A blinded, controlled comparison of methods for adjusting for covariate measurement error in regression modelling

A joint project of TG2 (Selection of Variables and Functional Forms) and TG4 (Measurement Error and Misclassification) of the STRATOS Initiative

Aris Perperoglou, Mohammed Sedki, Anne Thiébaut and Laurence Freedman
on behalf of TG2-TG4

Members of the TG2-4 Subgroup

TG2:	**Michal Abrahamowicz, Frank Harrell, Aris Perperoglou, Willi Sauerbrei**
TG4:	**Kevin Dodd, Raymond Carroll, Laurence Freedman, Paul Gustafson, Victor Kipnis, Douglas Midthune, Anne Thiebaut**
Affiliates:	**Brian Barrett, Matthew Chaloux, Nadja Klein, Amer Moosa, Mohammed Sedki, Steve Ferreira Guerra**

Outline

- TG2 – TG4 Partnership / Functional Forms & Measurement Error
- The project protocol
- Simulation
- Results
- Discussion

A joint project between TG2 and TG4

TG2

Selection of variables and functional forms in multivariable analysis

Aim: Derive guidance for variable and function selection in multivariable analysis.

Main focus: identify influential variables and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) main challenges are selection of variables for inclusion in a multivariable explanatory model, and **choice of functional forms** for continuous variables

TG4

Measurement error and misclassification

Aim: Increase awareness of problems caused by **measurement error and misclassification** in statistical analyses and remove barriers to use statistical methods that deal with such problems.

Key messages: Considering measurement error is necessary because it may have an impact on the study results.

Special statistical methods are used to account for measurement error.

Additional information is required about the type and size of the measurement error to adjust for measurement error.

Aim of the joint project

We are interested in learning the regression relationship between outcome and covariate(s) when is measured with error.

- Classical Measurement Error Model (CME)
  , whereis random variable with mean 0, independent of and .
- Impact on the regression relationship
  - **Attenuation Bias**: Measurement error leads to attenuation of the estimated regression coefficients.
  - **Loss of Precision**: Increased variance in the estimates. Effective sample size is reduced due to the error variance.
- **When X is not linearly related with Y**: $\mathrm{E}(Y \mid X) = f(X)$.
  - Function is unknown, requiring **flexible estimation methods**
  - Observingmeasured with error distorts the identification of the functional form

Available methods

when X is measured exactly

Popular methods:

- B-splines and natural splines
- P-splines
- Fractional polynomials

when X measured with error and f(X) is linear

Popular methods:

- Regression calibration
- Multiple imputation
- Bayesian estimation
- SIMEX

All these remove bias but do not recover lost precision.

Research objectives

To compare the following methods of estimating f(X) using simulated datasets:

- Regression Calibration
- Multiple Imputation
- Bayes
- SIMEX

X

- B-Splines
- P-splines
- Fractional Polynomials
- Natural Cubic Splines

Stage 1 Blinded Method Development

Data Generation Team → **Shares data** → **3 Methods Teams** → **Submit code** → **Data Generation Team**

Simulates 5 datasets and standardizes spline methods

- Cubic B-splines with one interior knot
- P-splines with 10 interior knots, penalty optimised by methods teams
- Fractional Polynomials (4df)

Create code to apply methods

- Regression Calibration & Multiple Imputation
- Bayes
- SIMEX

Runs code

Evaluates and presents blinded results

Stage 2 Extensive Unblinded Evaluation

Data Generation Team → **Discussion on DGM** → **Data Generation Team**

Simulates variants of 5 datasets

150 unique datasets, building on the 5 original functions but varying key parameters like sample size, measurement error variance, and the error distribution.

Addition of natural cubic splines
Runs code
Evaluates and presents unblinded results

Stage 3 Extension and Replication

Replication → **Data Generation Team**

150 datasets simulated 10 times, creating a total of 1500 datasets

Evaluates and presents unblinded results

Data Generation and Evaluation Team
(Anne Thiebaut, Laurence Freedman, Aris Perperoglou, Mohammed Sedki)

- **Data generation: Binary outcome *Y* linked to continuous *X* by logistic regression**

with undisclosed values or distribution of and undisclosed form of .

In place of , values of * (perturbed by classical measurement error) were provided. Variance and distribution of measurement error were undisclosed, but a subset of replicated values of * were provided.

- **Evaluation of results:** Mean squared error of estimated compared to true evaluated over the central 95% of the distribution of In stages 2,3 logMSE was used

Imputation methods

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

- **Regression calibration** estimates the conditional expectation of the function given the error prone covariate *X** and substitutes it for the true covariate in the logistic regression.
- **Multiple imputation:** The imputed consists of its conditional expectation given X* and Y plus the imputed value of the regression residual. Imputation is done several (usually 10) times using different model parameter values from the corresponding estimated distributions

Bayesian Method
(Paul Gustafson, Raymond Carroll, Frank Harrell, Nadja Klein)

The team specified:

- an outcome model (for *Y* given *X*)
- an exposure model for *X*
- a measurement error model for *X** given *X*
- prior distributions for 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.
- Given a dataset, off-the-shelf MCMC software yields (a Monte Carlo approximation to) this posterior distribution.
- Summaries of the posterior distribution used for inference, e.g., posterior means of parameters in the outcome model are point estimates.

Simulation-Extrapolation (SIMEX)
(Michal Abrahamowicz and Steve Ferreira Guerra)

A 2-step method, Cook and Stefanski (1994), adapted to various measurement error problems Carroll (2006)

General idea

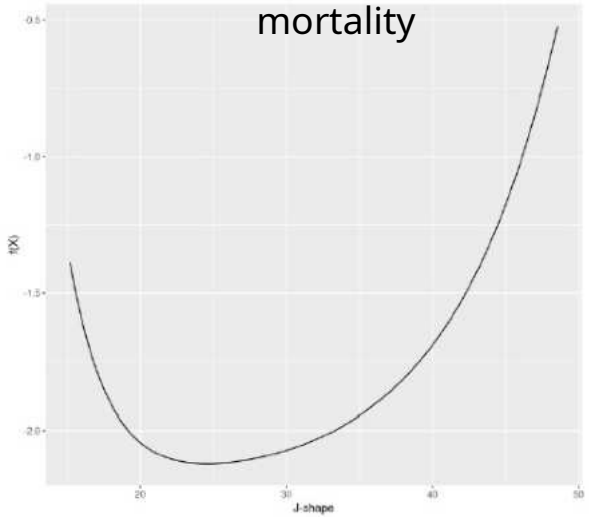
Sequentially **simulate** new variables with increasing measurement error. Use generated variables to estimate parameter of interest; each estimate being increasingly biased. This establishes a relationship between amount of bias and amount of measurement error. Finally, **extrapolate** this relationship back to the case of no error.

For this project, we used two alternative SIMEX approaches:

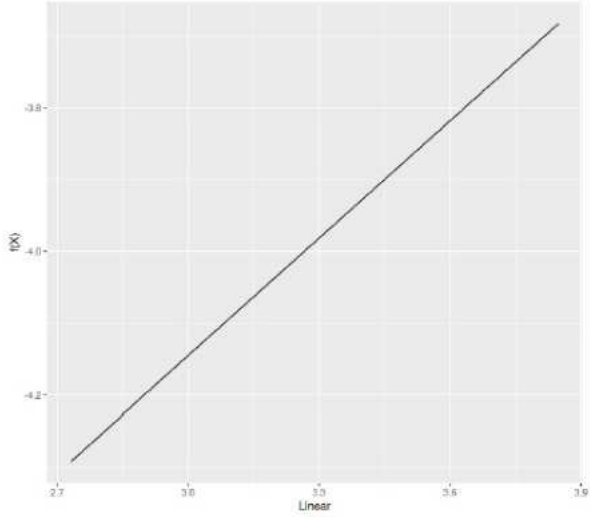
1) Apply SIMEX to the individual points on the curve
2) Apply SIMEX to the B-spline or FP coefficients (not for P-splines)

The forms of *f(X)* used in the simulations

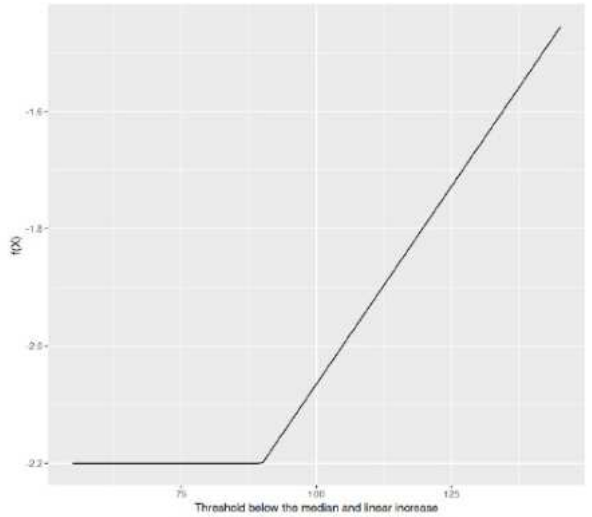
Inspired by the J-shaped relationship between body mass index (BMI) and mortality



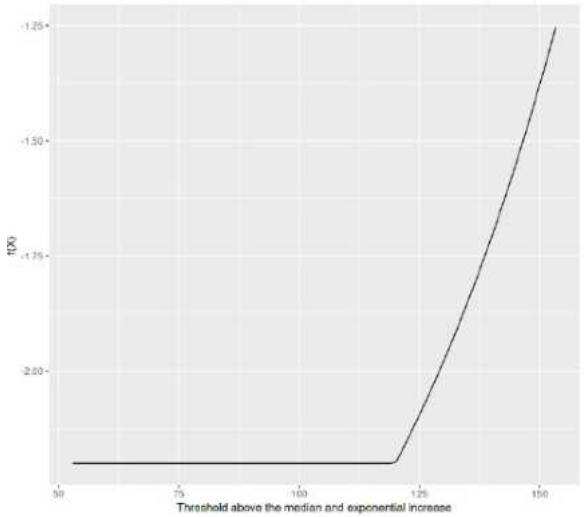
Association reported between dietary fat intake and breast cancer incidence



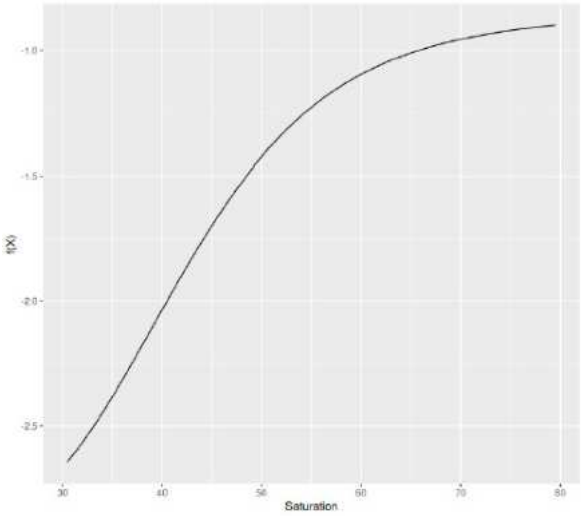
Based on models for air pollution and morta



Based on air pollution models but assumed an exponential increase in risk above the threshold,



Inspired by the Hill equation used in pharmacology to model a drug's dose-response relationship



Blinded results from Stage 1 & Benchmarks

Method	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
A	0.0051	0.00122	0.00518	0.0033	0.0084	**0.0046**
B	0.0034	0.00149	0.00454	0.0039	0.0103	0.0047
C	0.0078	0.00264	0.00278	0.0033	0.0156	0.0064
D	0.0089	0.00250	0.00400	0.0038	0.0143	0.0067
E	0.0058	0.00161	0.00822	0.0065	0.0130	0.0070
F	0.0054	0.00159	0.00893	0.0069	0.0137	0.0073
G	0.0068	0.00236	0.00430	0.0052	0.0223	0.0082
H	0.0081	0.00238	0.00576	0.0043	0.0257	0.0092
J	0.0074	0.00094	0.01079	0.0127	0.0141	0.0092
K	0.0067	0.00098	0.01078	0.0142	0.0131	0.0092
L	0.0082	0.00111	0.00550	0.0161	0.0181	0.0098
M	0.0111	0.00591	0.00445	0.0096	0.0190	0.0100
N	0.0083	0.00088	0.00663	0.0167	0.0184	0.0102
P	0.0106	0.00452	0.00440	0.0140	0.0182	0.0103
Q	0.0101	0.00080	0.00722	0.0150	0.0200	0.0106
R	0.0108	0.00040	0.00683	0.0157	0.0209	0.0109
S	0.0099	0.00073	0.00840	0.0165	0.0207	0.0112
T	0.0108	0.00047	0.00699	0.0160	0.0220	0.0113
U	0.0127	0.00090	0.00555	0.0170	0.0261	0.0124
V	0.0064	0.00097	0.00919	0.0188	0.0339	0.0139
W	0.0060	0.00102	0.01012	0.0166	0.0369	0.0141
X	0.0139	0.00135	0.01397	0.0326	0.0161	0.0156
Y	0.0137	0.00141	0.01457	0.0322	0.0167	0.0157
Z	0.0234	0.00345	0.01085	0.0447	0.0238	0.0212
AA	0.0318	0.00057	0.00597	0.0545	0.0171	0.0220
AB	0.0266	0.00057	0.00596	0.0634	0.0169	0.0227
AC	0.0320	0.00129	0.01277	0.0543	0.0135	0.0228
AD	0.0368	0.00177	0.01193	0.0531	0.0289	0.0265
AE	0.0448	0.00112	0.01355	0.0580	0.0160	0.0311
AF	0.0812	0.00359	0.00627	0.0697	0.0360	0.0394
AG	0.0626	0.00045	0.00646	0.1515	0.0339	0.0518

Two sorts of benchmarks:
1. MSEs based on exact X's (lower bound)
2. MSEs based on unadjusted spline methods

Method	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
Bench-BX	0.0029	0.00160	0.00203	0.0034	0.0040	**0.0028**
Bench-PX	0.0035	0.00008	0.00280	0.0029	0.0035	**0.0026**
Bench-BX*	0.0124	0.00449	0.00594	0.0028	0.0311	**0.0113**
Bench-PX*	0.0101	0.00418	0.00850	0.0023	0.0314	**0.0113**

Simulations

Stage 2

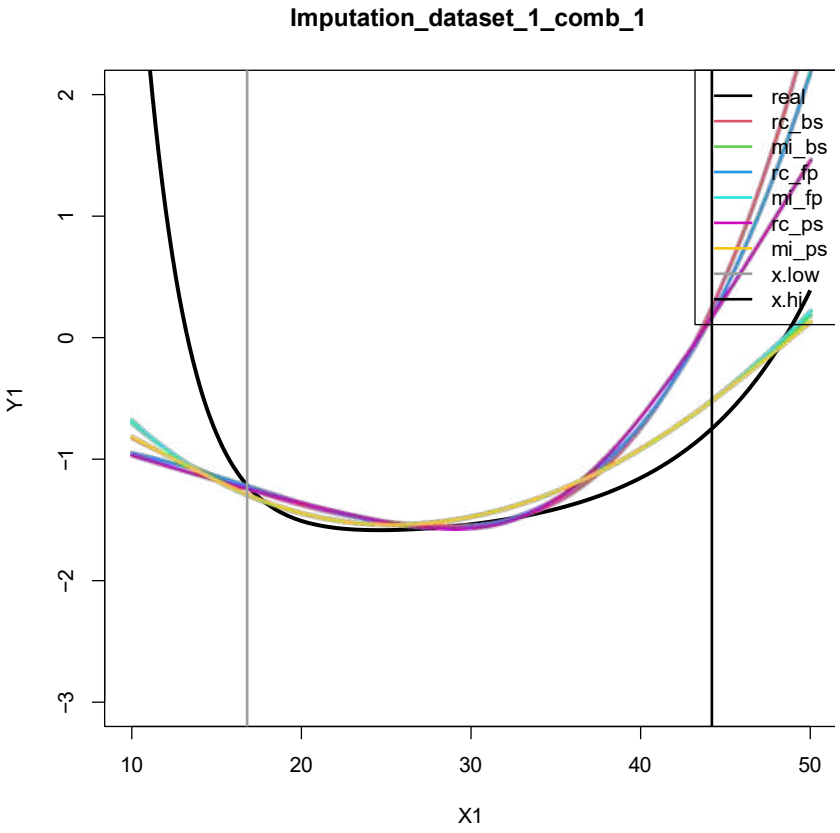
- Same 5 forms of Y-X relationships: logit($P(Y=1|X)$)=$f(X)$
- Main sample sizes: 30000, 15000, 5000, 2000
- Replication substudy sample sizes: 250, 750
- Measurement error variances: 0.5*var(X), 1.0*var(X)
- Error distribution: Normal, Gamma (shape parameter 3) adjusted to have mean 0
- All combinations of above, except the Stage 1 combination, leading to 150 datasets
- Code finalized after Stage 1 used by Data Generation and Evaluation Team to run on all 150 datasets
- Added natural cubic splines.

Stage 3

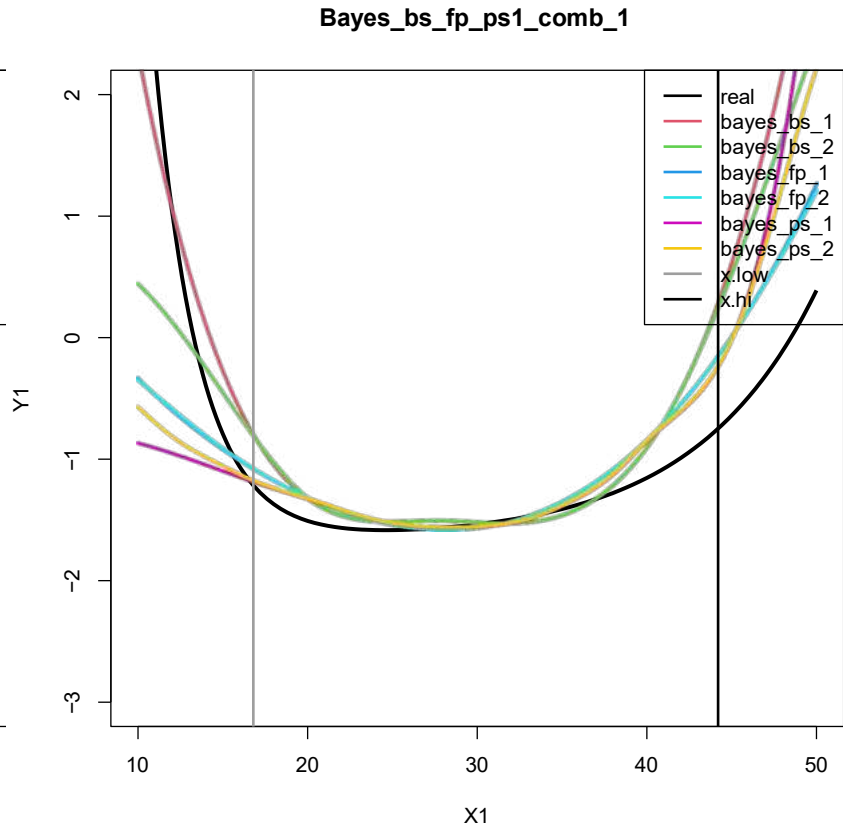
- Each of the 150 datasets was simulated 10 times to provide repeat observations for calculating standard errors and confidence intervals.
- The final comprehensive analysis was based on a total of 1,500 distinct datasets

Selected results Stage 2: Graphs of J-shape

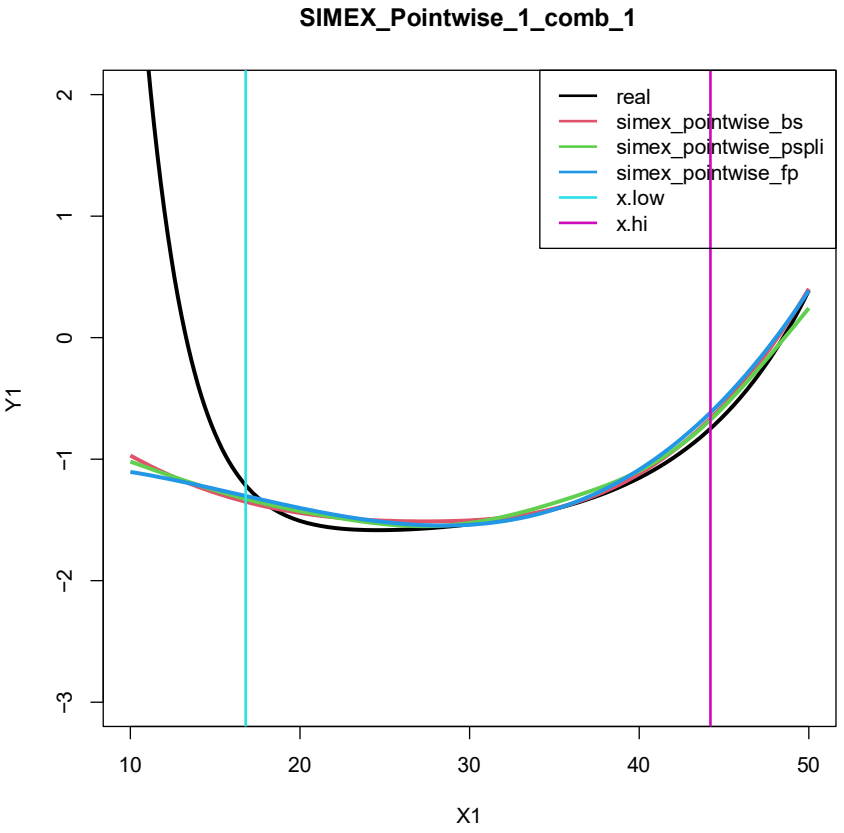
Multiple Imputation and Regression Calibration



Bayes

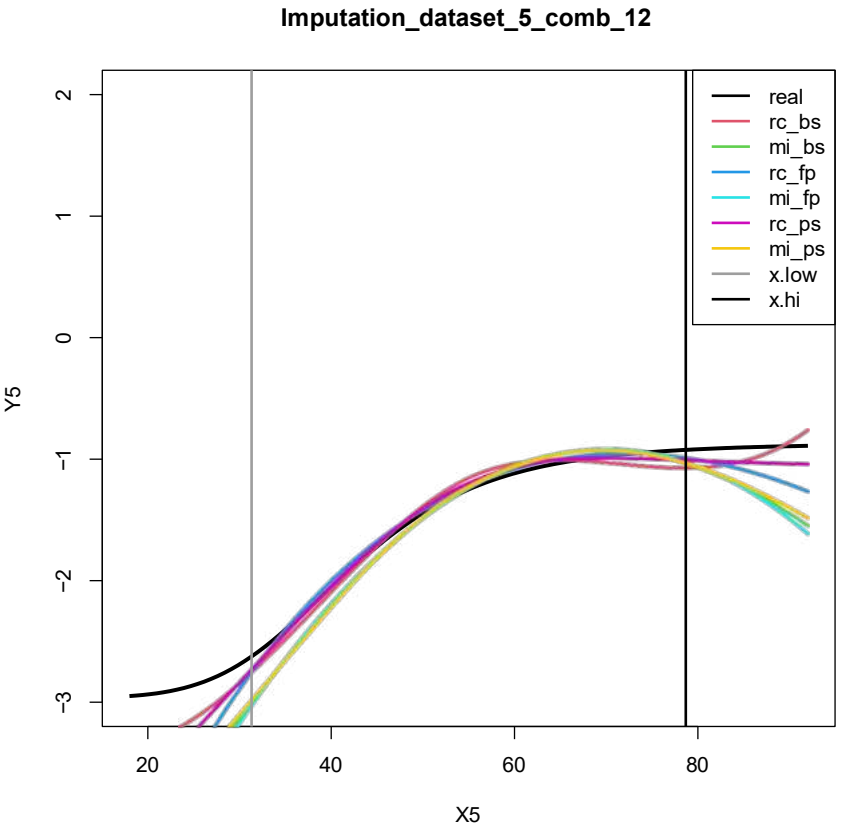


SIMEX (pointwise)

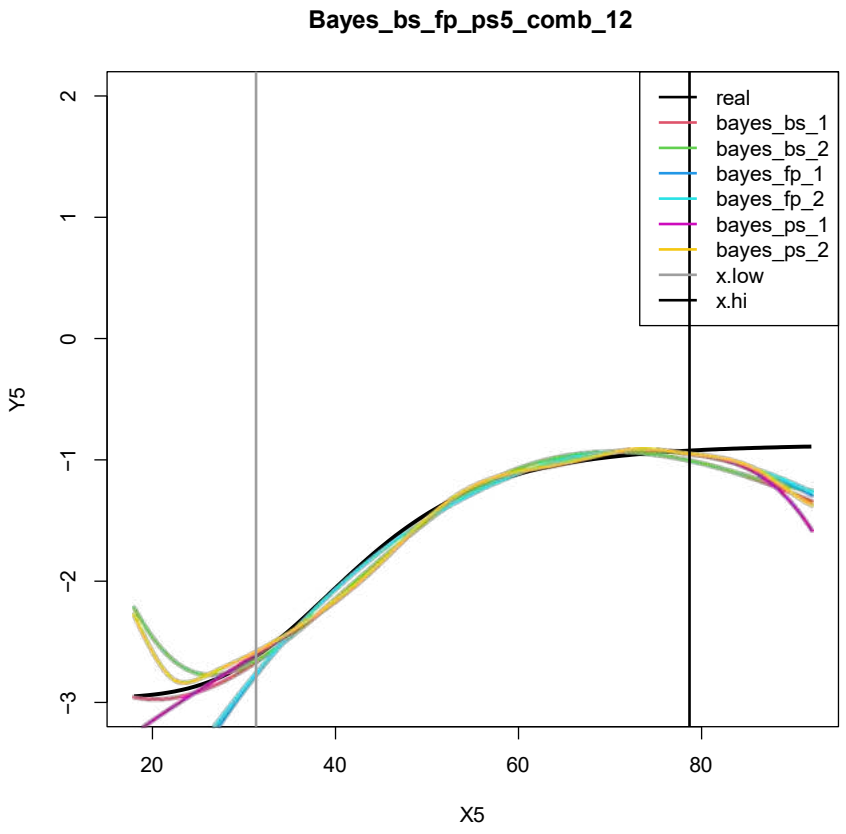


Selected results Stage 2: Graphs of Saturation

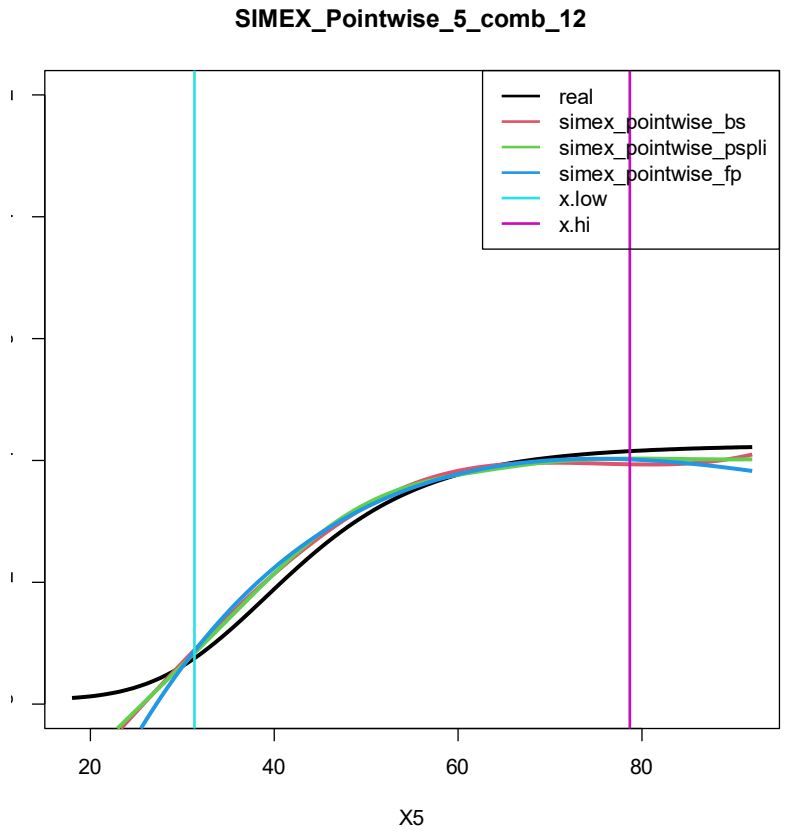
Multiple Imputation and Regression Calibration



Bayes

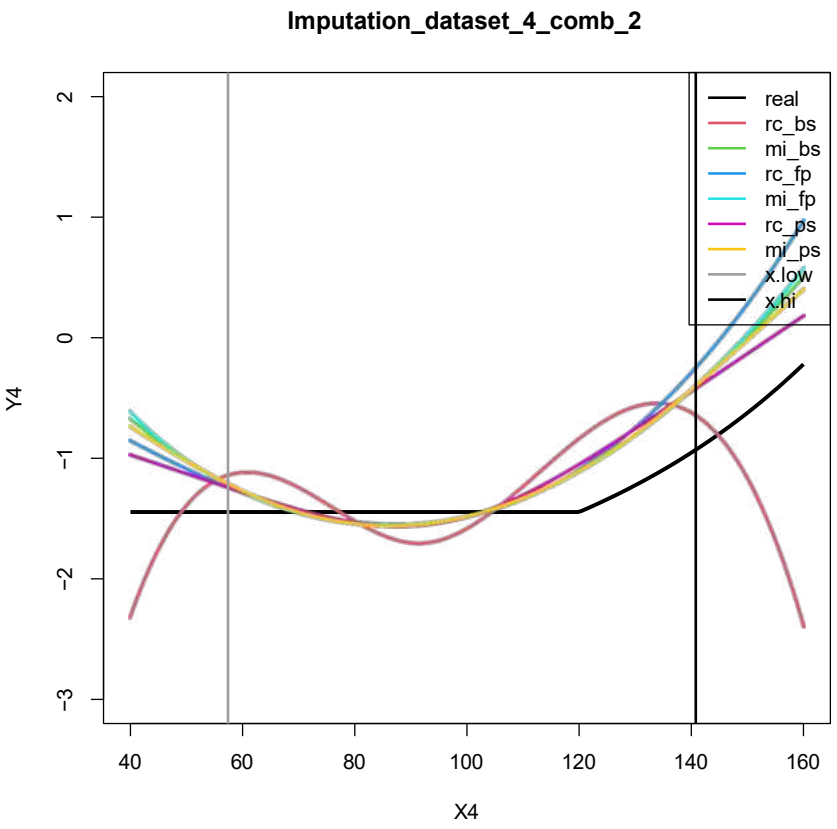


SIMEX (pointwise)

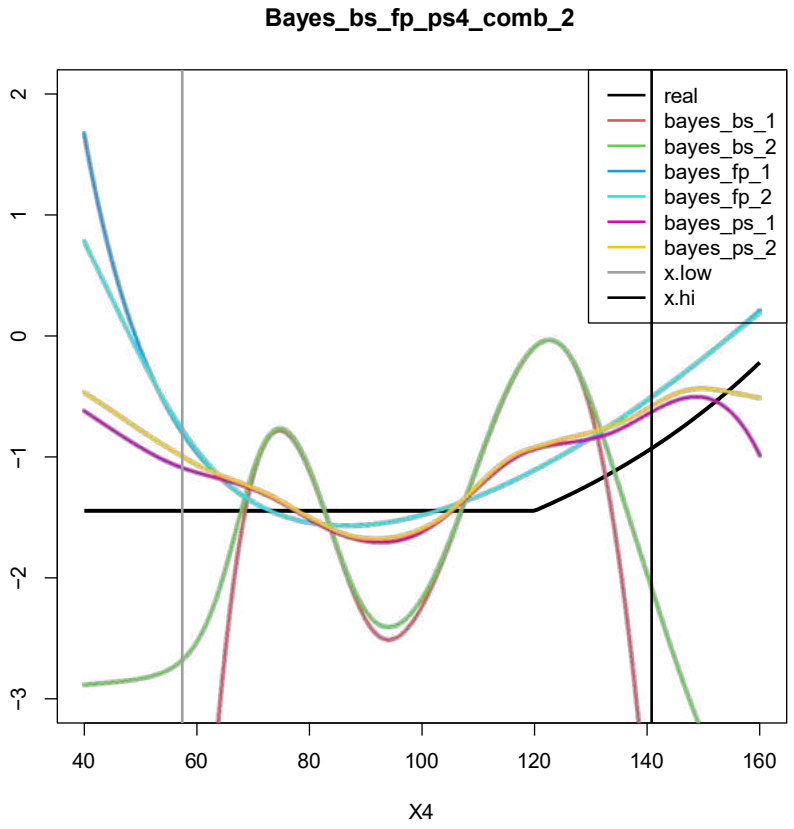


Selected results Stage 2: Graphs of Threshold2

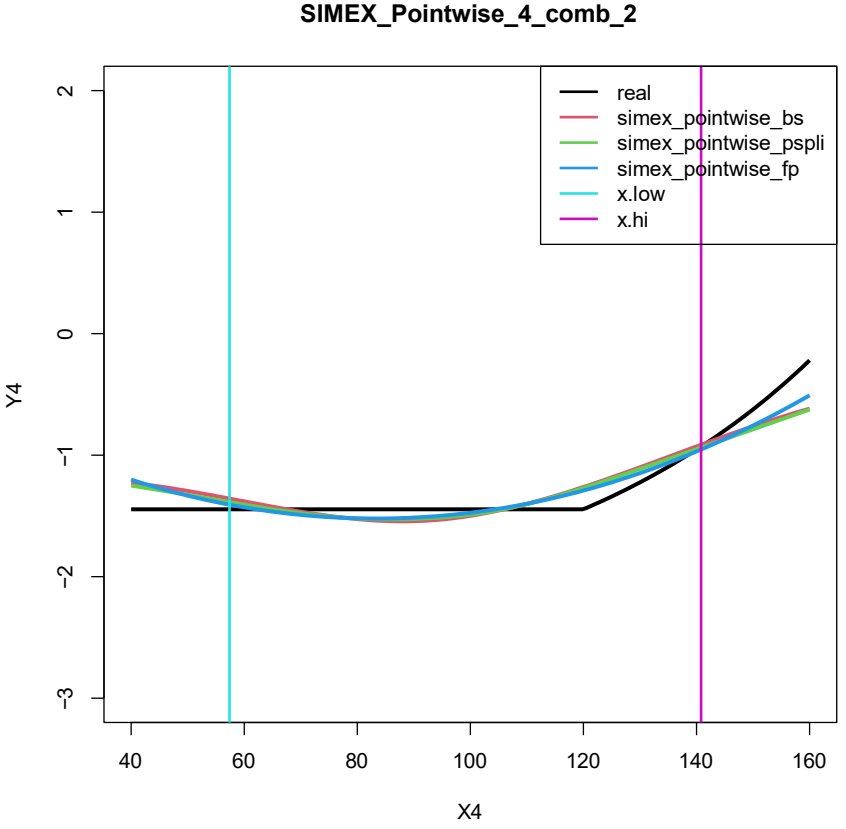
Multiple Imputation and Regression Calibration



Bayes



SIMEX (pointwise)



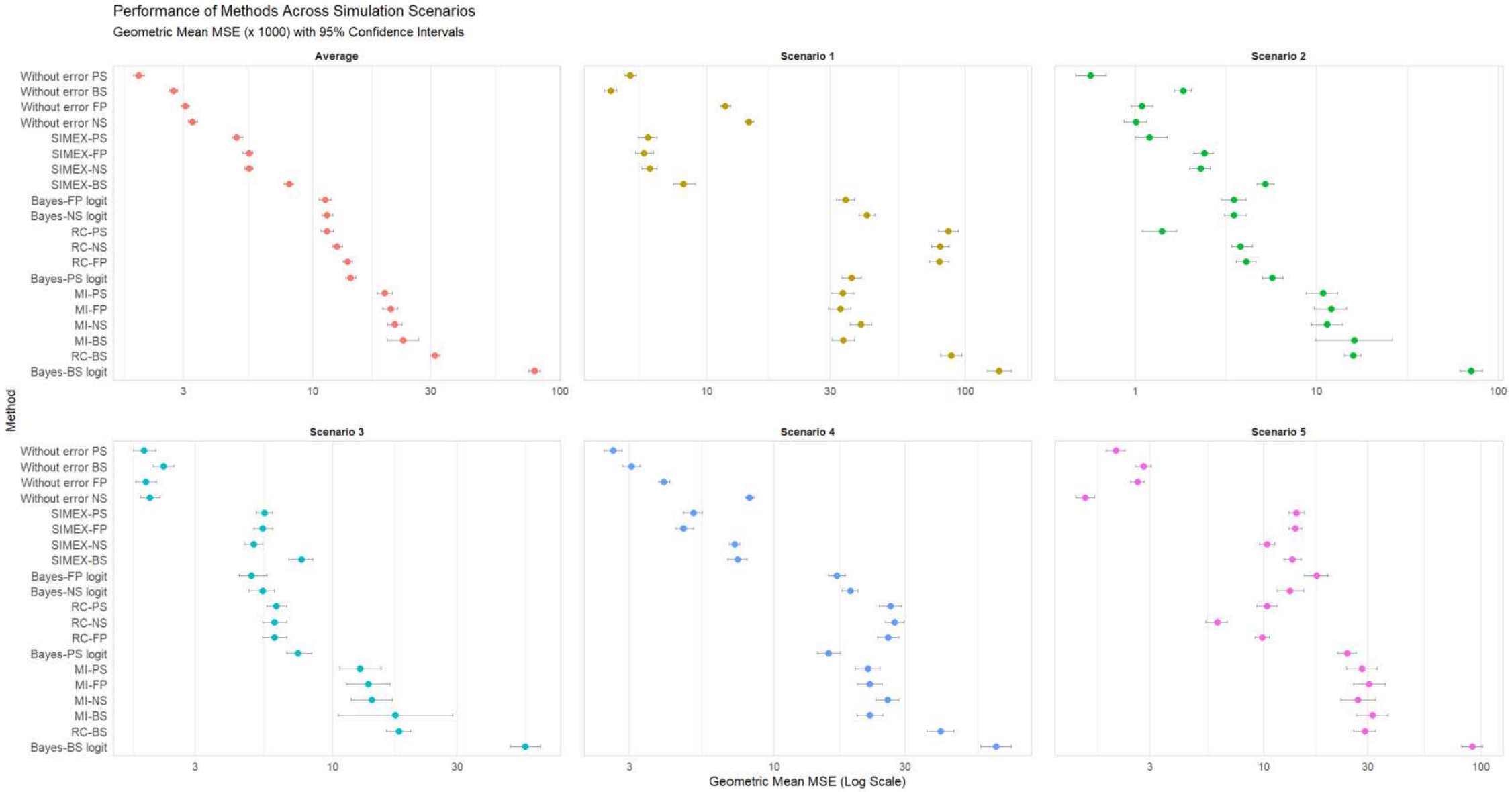
Stage 2 Extension: MSE means over combinations of smaller sample size scenarios

	Datasets 1	Datasets 2	Datasets 3	Datasets 4	Datasets 5	Mean
SIMEX-NS	**0.0148**	**0.0083**	**0.0096**	**0.0098**	**0.0199**	**0.0125**
SIMEX-PS	0.0196	0.0055	0.0084	0.0083	0.0229	0.0129
SIMEX-FP	0.0155	0.0083	0.0103	0.0098	0.0288	0.0145
SIMEX-BS	0.0309	0.0275	0.0165	0.0177	0.0346	0.0254
Bayes-NS logit	**0.0570**	**0.0299**	**0.0192**	**0.0199**	**0.0430**	**0.0338**
Bayes-PS logit	0.0580	0.0178	0.0174	0.0211	0.0583	0.0345
Bayes-FP logit	0.0549	0.0360	0.0261	0.0199	0.0522	0.0378
RC-PS	0.1441	0.0116	0.0119	0.0202	0.0276	0.0431
RC-NS	**0.1347**	**0.0266**	**0.0166**	**0.0327**	**0.0266**	**0.0474**
RC-FP	0.1733	0.0295	0.0170	0.0368	0.0314	0.0576
RC-BS	0.3035	0.1396	0.1163	0.1088	0.1000	0.1536
MI-NS	**0.0811**	**0.0387**	**0.2173**	**0.3864**	**0.2115**	**0.1870**
MI-PS	0.0767	0.0356	0.2486	0.7552	0.1589	0.2550
MI-FP	0.0768	0.0434	0.3831	0.9998	0.2126	0.3431
MI-BS	0.0797	0.0508	0.3438	4.0515	0.2191	0.9490
Bayes-BS logit	0.5421	0.4455	0.9507	0.8475	0.6738	0.6919

Main Results Phase 2

- SIMEX better than {MI, RC, Bayes}
- {P-Spline, FP, Natural spline} better than the B-Spline
- X-Y relationship: Linear > Threshold change-point below median > {Saturation, Threshold change-point above median} > J-shape
- Main study sample size: 30,000 > 15,000 > 5,000 > 2,000; but improvement was greatest between 2000 and 5000.
- Replicate sub-study sample size: 750 > 250, as expected, but not equally for all methods
- Measurement error magnitude: 0.5*Var(X) > 1.0*Var(X), as expected, but not equally for all methods

Selected results Stage 3: logMSE



Key findings

- **Surprising Results:** Our blinded study challenged conventional wisdom, revealing an unexpected performance hierarchy among methods. We suspect that SIMEX might be more robust in complex models.
- **Value of Neutral Comparison:** This work highlights that blinded, unbiased studies are crucial for rigorously evaluating statistical methods, much like clinical trials in medicine.
- **Top Performer:** SIMEX consistently proved to be the most accurate and robust method across most scenarios.
- **Observed Hierarchy:** The general performance ranking was:

  SIMEX > Bayes (with FP/NS) > MI / RC.

- **A Key Caution:** The Bayesian approach, when combined with B-splines, performed poorly and should be used with caution in this context.
- **Take-Home Message:** Adjusting for measurement error is critical. The choice of method has a profound impact, and this study provides some evidence to guide researchers, still further work is needed… Two papers to be made public out of this work…

Thank you & key publications

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and Prognostic Research

COMMENTARY Open Access

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei[1*], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], Georg Heinze[6] and for TG2 of the STRATOS initiative

1. Investigation and comparison of properties of **variable selection strategies**
2. **Comparison of spline procedures** in univariable & multivariable contexts
3. How to model one or more variables with a ‚**spike-at-zero**‘?
4. Comparison of **multivariable procedures for model and function selection**
5. **Role of shrinkage** to correct for bias introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. Adaptation of procedures for **very large sample sizes** needed?

Statistics in Medicine

TUTORIAL IN BIOSTATISTICS

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment

Ruth H. Keogh, Pamela A. Shaw, Paul Gustafson, Raymond J. Carroll, Veronika Deffner, Kevin W. Dodd, Helmut Küchenhoff, Janet A. Tooze, Michael P. Wallace, Victor Kipnis, Laurence S. Freedman

First published: 03 April 2020 | https://doi.org/10.1002/sim.8532 | Citations: 56

TUTORIAL IN BIOSTATISTICS

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics

Pamela A. Shaw, Paul Gustafson, Raymond J. Carroll, Veronika Deffner, Kevin W. Dodd, Ruth H. Keogh, Victor Kipnis, Janet A. Tooze, Michael P. Wallace, Helmut Küchenhoff, Laurence S. Freedman

First published: 03 April 2020 | https://doi.org/10.1002/sim.8531 | Citations: 28