Evaluating the impact of covariate measurement error on functional form estimation in regression modelling

Aris Perperoglou, **Michal Abrahamowicz**, Paul Gustafson, Victor Kipnis, Anne Thiebaut, Raymond Carrol, Kevin Dodd, Steve Ferreira Guerra, Frank Harrell, Nadja Klein, Douglas Midthune, Willi Sauerbrei and Laurence Friedman.

A joint project of Topic Groups 2 (Selection of Variables and Functional Forms) and TG4 (Measurement Error and Misclassification) of the STRATOS Initiative

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: Only a minority of published papers present estimates that are adjusted for measurement error.

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.

TG2: 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

Perperoglou *et al. BMC Medical Research Methodology* (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

BMC Medical Research Methodology

REVIEW Open Access

A review of spline function procedures in R

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of 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?

TG4: Key publications

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

SIGNIFICANCE ROYAL STATISTICAL SOCIETY ASA AMERICAN STATISTICAL ASSOCIATION Statistical Society of Australia

Analysis in an imperfect world

When we observe the world, we sometimes make mistakes. **Michael Wallace**, on behalf of the measurement error topic group of the STRATOS Initiative, explains the potentially severe consequences of this often overlooked issue, and how statistics can help bring us back – or at least a little closer – to the truth

Michael Wallace is an

First published: 29 January 2020 | https://doi.org/10.1111/j.1740-9713.2020.01353.x | Citations: 1

Measurement error in regression modelling

We are interested in learning the regression relationship between an outcome variable $\boldsymbol{Y}$ and a covariate(s) $\boldsymbol{X}$.

$$E(Y|X) = \beta_0 + \beta_X X$$

Measurement error can be seen in continuous covariates, categorical covariates (misclassification) or the outcome variable $\boldsymbol{Y}$.

Focus here on the first case, of a continuous covariate, for which the true value of $\boldsymbol{X}$ may be unobserved. Denote $\boldsymbol{X}^*$ the error-prone observed variable.

- **Classical Measurement Error Model (CME)**

  $\boldsymbol{X}^* = \boldsymbol{X} + \boldsymbol{U}$, where $\boldsymbol{U}$ is random variable with mean 0, independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

- **Non-differential error** stipulates that error distribution remain consistent across different levels of the outcome variable.

Effects of measurement error in studies

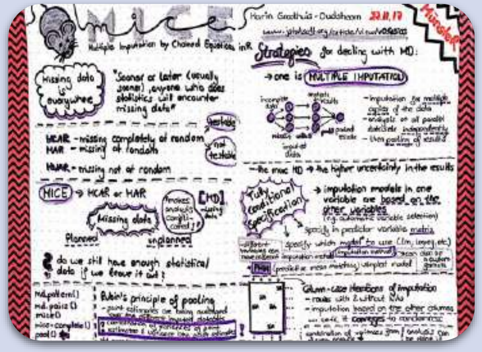
- Assume a simple linear regression, given as: $E(Y|X) = \beta_0 + \beta_X X$
- Because of measurement error we explore: $E(Y|X^*) = \beta_0^* + \beta_X^* X^*$
- Under non-differential and CME ($\boldsymbol{X}^* = \boldsymbol{X} + \boldsymbol{U}$) then $|\beta_X^*| \leq |\beta_X|$ with equality only when β_X=0
- The measurement error attenuates the estimated coefficient $\beta_X^* = \lambda\beta_X$, where $\lambda = \frac{var(X)}{var(X)+var(U)}$, the attenuation factor $[\, 0 < \lambda \leq 1 \,]$
- Larger var(U) → smaller λ → greater attenuation
- Measurement error also makes the estimate less precise relative to its expected value $\frac{E(\widehat{\beta_X^*})}{se(\widehat{\beta_X^*})} < \frac{E(\widehat{\beta_X})}{se(\widehat{X})}$
- Effective sample size is reduced by $\rho^2_{XX^*}$, the squared correlation coefficient between $\boldsymbol{X}$ and $\boldsymbol{X}^*$
- While measurement error in this setting results in bias and loss of power, null hypothesis $\beta_X^* = 0$, is still a valid test for β_X

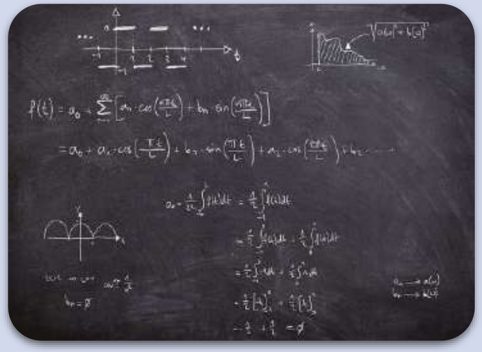

```
# Simulate data
n <- 500
X_true <- rnorm(n)
# Non-linear relationship for Y without measurement error
Y <- 0 + 1*X_true+ rnorm(n)
# Introduce measurement error to X
X_observed <- X_true + rnorm(n, sd=0.5)
# Simple linear regression using X with measurement error
lin_reg <- lm(Y ~ X_observed)
lin_reg_true <- lm(Y~X_true)
# Plot Y against X_observed
plot(Y ~ X_observed, main = "Scatterplot with Linear Fit ",
     xlab = "X_observed", ylab = "Y", col = "blue")
legend("bottomright", legend = c("Data", "On X", "On observed X*"),
       col = c("blue", 1, 2), lty = c(NA, 1, 1), pch = c(1, NA, NA))
# Add the regression line to the plot
abline(lin_reg, col = "red")
abline(lin_reg_true)
```

The impact of measurement error on functional form estimation

- Often we encounter cases where X is not linearly related with Y: $\mathrm{E}(Y|X) = f(X)$
  - Examples in dose-response studies, environmental exposures…
- Challenges
  - Function $f()$ is unknown, requiring flexible estimation methods
  - We observe X^*which is measured with error
- Consequences not fully understood
  - The observed X^* can introduce bias in the estimated $f()$ and mislead inference
- Objectives
  - Evaluate the impact of **measurement error** in a **continuous predictor** X, on its estimated, potential **non-linear** dose response relationship $f(X)$ with an outcome Y**.**
  - Compare different methods of estimating the true relationship between the outcome variable Y and the covariate X.
  - Validate "correction" strategies to reduce the impact of measurement error.

Framework of investigation





Data Generation & Evaluation

- Anne Thiebaut (lead)
- Laurence Freedman
- Aris Perperoglou

Bayesian Methods

- Paul Gustafson (lead)
- Raymond Carroll
- Frank Harrell
- Nadja Klein

Imputation

- Victor Kipnis
- Douglas Midthune

SIMEX

- Michal Abrahamowicz (lead)
- Steve Ferreira

The simulation

- Data generated from $\mathrm{logit}(\mathrm{P}(\mathrm{Y} = 1|X)) = f(X)$ where X **distribution** of X and $f(X)$ is undisclosed
- K datasets, each:
  - **Main study:** N={15000, 5000) independent realizations of a Y **binary outcome** and a **continuous covariate measured with error** X^*
  - **Validation sub-study:** n pairs of repeat observations of X^*.
- A classical measurement error model linking error-prone X^* to X, with **error term** having undisclosed variance and distributional form.
- Each dataset will be checked for outliers, and the outliers removed. The undisclosed aspects of these datasets will be varied across the K datasets.

Workflow

Data generation team → Methods Teams → Data generation team

Simulates and distributes data

- Defines flexible functions to be investigated:
  - Cubic b-splines with 1 interior knot at median of observed X*.
  - P-splines with 10 interior knots. Penalty optimised within groups.
  - Fractional Polynomials of second degree. Powers selected within groups.

Applies methods

- SIMEX
- Imputation
- Bayesian

Evaluates methods

- Mean squared error
- Weighted mean square error

Simulation-Extrapolation (SIMEX)

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 to the case of no error.

For this project, we propose two alternative SIMEX approaches:

1) Apply SIMEX directly on estimated curves
   → Let $\hat{f}(x_0)$ be an estimate of the NL relationship for selected X = x_0. $\hat{f}(x_0)$ will be estimated for increasing amounts of measurement error and then extrapolated to the case of no error, yielding the SIMEX corrected estimate of $\hat{f}(x_0)$.

2) Apply SIMEX on the spline or FP coefficients
   → For increasing amounts of measurement error, estimate the spline of FP coefficients and extrapolate each coefficient to the case of no error.

   → The SIMEX-corrected estimate of $\hat{f}(x_0)$ will then be obtained using the extrapolated coefficients.

Imputation methods

- **Regression calibration** estimates the conditional expectation of the function $f(X)$ given the error prone covariate X* and substitutes it for the true covariate in the logistic regression.
    - Assuming that there is a Box-Cox transformation g so that the model for g(X*) on the transformed scale in the calibration substudy is specified as a linear mixed model with random intercept, the conditional expectation of $f(X)$ on the original scale can be estimated by using the NCI method (see NCI BRG website)
- **Multiple imputation:** The imputed $f(X)$ 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
    - The method defers since the model for g(X*) in the calibration substudy should include a covariate being the output dichotomous variable Y in that substudy.

Bayesian Methods

We specify:

- an outcome model (for Y given X)
- an exposure model for X
- a measurement model for X* given X
- prior distributions for parameters in each of the three sub-models

* This defines 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.

Methods of evaluation

Let $f(x_1), f(x_2), \dots, f(x_m)$ the true values of the function and $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_m)$ the estimated values

For each dataset choose **undisclosed** evaluation “limit points” x_{low} and x_{high} that define range over which evaluation will be conducted compute:

- **Unweighted mean squared error on the log odds scale**: $\sum_{i=low}^{i=high}\{f(x_i) - \hat{f}(x_i)\}^2$/(high – low +1)
- **Weighted mean squared error on the log odds scale**: $\sum_{i=low}^{i=high} w_i\{f(x_i) - \hat{f}(x_i)\}^2$/(high – low +1), where the weight w_i is the density of x_i in the distribution of X.

Other evaluation functions that may be considered are:

Absolute error on the log odds scale, mean squared error on the risk scale and absolute error on the risk scale

The process

Stage 0: Preparation

- Data team **(DT)** defines simulation method, variations and combinations of parameters.
- Methods teams **(MT)** define their method of analysis.

Stage 1: First few datasets

- **Five** example datasets will be circulated to MTs, each with a different f(x).
- Teams pilot their methods and create scripts (R and/or SAS).
- Results back to data team along with first scripts for running code.

Stage 2: More data

- DT produce combinations on:
  - *main sample size*
  - *replication sub-study sample size*
  - *measurement error variance*
  - *distributional form of X*
- Datasets fed back to MTs.
- MTs fine-tune their code and ensure it can be applied by the data team.

Stage 3: Methods applied by data team

- DT produces further replicates of each scenario variant (approx 20 – leading to 800 to 1600 datasets).
- **DT uses computer code provided by the three teams to run methods and evaluate results.**
- At each stage the results will be compared.
- Methods that appear clearly inferior, may be dropped from further simulations.

Stage 4: Optional

- If there are competing methods that appear optimal but indistinguishable for a scenario variant, then further simulations up to 500 will be run on them, to discover whether one is optimal.

today

Assessing the Impact of Measurement Error on B-spline & FP2 Estimates of Non-Linear Functions: *preliminary Simulation results*

Michal Abrahamowicz * & Steve Ferreira Guerra
McGill University
Montreal, CANADA

(on behalf of All Participants of the joint TG2.&.TG4 STRATOS project)

44TH ISCB cONFERENCE

Milan, August 31, 2023

Overall Aim

This project is the 1^{st} step in the TG2-TG4 collaboration.

Specific goal :

To use simulations to assess the impact of measurement error (ME) in a continuous 'covariate' *X* on B-spline and fractional polynomial (FP) estimates of its possibly non-linear (NL), relationship *f(X)* with the outcome in univariate logistic regression.

Data generation

- Classical 'random' ME model:
  - Observed = Truth + error ($X_i^* = X_i + e_i$)

2 distributions of X:
  - X ~ Unif(80,150)
  - X resampled from real-world values of SBP at baseline from the Framingham Heart Study

4 strengths of ME:
  - $e_i \sim N(0,\sigma_e)$
    - σ_e / σ_X = ¼, ½, ¾, 1
- Logistic regression with a Binary outcome : logit[P(Y=1|X)] = *f(X)*
  - Different shapes of true f(X)
    - Created using complex functions of X (e.g., asymmetrical sigmoidal or 5-degree polynomials)
- Sample sizes of N = 250, 500, 1000, 2000 (with ~30% cases: Y=1)

Analysis methods

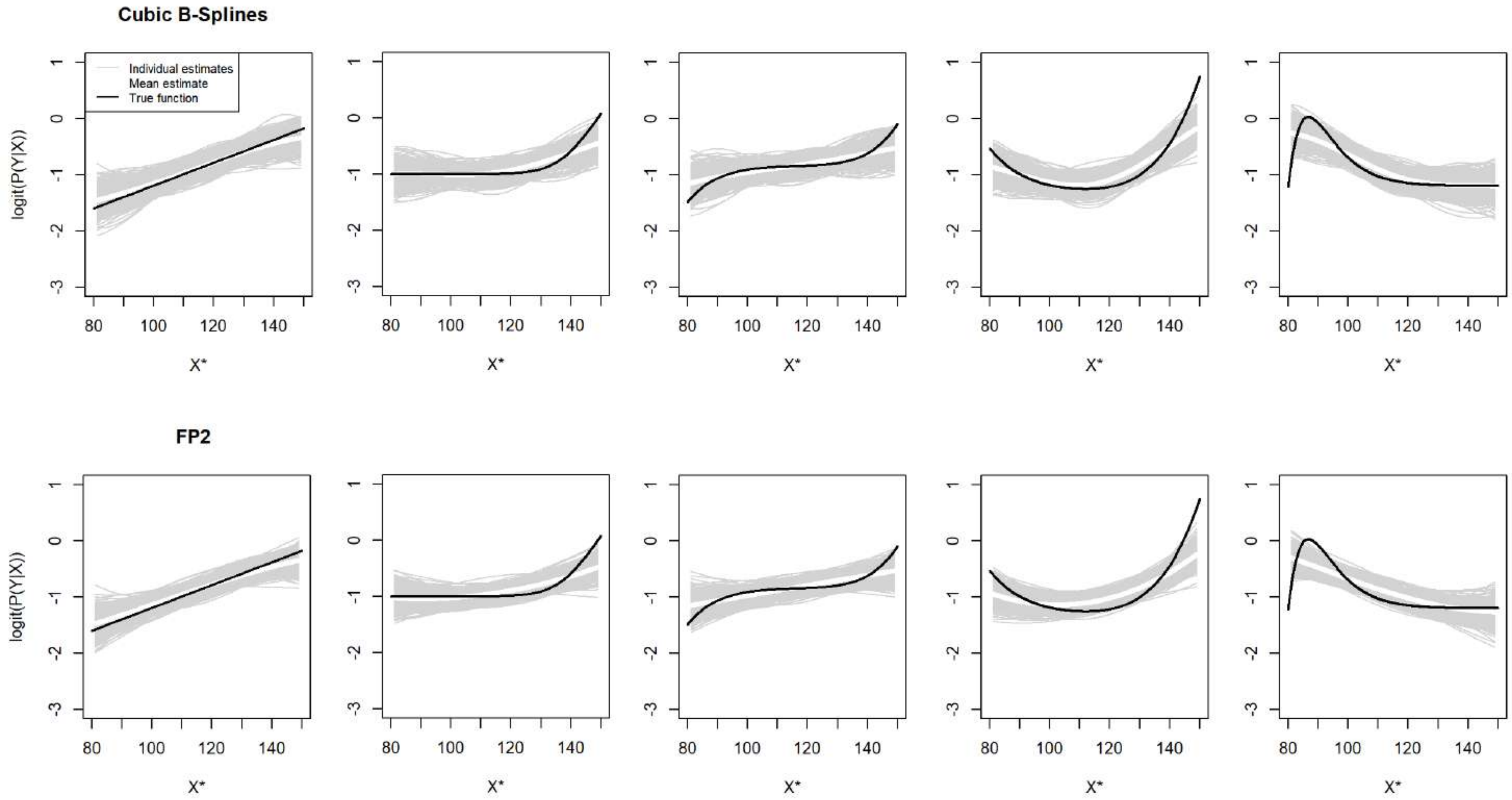
We compared estimated NL curves for (i) f(X) for true X, vs (ii) f(X*) for error-prone X* using **2 flexible estimation methods**:

- **Unpenalized cubic regression B-splines** with 1 interior knot (4 df) placed at the median of X*
- **Fractional polynomials of degree 2 (FP2**, 4 df), using the MFP algorithm to select the two powers *(FP2 was "forced" regardless of test results)*

SELECTED RESULTS

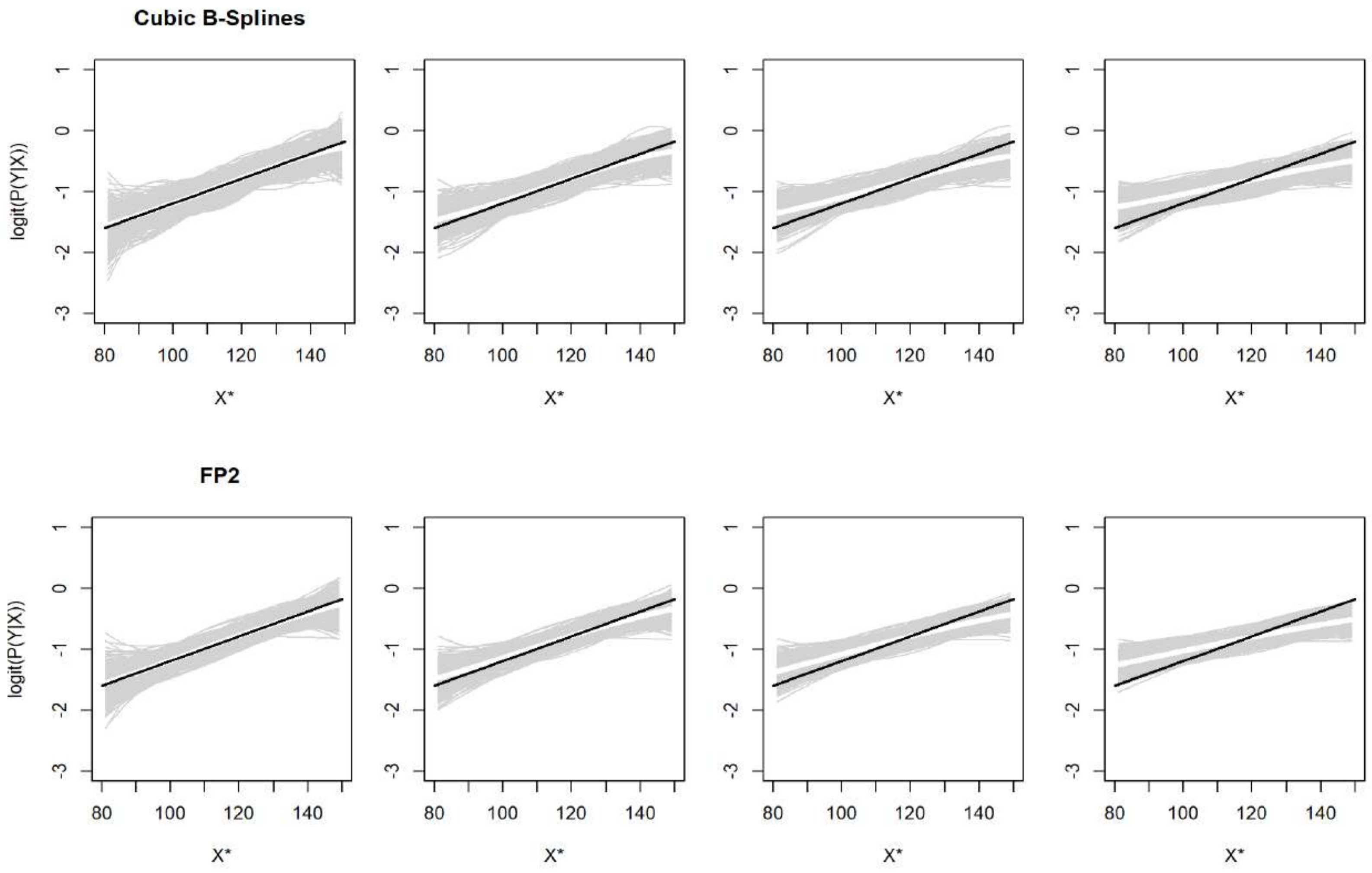
Uniform Distribution of x

5 scenarios for true f(X): small local biases



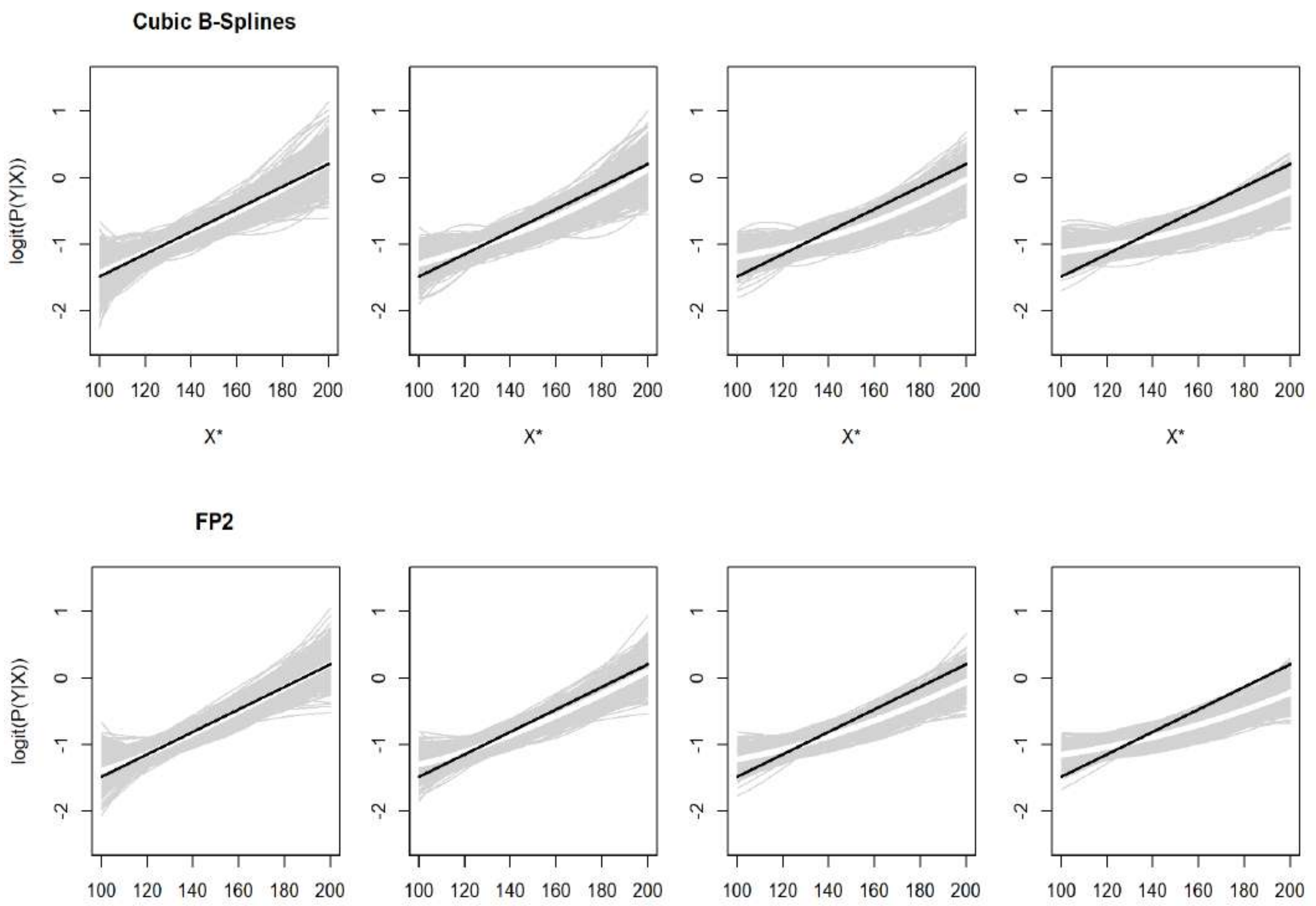
(**Black** = True f(X), **Grey** = Individual estimates, **White** = Mean estimate)
*N = 1000, X~Unif(80,150), $\sigma_e / \sigma_X = ½$

Attenuation increases with increasing ME in X* (L -> R) (Uniform X, N= 1,000)



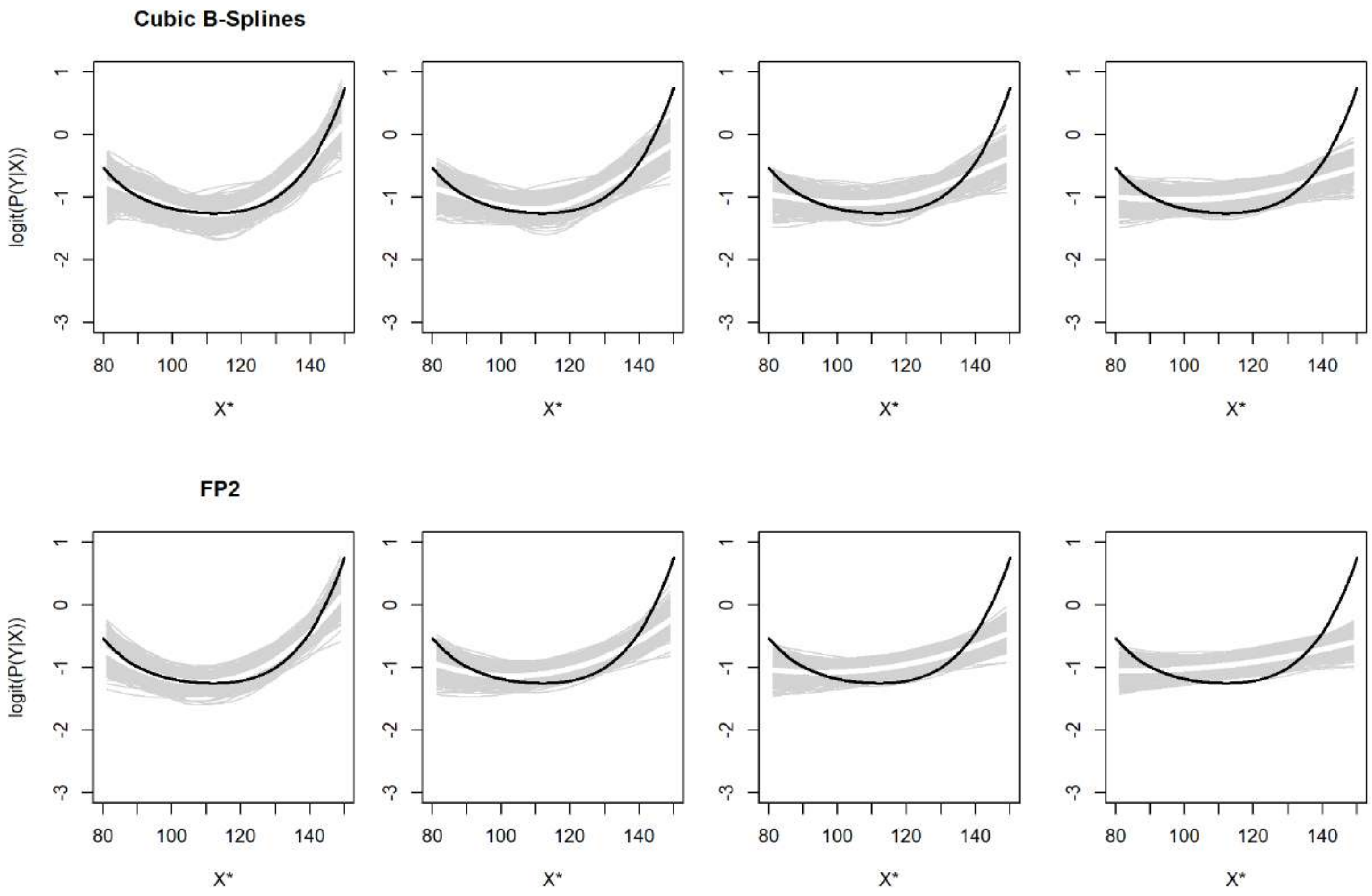
σ_e / σ_X = ¼, ½, ¾, 1, respectively
*N = 1000, X~Unif(80,150)

Non-uniform (+ skewed) X, based on SBP data: Spurious Non-linearity in upper half of f(X*)



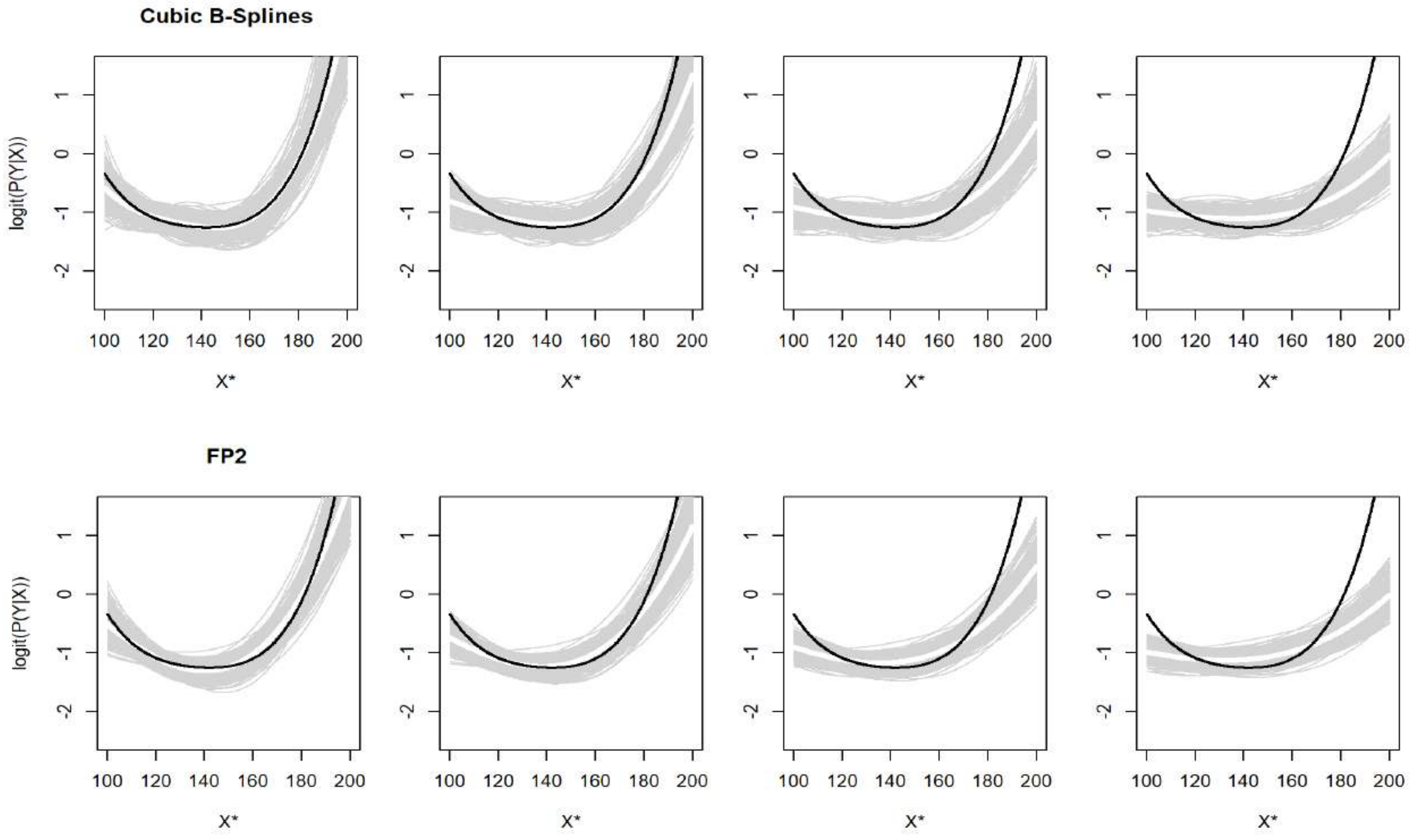
σ_e / σ_X = ¼, ½, ¾, 1, respectively
*N = 1000, X = SBP

Flattening increases with increasing ME in X* (L -> R)(Uniform X, N= 1,000)



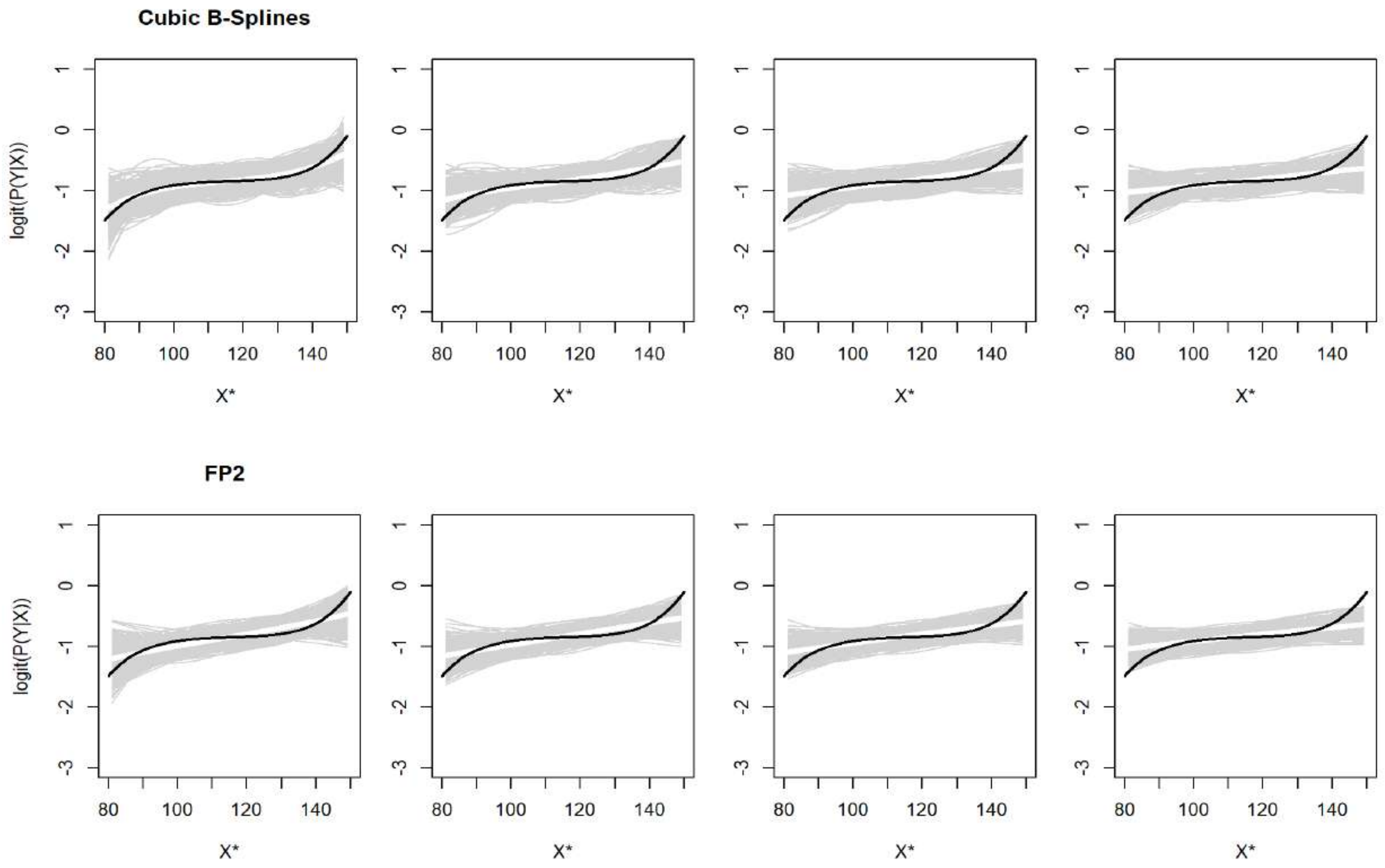
σ_e / σ_X = ¼, ½, ¾, 1, respectively
*N = 1000, X~Unif(80,150)

Flattening increases with increasing ME in X* (L -> R) (+ skewed X distrib., based on SBP data, N= 1,000)



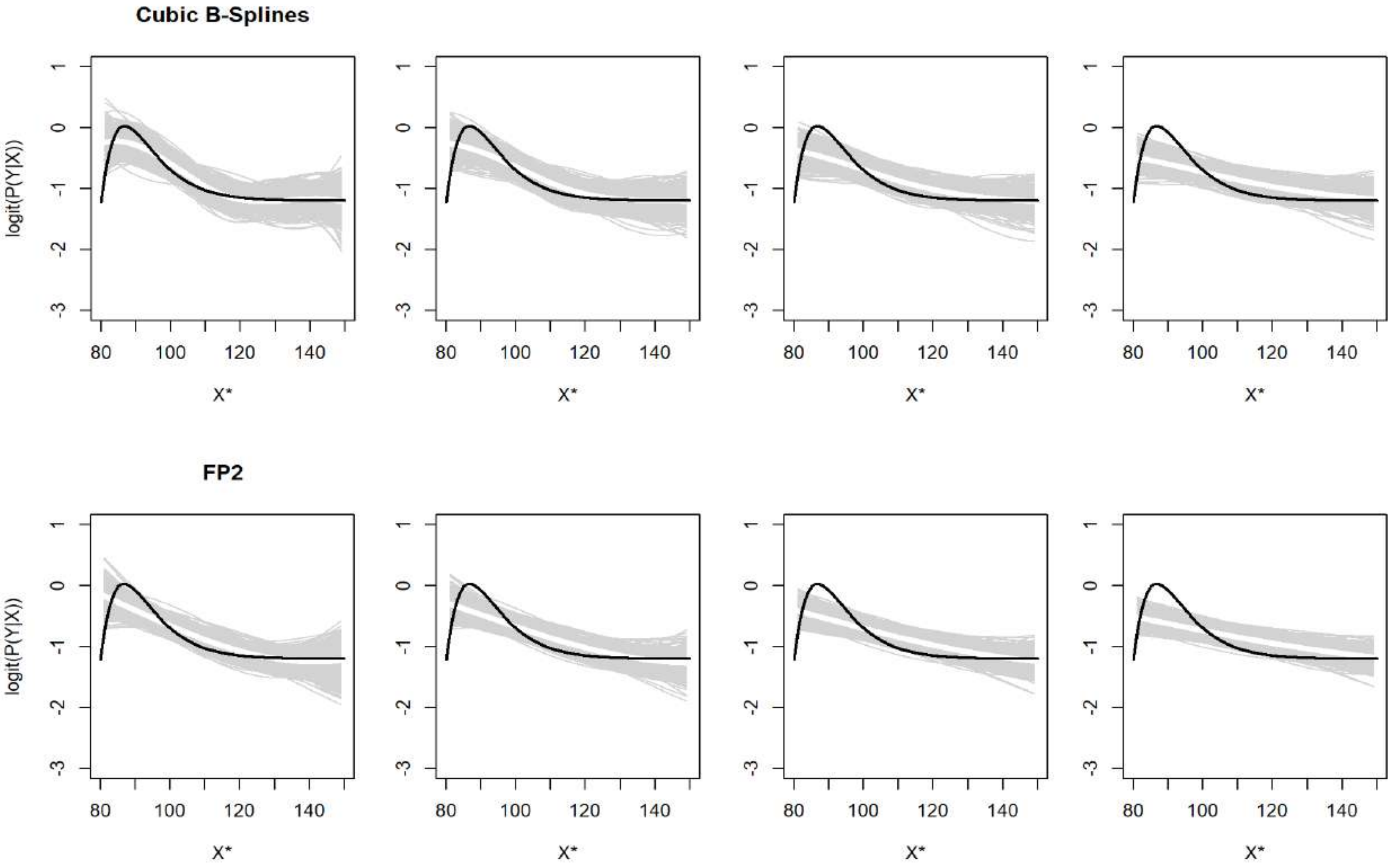
σ_e / σ_X = ¼, ½, ¾, 1, respectively
*N = 1000, X~SBP

“Linearization” increases with increasing ME in X* (L -> R) Uniform X, N= 1,000



σ_e / σ_X = ¼, ½, ¾, 1, respectively
*N = 1000, X~Unif(80,150)

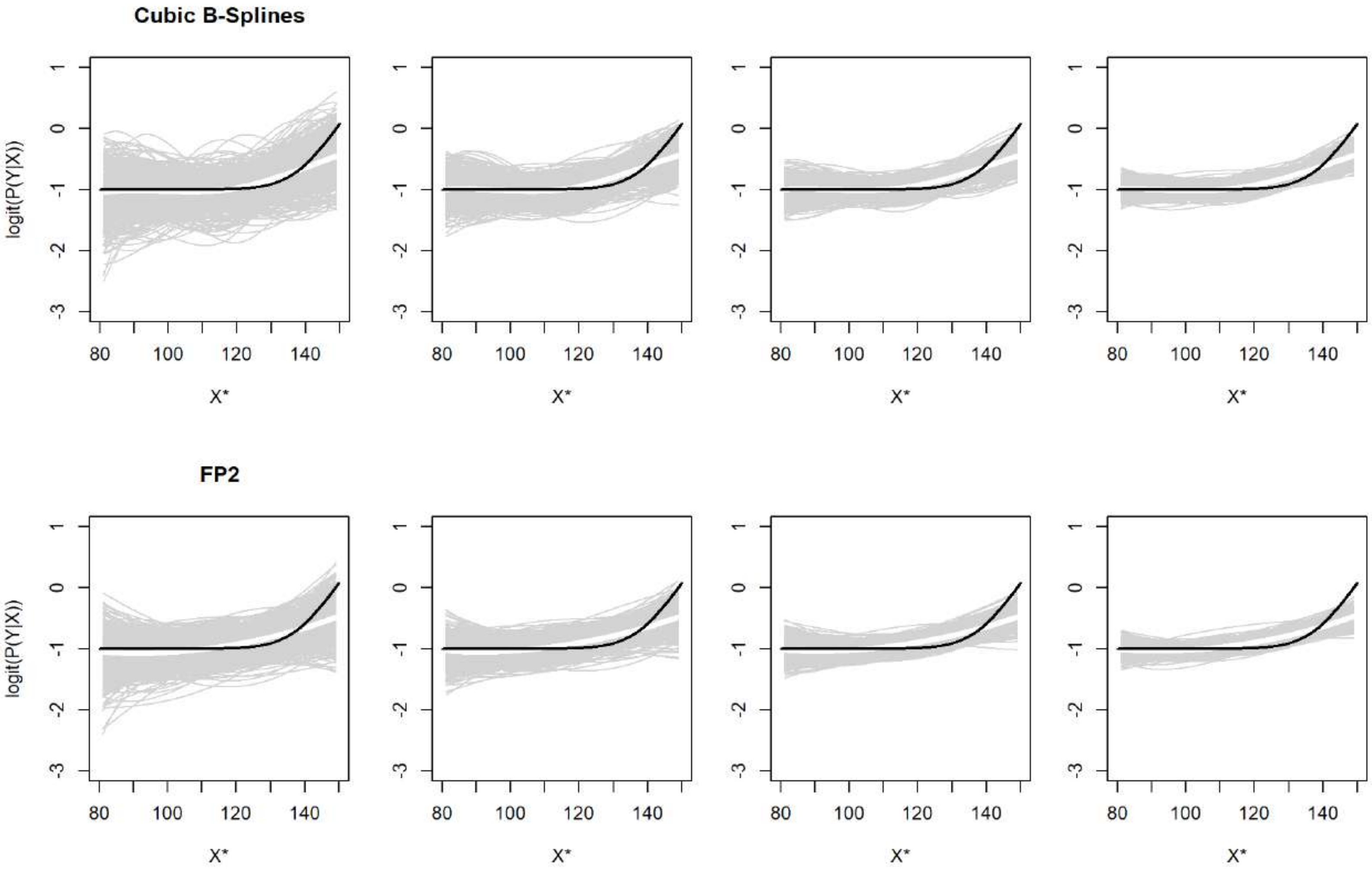
“Linearization” increases with increasing ME in X* (L -> R) Uniform X, N= 1,000



σ_e / σ_X = ¼, ½, ¾, 1, respectively
*N = 1000, X~Unif(80,150)

Increasing N (L->R) affects only variance but Bias in f(X*) remains

STRATOS INITIATIVE



N = 250, 500, 1000, 2000, respectively
*X~Unif(80,150), $\sigma_e / \sigma_X = ½$

Preliminary Conclusions

- **Random Measurement Errors (ME) in X may affect flexible estimates of Non-Linear (NL) associations in a complex way**
- **Generally,** ME induces **both «Linearization» & «Flattening (attenuating)»** of the NL estimates
- however, **Locally,** the ME-prône estimate **may over-estimate the local slope and induce spurious NL**
- Results **may dépend on the True Distribution of X**
- **Cubic B-splines vs Fractional Polynomials (FP2) produce very similar estimates**
- **Splines & FP2 (both with 4 df) somewhat biased even for True X if f(X) complex …**
- With low N (250, about 80 events/cases) FP2 estimates more stable than splines

STRATOS INITIATIVE

GRAZIE THANK YOU