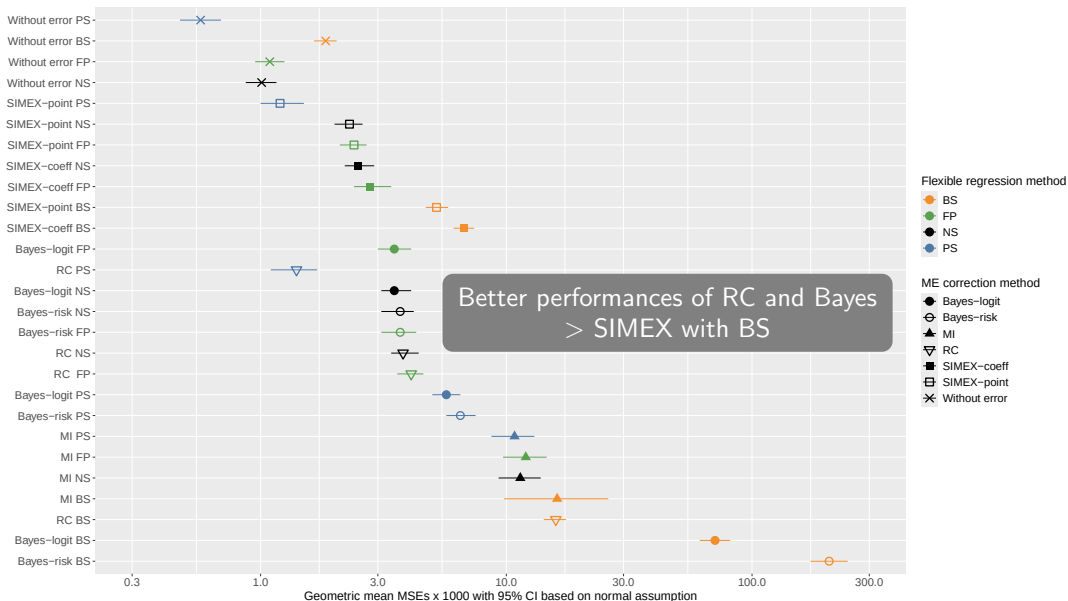
Results: J-shape scenarios



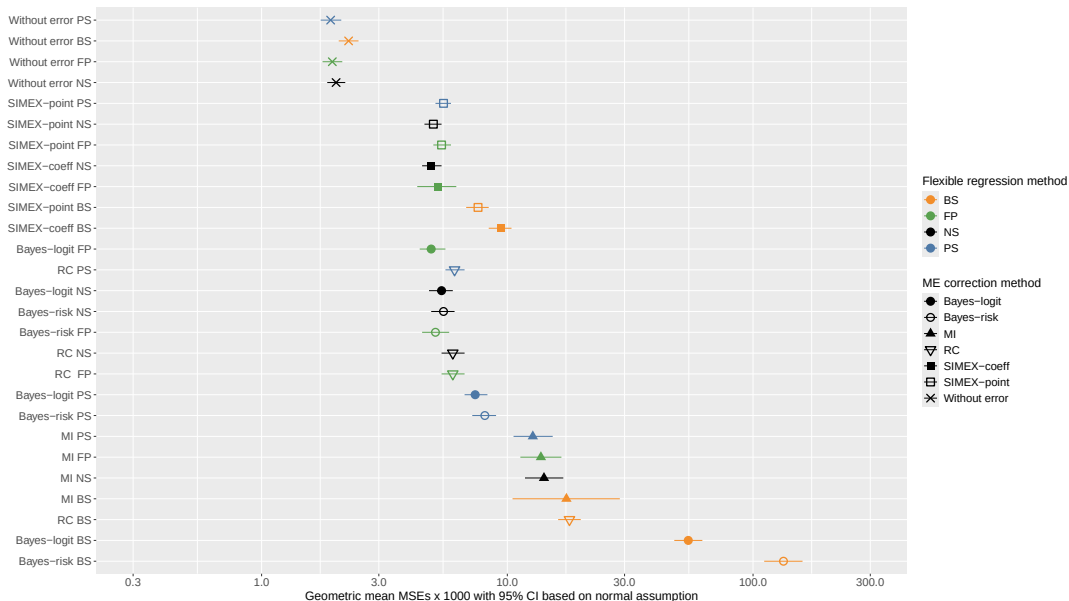
Results: Linear scenarios



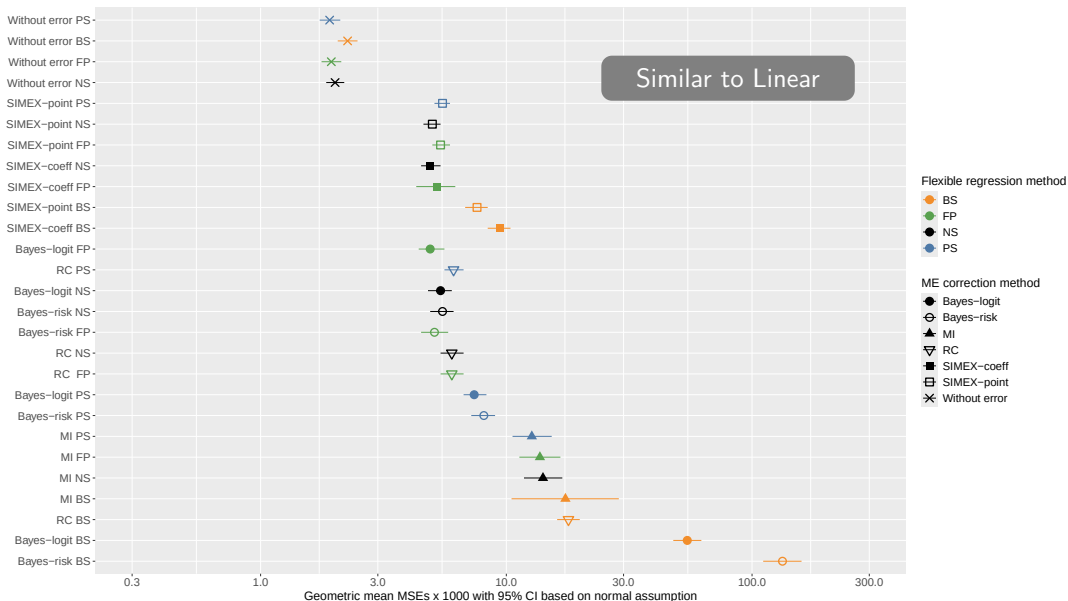
Results: Linear scenarios



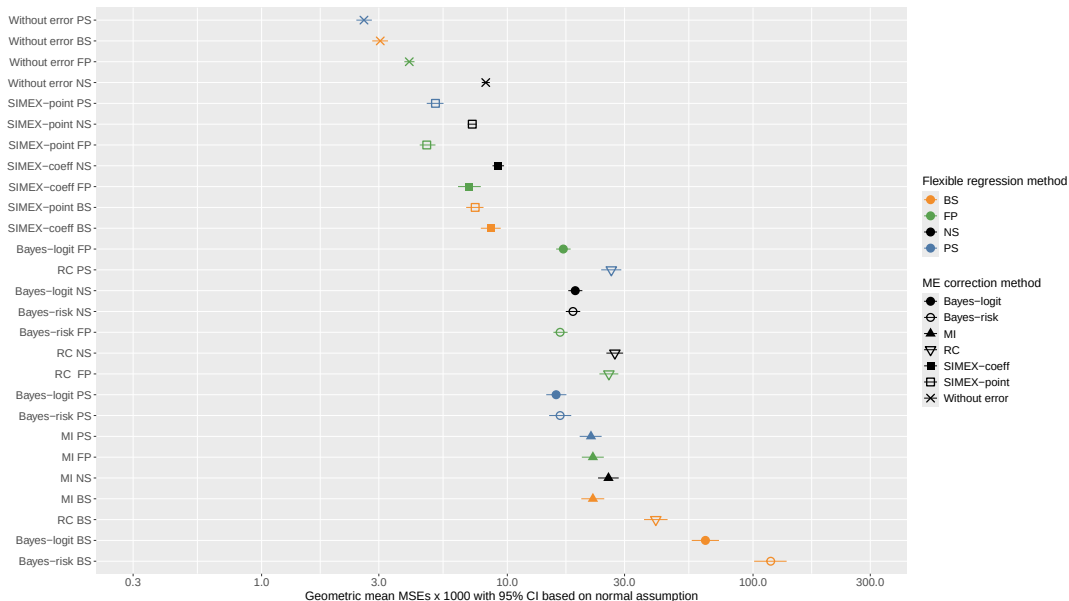
Results: Threshold below median and linear increase scenarios



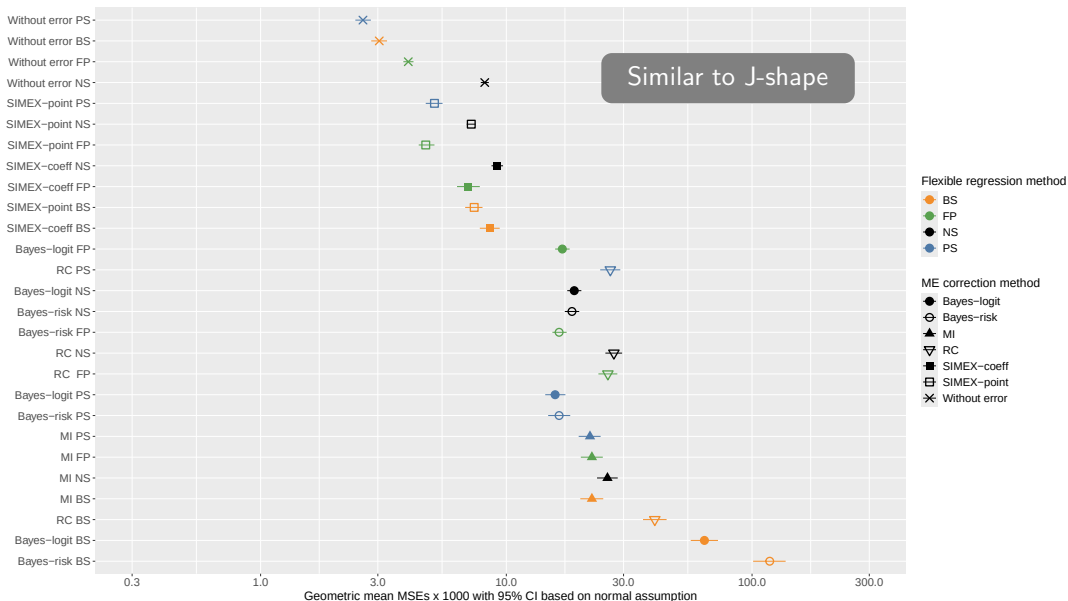
Results: Threshold below median and linear increase scenarios



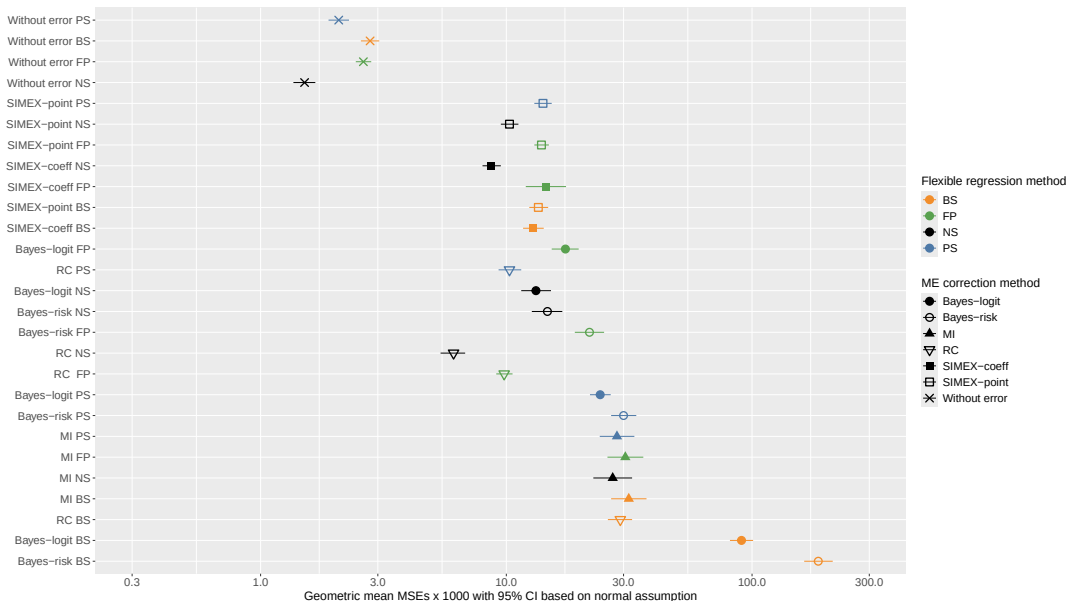
Results: Threshold above median and exponential increase scenarios



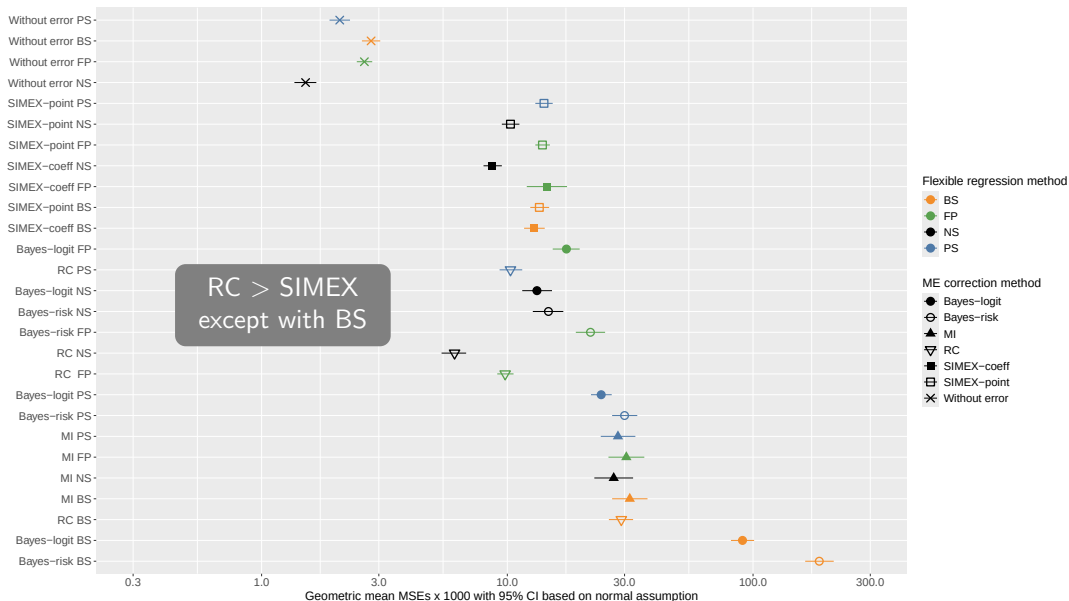
Results: Threshold above median and exponential increase scenarios



Results: Saturation scenarios



Results: Saturation scenarios



Discussion: Simulation results

- Surprising results

- ▶ Top performer: (pointwise) SIMEX vs. expected Bayes and $MI > RC > SIMEX$
- ▶ B-splines consistently exhibited highest MSEs

⇒ Neutral comparison design was crucial!

- Further investigations

- ▶ Leave-one-out influence analyses and extreme simulation scenarios (smaller samples)
- ▶ Accuracy of recovering specific features of the curve (nadir, changepoint)

- Limitations of simulation plan and possible extensions

- ▶ Single error-prone covariate → confounders?
- ▶ Nondifferential classical ME → more complex ME?
- ▶ Other flexible modeling or ME correction approaches?

Discussion: Study design

- **Neutral comparison design** advocated to provide unbiased, controlled comparison of competing statistical methods (Boulesteix et al., Biom J 2024)
 - ▶ Blind analysis → avoid confirmation bias
 - ▶ Adversarial collaboration of independent international teams (Cowan et al., Persp Psy Sci 2020)
 - ▶ Strong commitment of a chair person
- **Choice of simulation scenarios**
 - ▶ Based on examples from real health studies
 - ▶ Not driven by expected performances, except linear scenario → prevent selective reporting
 - ▶ Some variations in Stage 2 motivated by intuition that one method may perform more poorly
- **Limitations**
 - ▶ Natural splines introduced after unblinding
 - ▶ Flexible modeling methods not fully standardized across Methods teams
 - ▶ Each competing method implemented by a single, most expert team → multiple, less expert teams?

Conclusion

- Value of neutral comparison

- ▶ Blinded, unbiased studies are crucial for rigorously evaluating statistical methods, much like clinical trials in medicine
- ▶ This experiment demonstrates their feasibility through a large collaborative project

- Take-home message

- ▶ Adjusting for measurement error is critical
- ▶ The choice of method has a profound impact
- ▶ This study provides some evidence to guide researchers, still further work is needed

Project members



- Data Generation and Evaluation team

- ▶ TG4: Laurence Freedman, Anne Thiébaud
- ▶ TG2: Aris Perperoglou
- ▶ Affiliate: Mohammed Sedki

- SIMEX team

- ▶ TG2: Michal Abrahamowicz
- ▶ Affiliate: Steve Ferreira Guerra

- Imputation team

- ▶ TG4: Kevin Dodd, Victor Kipnis, Douglas Midthune
- ▶ Affiliates: Brian Barrett, Matthew Chaloux, Amer Moosa

- Bayesian team

- ▶ TG4: Raymond Carroll, Paul Gustafson
- ▶ TG2: Frank Harrell
- ▶ Affiliate: Nadja Klein

Further details

● TG2

- ▶ Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. *BMC Medical Research Methodology* 2019; 19(1): 46
- ▶ Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell FE Jr, Royston P, Heinze G. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagnostic and Prognostic Research* 2020; 4: 3

● TG4

- ▶ Keogh RH, Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Küchenhoff H, Tooze JA, Wallace MP, Kipnis V, Freedman LS. 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* 2020; 39(16): 2197–2231
- ▶ Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Keogh RH, Kipnis V, Tooze JA, Wallace MP, Küchenhoff H, Freedman LS. 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* 2020; 39(16): 2232–2263

● TG2 & TG4

- ▶ Ferreira Guerra S, Freedman L, Gustafson P, Sauerbrei W, Harrell FE Jr., Abrahamowicz M. Assessing the impact of measurement error on fractional polynomials and spline estimates of non-linear dose-response functions: A simulation study. Provisionally accepted in *Biometrical Journal*
- ▶ Thiébaud ACM, Perperoglou A, Sedki M, Ferreira Guerra S, Gustafson P, Harrell FE Jr., Sauerbrei W, Abrahamowicz M, Freedman LS. Methods for adjusting for covariate measurement error in flexible modelling of functional form: designing a blinded, controlled neutral comparison simulation study. *Biometrical Journal* 2026; 68(4): e70155 – arXiv:2606.02130
- ▶ Sedki M, Perperoglou A, Thiébaud ACM, Ferreira Guerra S, Gustafson P, Harrell FE Jr., Sauerbrei W, Abrahamowicz M, Freedman LS. Methods for adjusting for covariate measurement error in flexible modelling of functional form: results of a blinded, controlled neutral comparison simulation study. Submitted – arXiv:2606.10123

Data generation: Stage 2 – larger sample sizes

Combination	Main study N	Error distribution	Variance ratio	Replication n
Stage 1	15,000	Normal	0.5	250
1	30,000	Normal	0.5	250
2	15,000	Normal	1	250
3	30,000	Normal	1	250
4	15,000	Shifted Gamma	0.5	250
5	30,000	Shifted Gamma	0.5	250
6	15,000	Shifted Gamma	1	250
7	30,000	Shifted Gamma	1	250
8	15,000	Normal	0.5	750
9	30,000	Normal	0.5	750
10	15,000	Normal	1	750
11	30,000	Normal	1	750
12	15,000	Shifted Gamma	0.5	750
13	30,000	Shifted Gamma	0.5	750
14	15,000	Shifted Gamma	1	750
15	30,000	Shifted Gamma	1	750

Shifted Gamma distribution with shape parameter 3, scaled to have the target variance

Data generation: Stage 2 – smaller sample sizes

Combination	Main study N	Error distribution	Variance ratio	Replication n
16 (comb1)	5,000	Normal	0.5	250
17 (comb2)	2,000	Normal	0.5	250
18 (comb3)	5,000	Normal	1	250
19 (comb4)	2,000	Normal	1	250
20 (comb5)	5,000	Shifted Gamma	0.5	250
21 (comb6)	2,000	Shifted Gamma	0.5	250
22 (comb7)	5,000	Shifted Gamma	1	250
23 (comb8)	2,000	Shifted Gamma	1	250
24 (comb9)	5,000	Normal	0.5	750
25 (comb10)	2,000	Normal	0.5	750
26 (comb11)	5,000	Normal	1	750
27 (comb12)	2,000	Normal	1	750
28 (comb13)	5,000	Shifted Gamma	0.5	750
29 (comb14)	2,000	Shifted Gamma	0.5	750
30 (comb15)	5,000	Shifted Gamma	1	750
Not included	2,000	Shifted Gamma	1	750

Shifted Gamma distribution with shape parameter 3, scaled to have the target variance