# 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

STRATOS INITIATIVE

# Members of the TG2-4 Subgroup

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

# Conflict of interest disclosure

I have no current or past relationships with commercial entities

# Outline

- Neutral Comparison Simulation Studies
- TG2 – TG4 Partnership / Functional Forms & Measurement Error
- The project protocol
- Results of Stages 1 & 2
- Next steps
- References

# Neutral Comparison Studies

- STRATOS aims to provide guidance on which statistical methods should be used for various types of observational studies
- This guidance requires reliable evidence on the comparative advantages of different competing methods
- Often this evidence is provided by simulation studies
- Biostatisticians teach clinicians the importance of unbiased, controlled evaluation of treatments, but rarely subject their own methods to the same standards
- **Neutral comparison studies** are those that are designed to provide such an unbiased, controlled comparison of competing statistical methods
- This talk describes one such study

# 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

an exposure variable $\boldsymbol{X}$ and an outcome variable $\boldsymbol{Y}$:

$$E(\boldsymbol{Y}|\boldsymbol{X}) = f(\boldsymbol{X})$$

when $\boldsymbol{X}$ is measured with error.

$f(\boldsymbol{X})$ is thought likely to be a non-linear function.

Various statistical methods are available to do this.

Which method(s) should be recommended?

# Methods available when $\boldsymbol{X}$ is measured exactly

**Popular methods (flexible regression)**

B-splines

P-splines

Fractional Polynomials

# When $\boldsymbol{X}$ is measured with error and $f(\boldsymbol{X})$ is linear

Suppose our measurement is $\boldsymbol{X}^* = \boldsymbol{X} + \boldsymbol{U}$, where $\boldsymbol{U}$ is random variable with mean 0, independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (Classical non-differential measurement error)

**Impact on the regression relationship**

- **Attenuation Bias**: Measurement error leads to attenuation of the estimated regression coefficients when usual estimation methods are used that do not account for the error in $\boldsymbol{X}^*$. The estimated coefficient is biased towards zero, reducing its magnitude.
- **Loss of Precision**: Increased variance in the estimates, making them less precise. Effective sample size is reduced due to the error variance.

# Methods available when $\boldsymbol{X}$ is measured with error and $f(\boldsymbol{X})$ is linear

**Popular methods**

Regression calibration

Multiple imputation

Bayesian estimation

Simulation-Extrapolation (SIMEX)

**Note**: All of these remove the bias but usually do not recover lost precision

# Research objectives

To compare the following methods of estimating $f(\boldsymbol{X})$ using simulated datasets:

- Regression Calibration
- Multiple Imputation
- Bayesian Estimation
- SIMEX

X

- B-Splines
- P-Splines
- Fractional Polynomials

# Study structure

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

**Simulates data and standardizes spline methods**

- Cubic B-splines with 1 interior knot at median of observed *X**
- P-splines with 10 interior knots, penalty optimised by methods teams
- Fractional Polynomials of second degree (4df), best FP2 power selected by methods teams

**Create code to apply methods**

- Regression Calibration & Multiple Imputation
- Bayes
- SIMEX

**Evaluates and presents results**

# Data Generation and Evaluation Team

(Anne Thiébaut, Laurence Freedman, Aris Perperoglou, Mohammed Sedki)

- Data generation
  - ✓ Binary outcome $\boldsymbol{Y}$ linked to continuous $\boldsymbol{X}$ by logistic regression. Case-control ratio 1:4

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

  - ✓ Undisclosed values or distribution of $\boldsymbol{X}$ and undisclosed form of $f(\boldsymbol{X})$
  - ✓ In place of $\boldsymbol{X}$, values of $\boldsymbol{X}^*$ ($\boldsymbol{X}$ perturbed by classical measurement error) were provided
  - ✓ Variance and distribution of measurement error were undisclosed, but a subset of replicated values of $\boldsymbol{X}^*$ were provided

- Evaluation of results: Mean squared error of estimated $f(\boldsymbol{X})$ compared to true $f(\boldsymbol{X})$ evaluated over the central 95% of the distribution of $\boldsymbol{X}$

# Imputation Methods Team

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

- **Regression calibration** estimates the conditional expectation of the function $f(\boldsymbol{X})$ given the error prone covariate $\boldsymbol{X}^*$ and substitutes it for the true covariate in the logistic regression

- **Multiple imputation:** The imputed $f(\boldsymbol{X})$ consists of its conditional expectation given $\boldsymbol{X}^*$ and $\boldsymbol{Y}$ 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 Team

(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) Method Team

(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)

# Stage 1: Data, code creation and evaluation

## Data Generation: 5 Datasets

- **Main Study:** N=15,000 independent realizations of a binary outcome $\boldsymbol{Y}$ and a continuous covariate measured with error $\boldsymbol{X}^*$
- **Replication substudy:** sample size 250
- **Measurement error variance:** 0.5*var($\boldsymbol{X}$)
- **Error distribution:** normal

Code generation on distributed "blind data" and Resulting estimates of $f(\boldsymbol{X})$

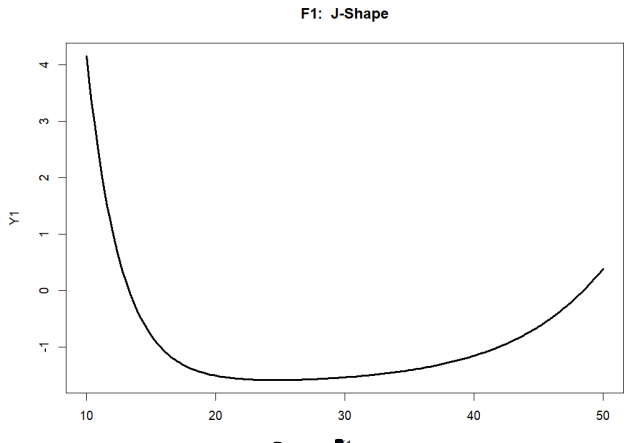
## Evaluation: Mean Squared Error

**Data Generation Team**

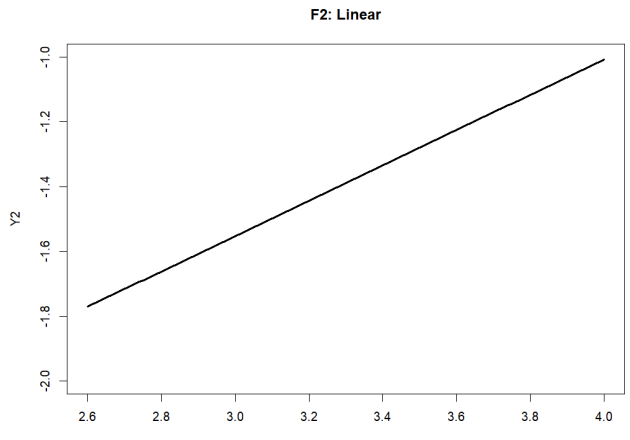
**Methods Teams**

**Data Generation Team**

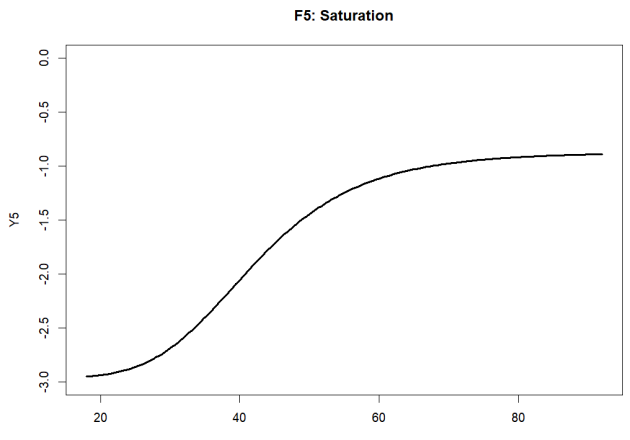
# The forms of $f(\boldsymbol{X})$ used in the simulations



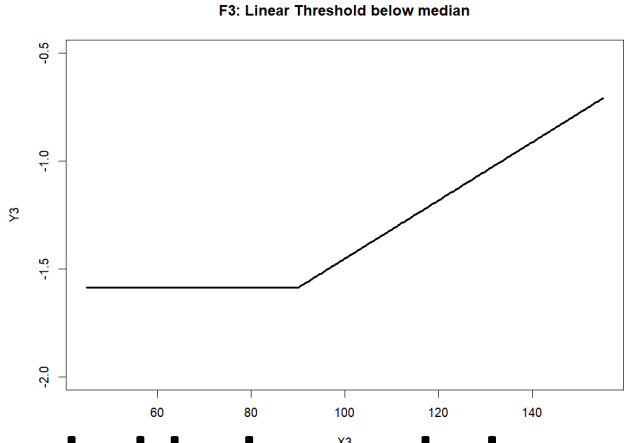

**J-shape**



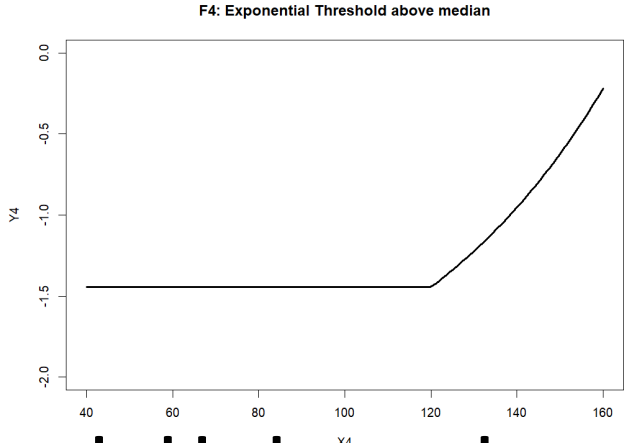

**Linear**




**Saturation**




**Threshold: change below median**




**Threshold: change above median**

# 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 |
| AH | 0.0688 | 0.00417 | 0.01189 | 0.2070 | 0.0400 | 0.0664 |
| AJ | 0.0134 | 0.00187 | 0.14832 | 0.1047 | 0.0868 | **0.0710** |
| AK | 0.0130 | 0.00210 | 0.38618 | 0.1102 | 0.1093 | **0.1242** |

Two sorts of benchmark:

1. MSEs based on exact *X*'s (lower bound)
2. MSEs based on unadjusted spline methods with *X**

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

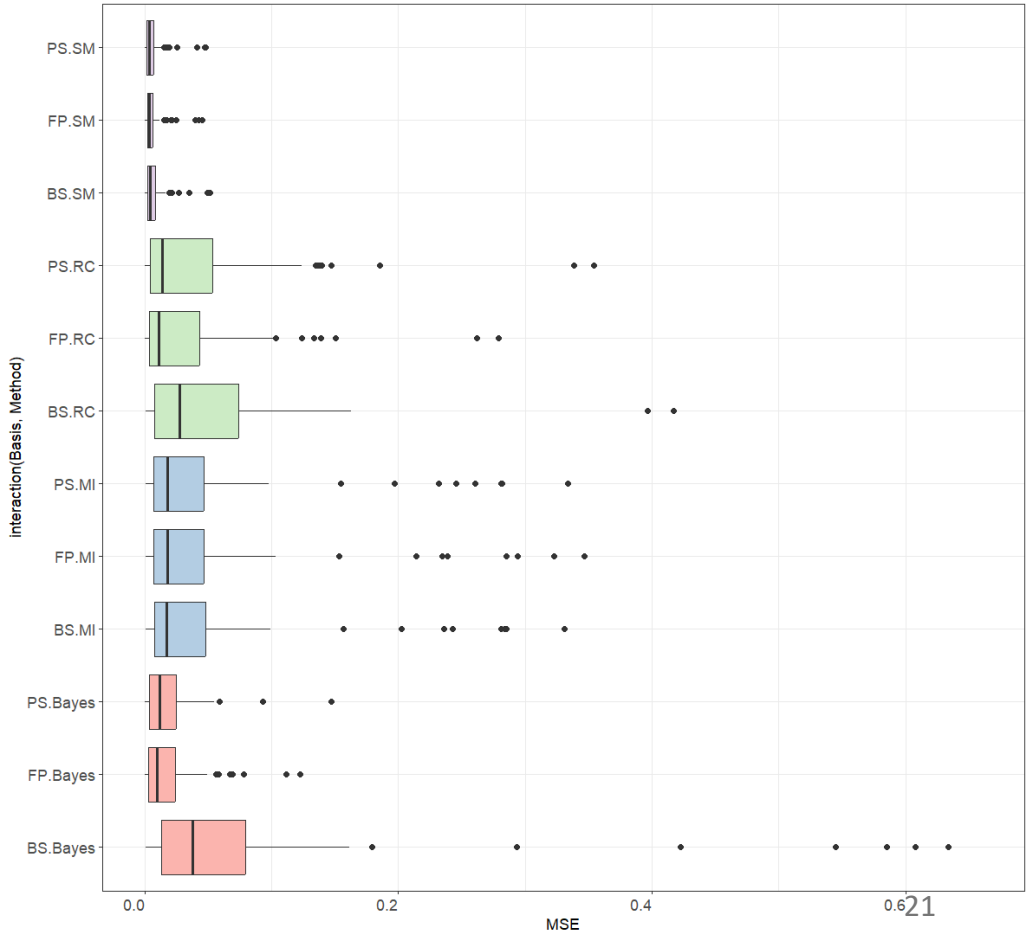
# Stage 2 Simulations

- Same 5 forms of $\boldsymbol{X}$-$\boldsymbol{Y}$ relationships: logit(P($\boldsymbol{Y}$=1|$\boldsymbol{X}$)) = $f(\boldsymbol{X})$
- Main sample sizes: 15000, 30000
- Replication substudy sample sizes: 250, 750
- Measurement error variances: 0.5*var($\boldsymbol{X}$), 1.0*var($\boldsymbol{X}$)
- Error distribution: Normal, Gamma (shape parameter 3) adjusted to have mean 0
- All combinations of above, except the Stage 1 combination, leading to 15 x 5 = 75 datasets: 15 for each of the 5 forms of relationship
- Code finalized after Stage 1 used by Data Generation and Evaluation Team to run on all 75 datasets

# Stage 2 Selected results: MSE

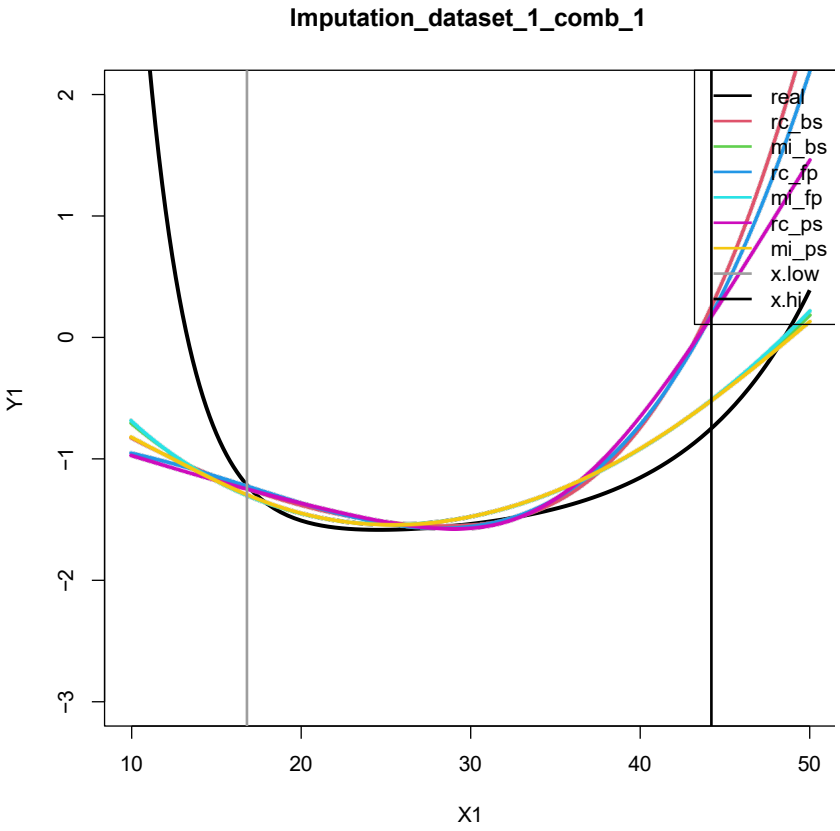
**Key**: **MI - Multiple Imputation**, **RC - Regression Calibration**, **Bayes (logit of posterior mean)**, **SIMEX (Pointwise)**

| Method | Data 1 | Data 2 | Data 3 | Data 4 | Data 5 | Average |
|---|---|---|---|---|---|---|
| **SIMEX-PS** | 0.006 | 0.001 | 0.005 | 0.004 | 0.018 | **0.0065** |
| **SIMEX-FP** | 0.004 | 0.002 | 0.005 | 0.004 | 0.019 | **0.0065** |
| **SIMEX-BS** | 0.006 | 0.004 | 0.006 | 0.005 | 0.016 | **0.0073** |
| **Bayes-FP** | 0.047 | 0.004 | 0.004 | 0.020 | 0.017 | **0.0183** |
| **RC-FP** | 0.120 | 0.003 | 0.005 | 0.038 | 0.012 | **0.0356** |
| **RC-PS** | 0.146 | 0.005 | 0.006 | 0.048 | 0.016 | **0.0440** |
| **MI-PS** | 0.068 | 0.035 | 0.023 | 0.036 | 0.072 | **0.0469** |
| **MI-BS** | 0.068 | 0.036 | 0.025 | 0.036 | 0.073 | **0.0473** |
| **MI-FP** | 0.067 | 0.036 | 0.025 | 0.035 | 0.078 | **0.0481** |
| **RC-BS** | 0.132 | 0.021 | 0.021 | 0.060 | 0.031 | **0.0531** |
| **Bayes-BS** | 0.227 | 0.261 | 0.051 | 0.069 | 0.084 | **0.1383** |

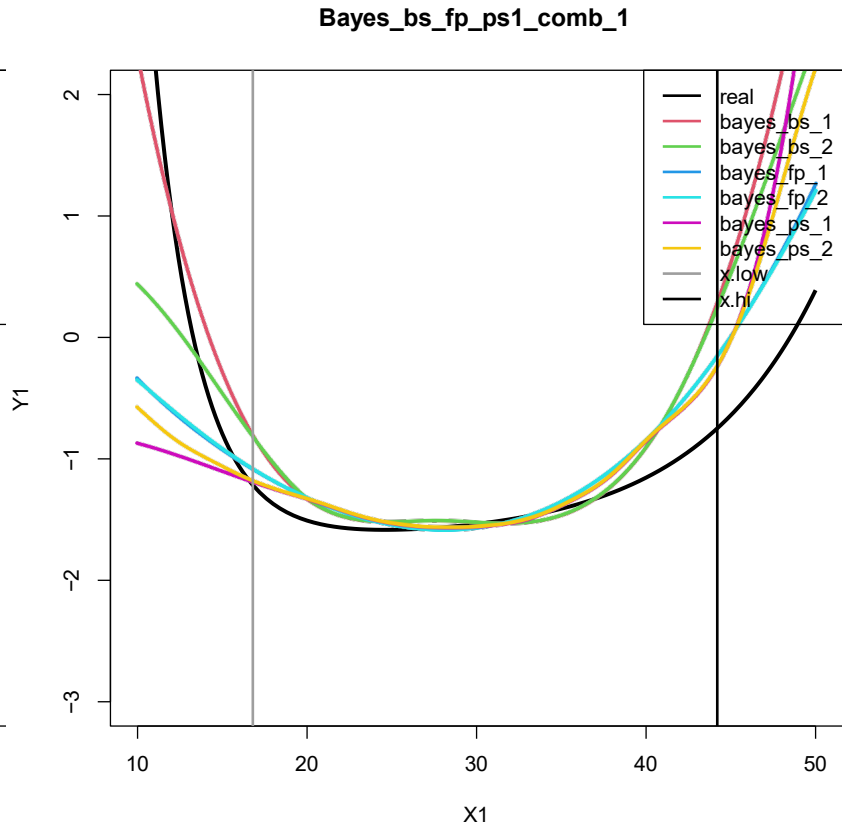


# Stage 2 Selected results: J-shape

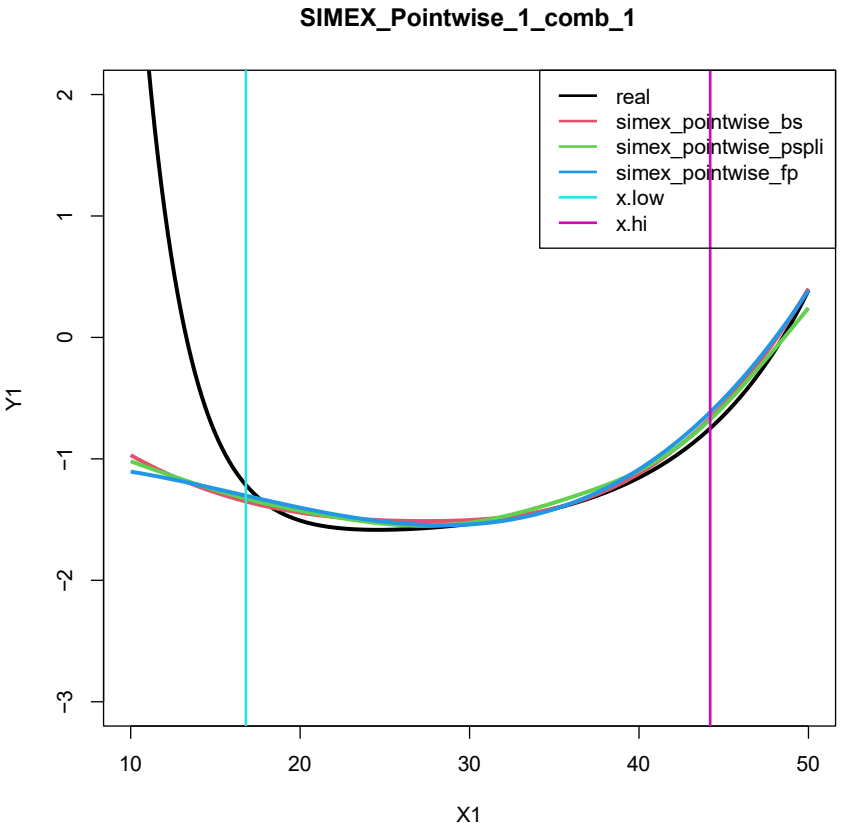
**Multiple Imputation and Regression Calibration**




**Bayes**

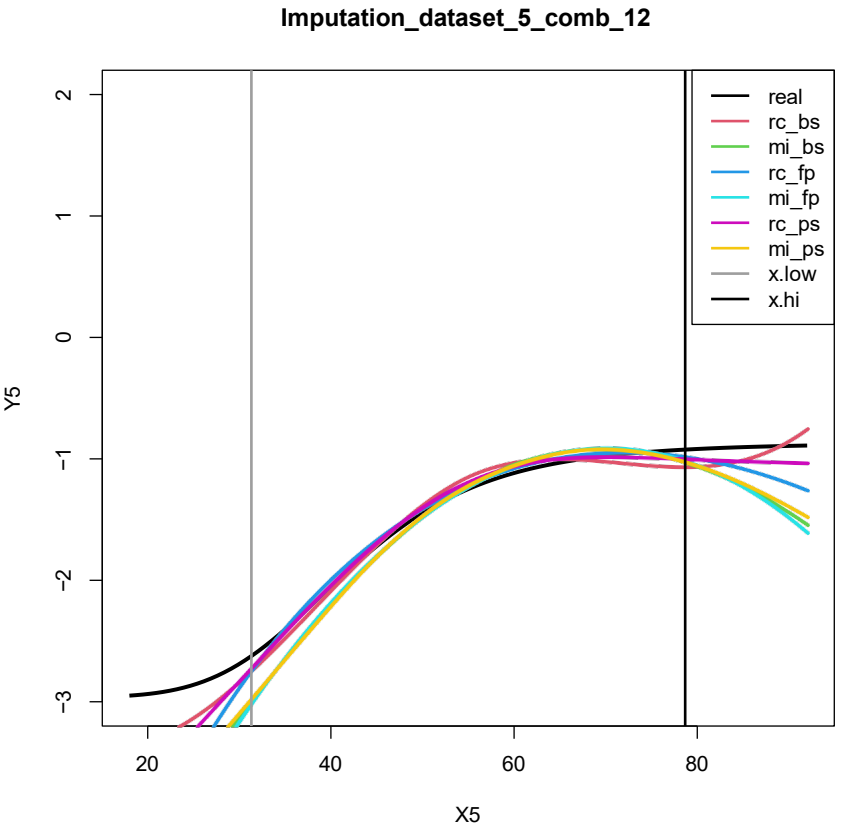



**SIMEX (pointwise)**

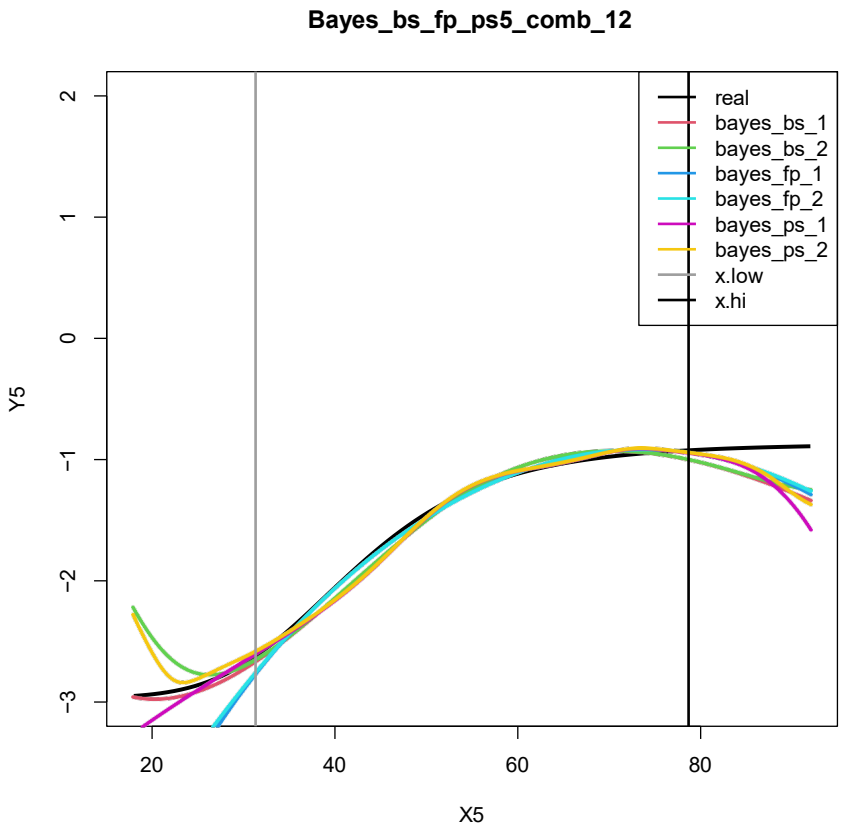


# Stage 2 Selected results: Saturation

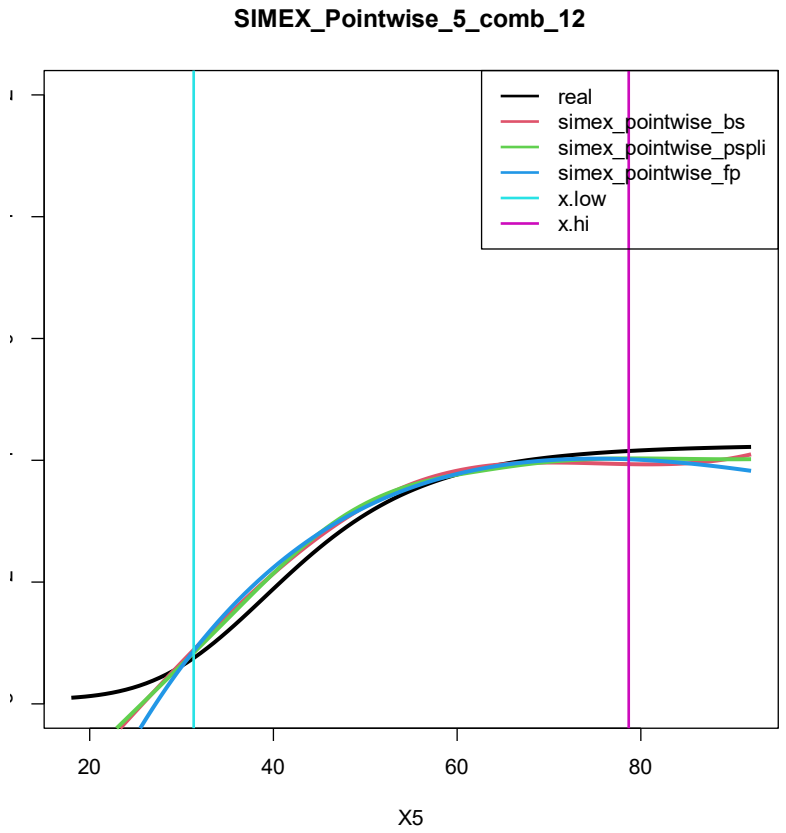
**Multiple Imputation and Regression Calibration**




**Bayes**

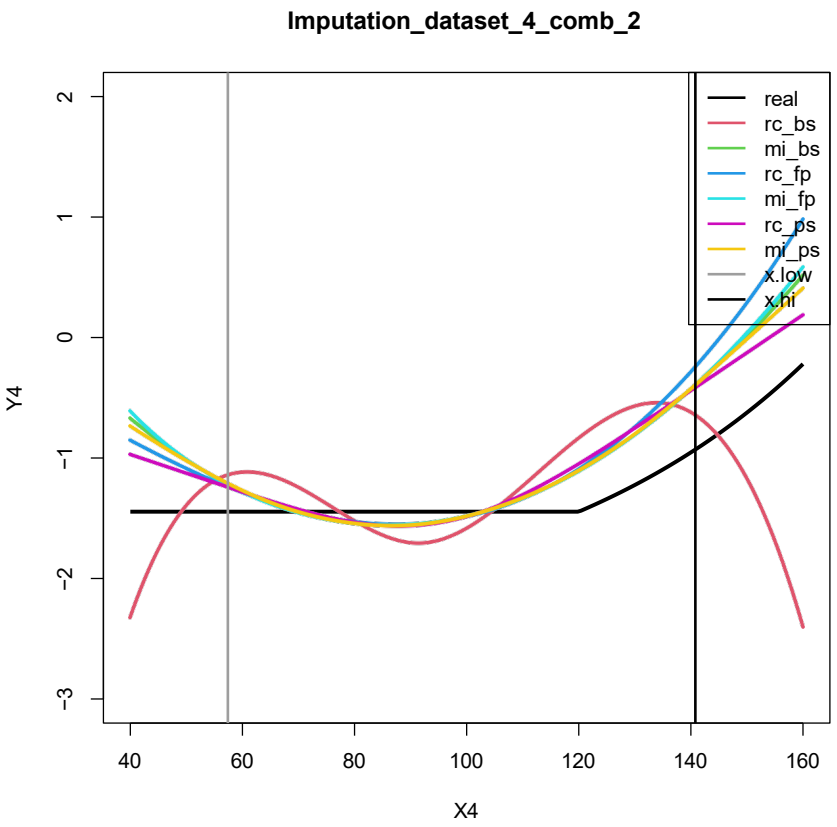



**SIMEX (pointwise)**

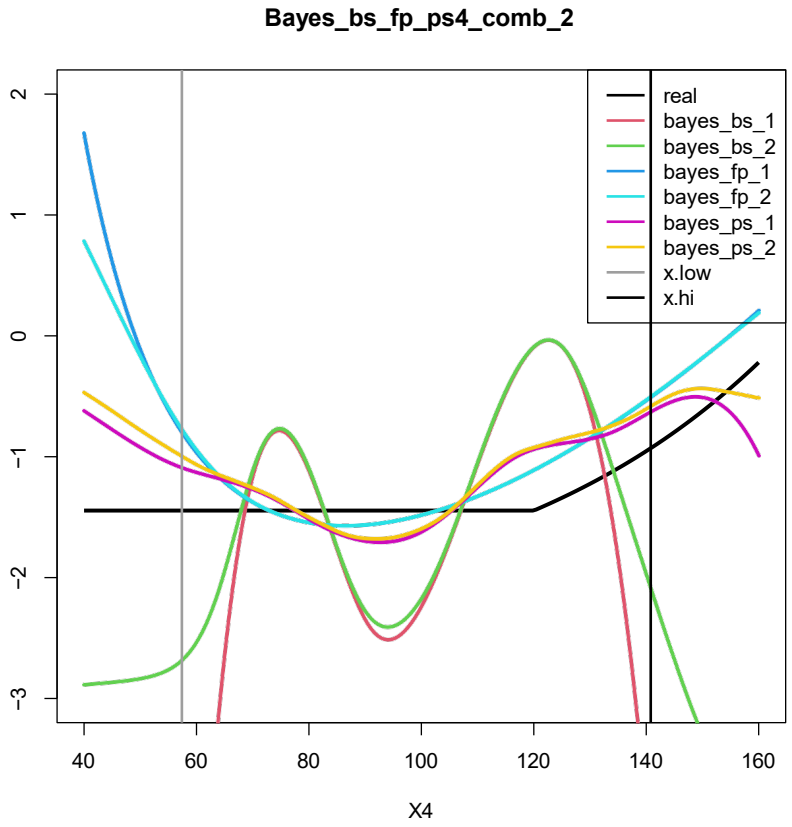


# Stage 2 Selected results: Threshold above median

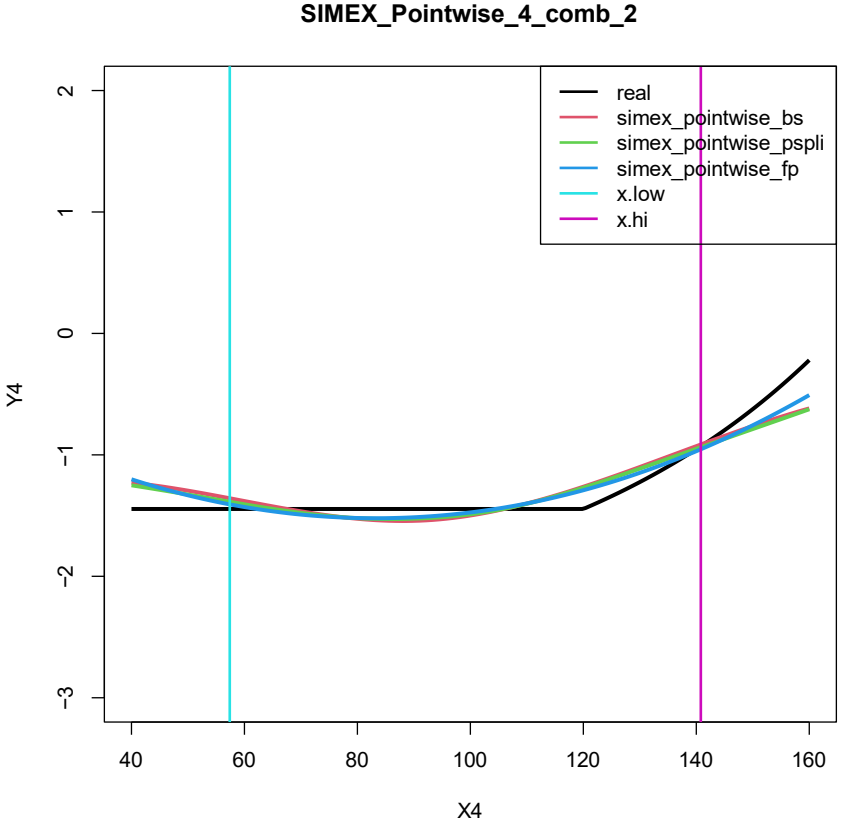
**Multiple Imputation and Regression Calibration**




**Bayes**




**SIMEX (pointwise)**



# Main Summary

- SIMEX performed best
- Bayes with fractional polynomials or P-splines was next best
- Multiple imputation and regression calibration performed similarly and were in third place
- Bayes with B-splines performed poorly

The results were surprising

- Most of us expected: Bayes + MI > RC > SIMEX
- We do not yet have a clear explanation of why it happened
- We suspect that SIMEX might be more robust in complex models

# Other matters

We have performed more in-depth analysis of the results relating estimation accuracy to dataset characteristics, and looking for interactions with estimation methods. Results available on request.

Next steps:

(a) Further investigation of reasons for the results

(b) Expand Stage 2 to smaller sample sizes (going down from 15000 to 2000)

(c) Stage 3: Perform replication to understand variances of the estimates

(d) Stage 4: Broaden to more realistic scenarios, e.g. where there are several other covariates

# References to Neutral Comparison Studies


1. Towards neutral comparison studies in methodological research.

   Boulesteix AL et al (eds). Biometrical Journal 2024; Vol.66 (Issue 2)

2. STRATOS: Neutral comparison studies as the cornerstone to compare statistical methods.
   Boulesteix AL et al. Biometric Bulletin 2024. Issue 2.