Designing Simulation studies for accurate, generalizable conclusions: recent developments

Michal Abrahamowicz * [1]
& Anne-Laure Boulesteix [2]

for the STRATOS Simulation Panel

[1] McGill University, Montreal, Canada

[2] University of Munich, Germany

Support:
Natural Sciences & Engineering Research Council of Canada (NSERC)

McGill

STRATOS Simulation Panel

Co-Chairs:

Michal Abrahamowicz, McGill, Montreal, Canada &

Anne-Laure Boulesteix, Munich, Germany

Members:

Harald Binder, Freiburg, Germany

Jessica Franklin, Harvard, Boston, USA

Rolf Groenwold, Leiden, the Netherlands

Victor Kipnis, NIH, Bethesda, USA

Tim Morris, UCL, London, UK

Willi Sauerbrei, Freiburg, Germany

Pamela Shaw, U Penn, Philadelphia, USA

Ewout Steyerberg, Leiden, the Netherlands

Ingeborg Waernbaum, Uppsala, Sweden

Outline

- **WHY we Need More (and Better) Simulations?**
- **General Recommendations** from STRATOS Simulation Panel
- **Reproducilibility** of Simulations

- Some **"Specialized" issues:**
  - Making Design & Assumptions *Clinically Relevant/Plausible*
  - Assessing *Bias/Variance Trade-off*
  - Digging deeper into results (*Beyond "Average"* performance)
  - Considering *Unfavorable Assumptions*
  - Watching for a *Tip of the Iceberg*

- **Practical Conclusions**

Why we need *more (and Better)* Simulations?

- STRATOS' overarching goal:
  provide Evidence-based Guidance re: choice of method(s) to deal with specific analytical challenges of observational studies

- **Yet, solid Evidence is often lacking, because:**
  - *Proliferation* of new, ever *more sophisticated/complex methods*
  - Alternative methods address *similar issues with different techniques*
  - *Difficult to analytically prove properties* of complex methods
  - *Relative Performance* of alternative methods – *and their Validity* – may *depend on the data structure/quantity/quality*
- *Thus, End-Users are often at loss re:*
  - *Will this method work for my data?*
  - *Which of the alternative methods I should use? etc.*

- **These challenges can be addressed by well designed, executed and interpreted Simulation studies**

Limitations of most published Simulation studies

- **Design and scope of simulations** reported in recent statistical papers **often raise concerns** (arbitrary assumptions, univariate…)
- E.g. *developers of the new methods often attempt to demonstrate its 'superiority'* over the existing alternatives, but consider *only a limited range of 'favorable'* scenarios,

raising concerns about specific

"Publication Bias"

Recent Progress:
ADEMP framework

- Several **'generic' issues related to design, conduct and reporting of simulations comparing alternative statistical methods** are addressed by **Morris, White & Crowther** [*Statistics in Medicine* 2019. 38: 2074-2102], who propose very useful:

  ADEMP reporting framework (**Aims; Data-generation; Estimands; Methods & Performance measures**).

➢ **Adherence to ADEMP rules will enhance the Validity & Transparency of statistical simulations**

General Recommendations from the STRATOS Simulation Panel

- (Based on the recent *Letter*:

  Boulesteix, Binder, Abrahamowicz & Sauerbrei [*Biometrical J* 2018]):

  The overarching Goal of the Simulation Panel is to advocate more wide-spread use of ***neutral* (unbiased) and '*realistic*' comparison studies** evaluating the performance of existing and new statistical methods using (mostly) Simulated or Real-life data

- Similar to the guidelines for RCT's [e.g. CONSORT] evaluating new treatments, **statistical simulations should meet several criteria, and address important design/reporting issues:**
  - (i) How to **simulate data in a Realistic way**, inspired from relevant real datasets?
  - (ii) How to ensure the **Reproducibility and Transparency of the methods used for Data Generation and Analyses?**
  - (iii) What are typical **sources of potential Biases** and how can they be avoided?
  - (iv) How can the **results be Interpreted**, without **the risk of over-interpretation**?
  - **(v) What Parameters & Assumptions should be Varied** across simulated scenarios?
  - (vi) What **range of Sample Sizes** should be considered?
  - **(vii) Which "Competing Methods"** should be considered?
  - **(viii) Etc. ...**

Essential criteria for "Informative Simulations":

Reproducibility / Availability of the Software

- **Simulation studies should be fully reproducible, i.e. all necessary information should be provided**

  (assumptions, relevant parameters, data generation algorithms, analysis methods, criteria to assess/summarize the results)

- **Space restrictions no longer an excuse**, as this information can go to the **Supplementary Materials**

- **Providing scripts** (with clear instructions for the users) e.g. on *github* (or even R packages) that allow to re-do simulations **ensures full Transparency and Reproducibility**.

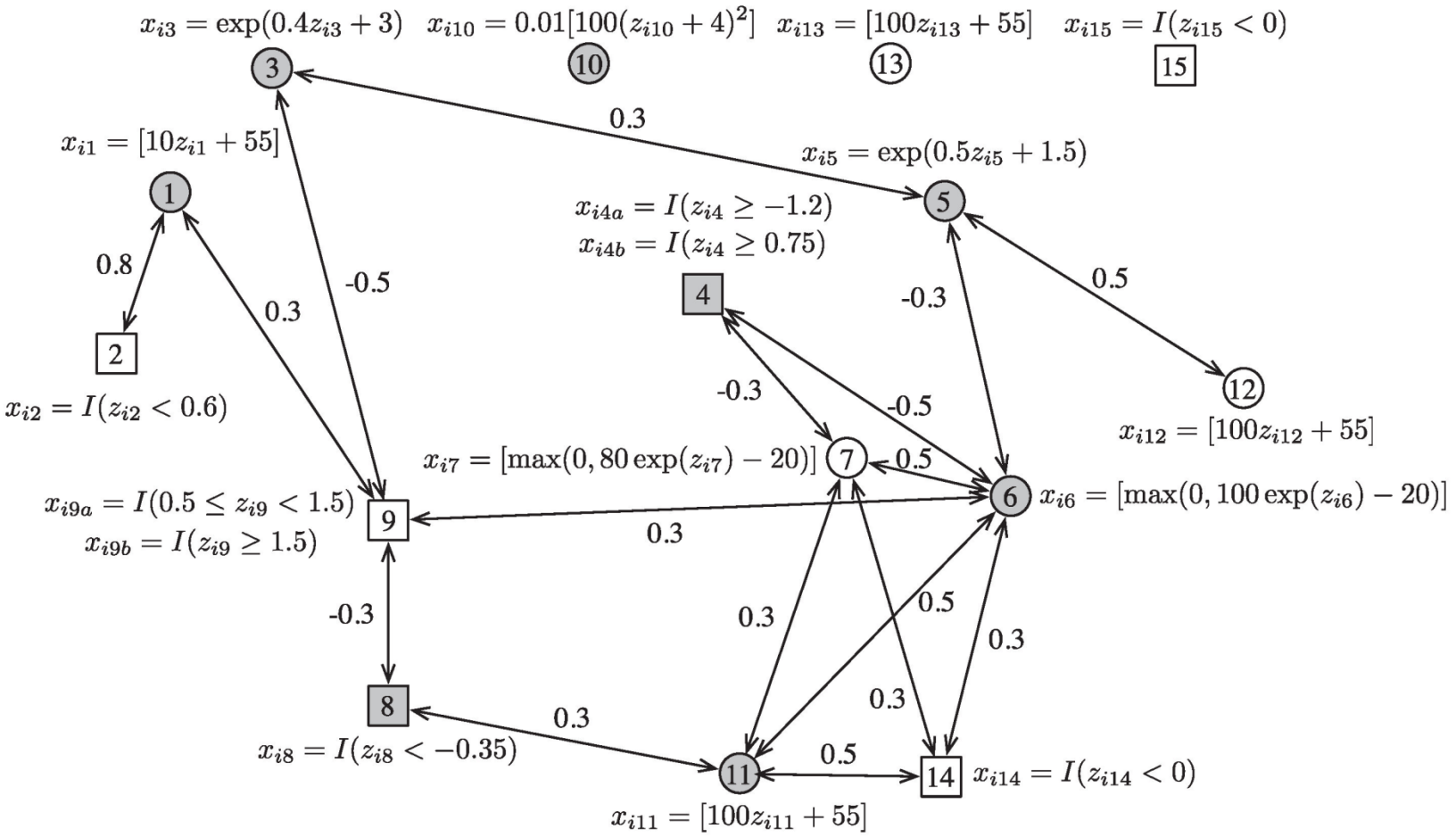
(rare) **Example of Realistic, Complex Simulations**
[**H. Binder, W. Sauerbrei, P. Royston. *Stat Med* 2013**]

- **3 "Competing** Goal:
  Compare flexible model building approaches for selection of (i) important variables and (ii) functional forms (for continuous variables)
  (in exploratory multivariable linear regression analyses)

- **Methods" compared:**
  - Multivariable fractional polynomial (MFP)
  - Restricted cubic splines (RCS)
  - Penalized splines (PS)

- Design of simulation study:
  - Based on **Realistic biomedical data**
    (**Distribution of Variables & their Correlations based on the Rotterdam Breast Cancer study**)
    15 variables (mix continuous/binary), **with complex correlation structure** (Figure 1 on the next slide)
  - **Alternative Sample Sizes** (N= 200, 500, or 1,000)

Figure 1. Simulation design: correlation structure of the underlying 15 variables (circles/squares) is indicated by arrows, where the numbers indicate the correlation coefficients. The formulae for obtaining the covariates from the underlying variables are adjacent to the circles/squares. [] indicates that the non-integer part of the argument is removed, and *I*() is the indicator function, taking the value 1 if its argument is true and 0 otherwise. **Continuous constructed covariates are indicated by circles**, and **categorical covariates by rectangles**. If a covariate has an effect on the response, the circle or rectangle is shaded grey. Note that some of the underlying variables correspond to several covariates or model components, for example, variable 4 corresponds to x_{4a} and x_{4b}, but only x_{4a} has a non-zero effect.



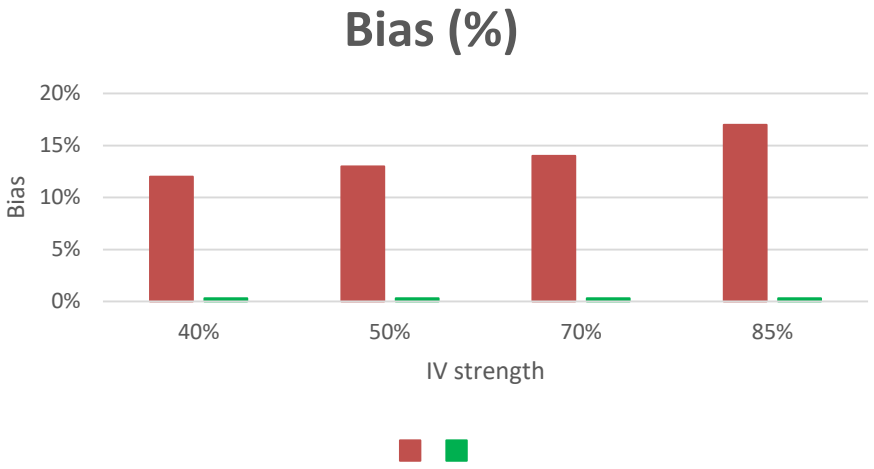
Importance of comparing the Bias/Variance trade-off (*via* RMSE)

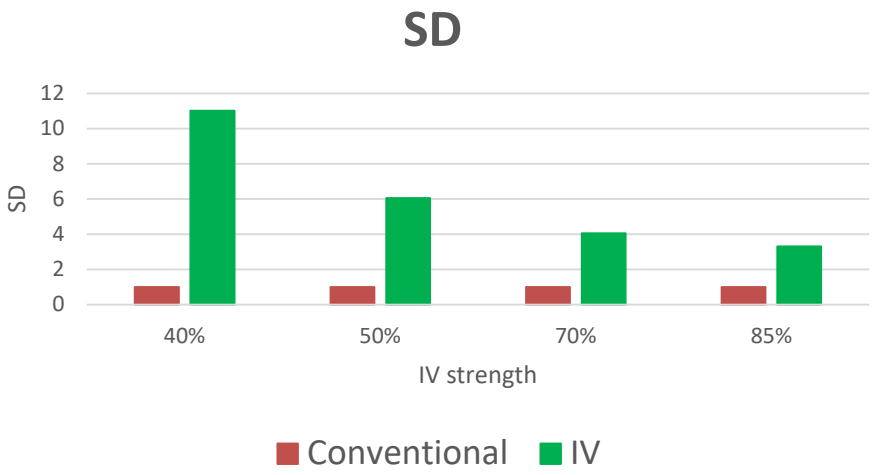
- **Many recent, complex methods aim at reducing different Biases**
- These methods typically use additional information and/or often involve **estimating additional 'auxiliary parameters or meta-parameters** (often using a 2-step estimation) (Examples: *Instrumental Variables (IV), MSM's with IPT weights, Fractional Polynomials, Missing Cause, SIMEX, Net Survival*...)
- **This tends to Inflate the Variance** of the estimates
- Thus, it is **Essential to report both the SD's** and the RMSE of the estimates (Lower RMSE = better Bias/Variance trade-off)**
- **Yet, relative RMSE's of alternative methods may depend strongly on some design parameters (see Next Slide,** which shows how **Instrumental Variables (IV) ability to reduce Unmeasured Confounding Bias depends on the Instrument's Strength)**

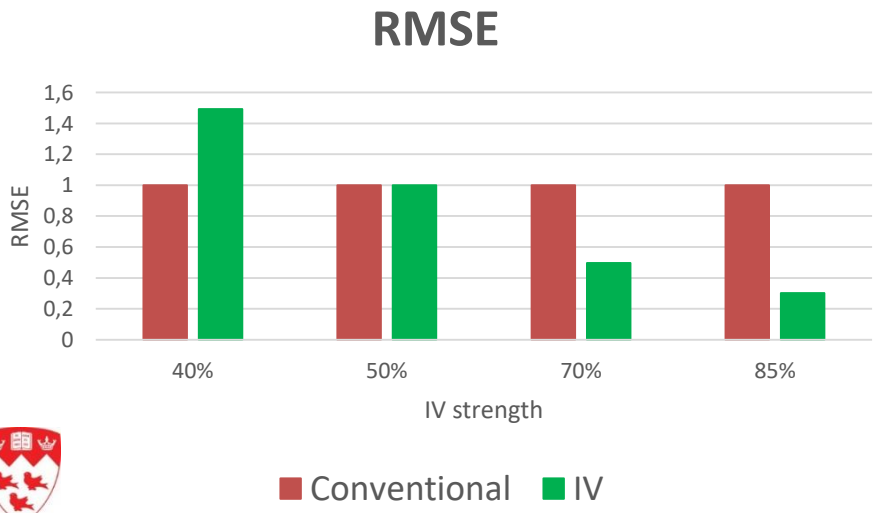
  [Ionescu-Ittu et al, *Pharmacoepi & Drug Safety (PDS)* 2009]

** Also: Hessian-based "analytical" Variance estimates are NOT accurate for 2-stage estimators and simulations can help assess the resulting under-estimation of the variance

Bias, SD and RMSE of IV-corrected vs Conventional estimates in presence of Unmeasured Confounding

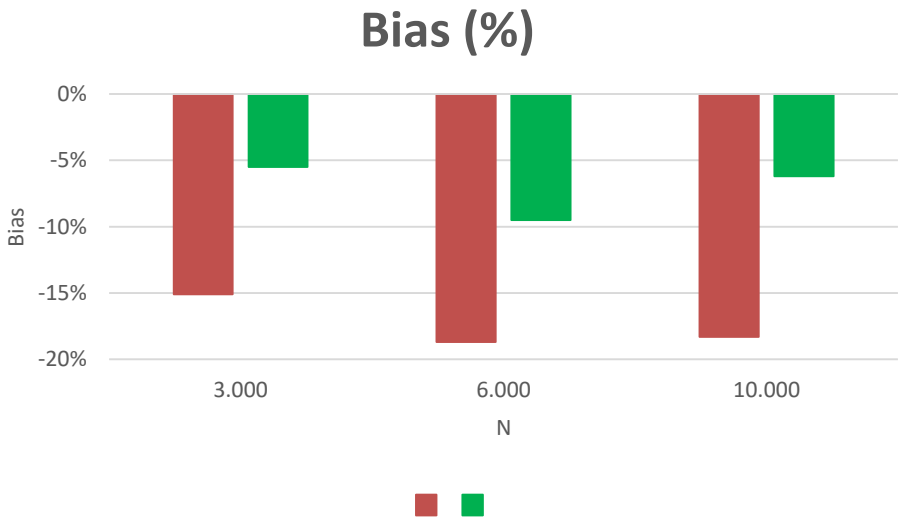


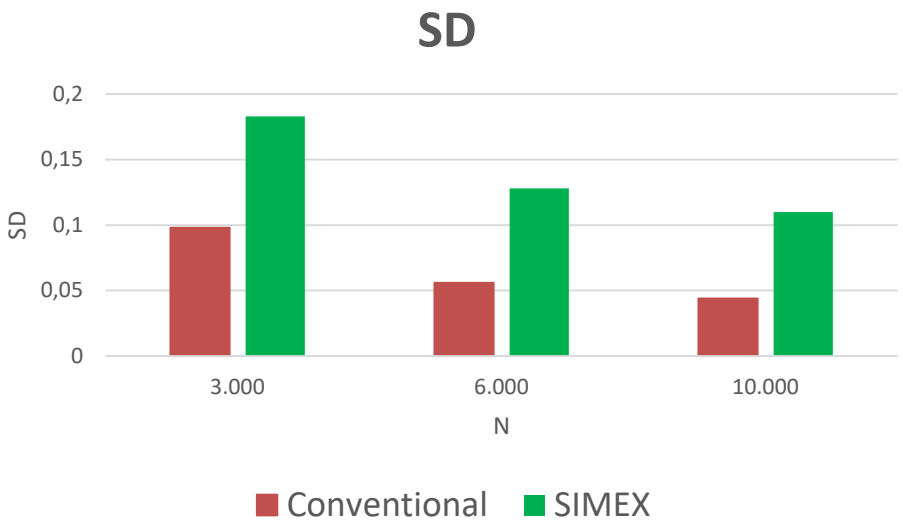


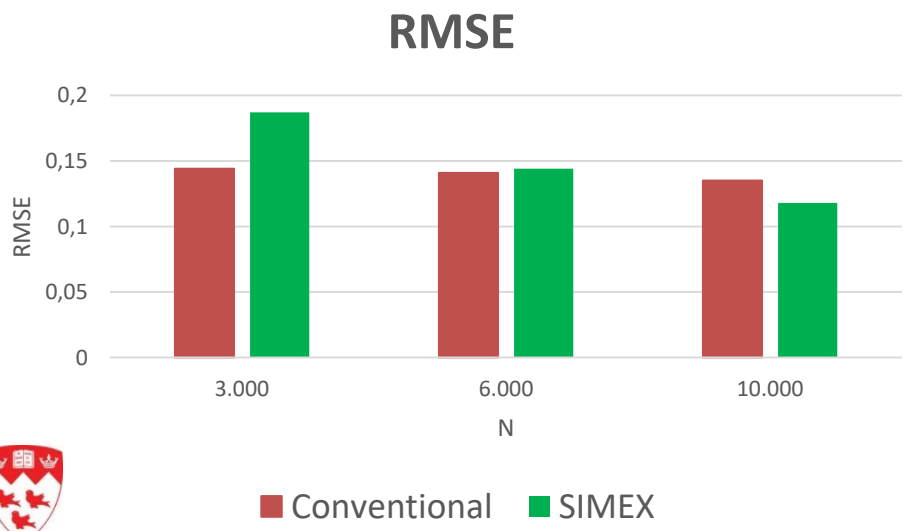


Bias/Variance trade-off (RMSE ratio) depends strongly on the IV Strength (X-axis). IV estimates are Un-Biased but have Higher SD's, which decrease with stronger instrument. *Using only 1 or 2 values of IV strength could give a wrong picture!*...

SIMEX correction for Measurements Errors in Outcome: **Bias/Variance Trade-off** (relative RMSE) **depends on N**







Here, **Bias/Variance trade-off (RMSE ratio) of SIMEX vs Conventional estimates depends strongly on sample size ("N" on X-axis)**. SIMEX reduces Bias but increases SD, which become smaller for larger N's, resulting in **lower RMSE only at N=10,000**. Again, *using only 1 or 2 values of N could give a wrong picture!*..."

"Digging deeper" into Simulation Results: Beyond the Average performance

- Results for SIMEX vs Conventional (previous Slide)
- Interpretation: **lower "mean" error (RMSE) does NOT imply one method always better than another** (e.g. SIMEX "wins" in 34% of samples for N=3K even if RMSE higher by ~ 35%)
- For **N=6K RMSE's equal** but **SIMEX wins in 66% of samples** $\Rightarrow$ Variance Inflated by '*outliers*' $\Rightarrow$ ***hint* to improve the method?**

N	RMSE Convent.	RMSE SIMEX	% samples SIMEX Closer to True β
3,000	0.14	< 0.19	34%
6,000	**0.14**	**= 0.14**	**66%**
10,000	0.14	> 0.12	72%

Flexible Weighted Cumulative Exposure (WCE) modeling of **Cumulative effects of Time-Varying exposures** on hazard

$$WCE\left(u\right)=\sum_{t\le u} w\left(u-t\right)* X\left(t\right)$$

u= current time (when Risk is being assessed)
X(t)= exposure intensity (dose) at time *t(t≤u)*
WCE(u)= **Weighted Cumulative Effect of the Past Exposures** on hazard at time *u,*
defined as a ***Weighted Sum of Past Exposures***
u-t= time elapsed since exposure *X(t)*
w(u-t)*= estimated Weight (Relative Importance) assigned to exposure *X(t)* as a function of Time-since-Exposure *(u-t) **

** weight function *w(u-t)* is modeled using Cubic Splines

>> **Next Slide evaluates the Accuracy of the 'individual' Weight function estimates (from 100 replicates) in 6 alternative simulated scenarios**

[Sylvestre & Abrahamowicz, *Statistics in Medicine* 2009]

Digging deeper: **100 individual *w(u-t)* estimates (black) vs their Mean (white)** for 6 scenarios (~250 events)

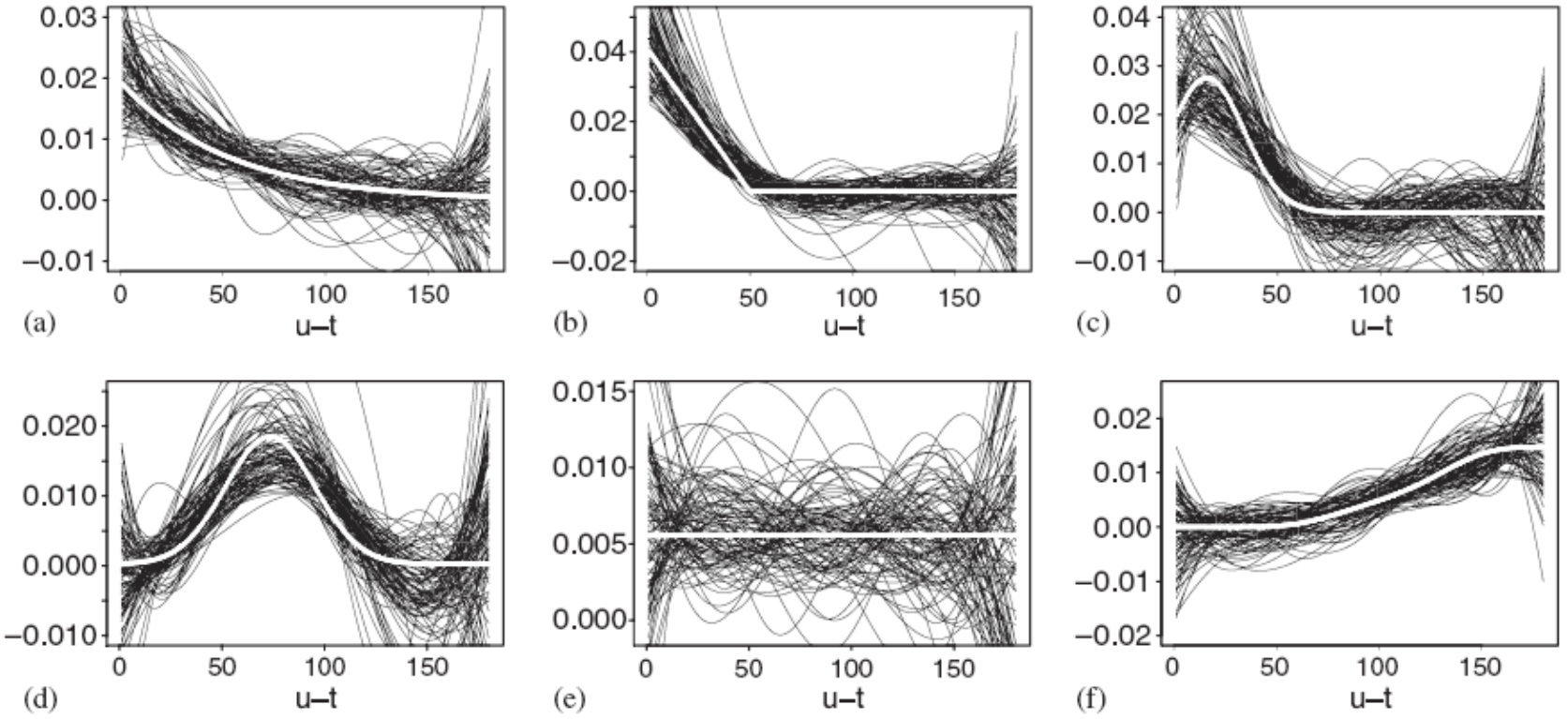


Figure 1. A random sample of 100 normalized estimated weight functions for the unconstrained models with the true weight function in thick white: (a) exponential; (b) bi-linear; (c) early peak; (d) inverted U; (e) constant; and (f) hat. Note that, to make the label of the X-axis readable, we show time in days, while in the text, we use 1 year as the unit of time, so that the values on the axes should be divided by 365.

From Sylvestre and Abrahamowicz (2009, Statistics in Medicine, Figure 1)

Considering **Unfavorable Assumptions** (2 Last Panels): Huge Bias if *w(u-t)* wrongly constrained to decay to null

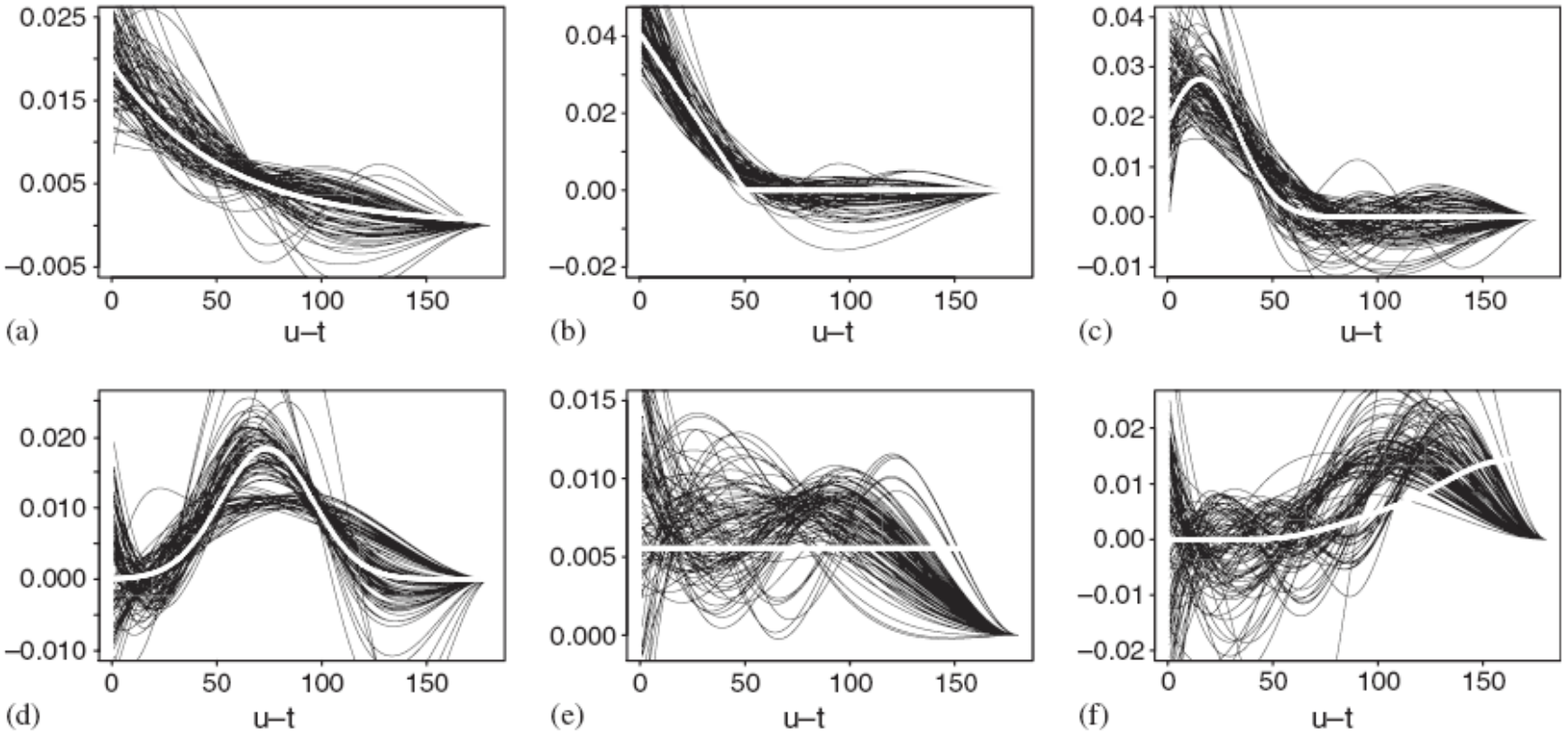


Figure 2. A random sample of 100 normalized estimated weight functions for the constrained models with the true weight function in thick white: (a) exponential; (b) bi-linear; (c) early peak; (d) inverted U; (e) constant; and (f) hat. Note that, to make the label of the X-axis readable, we show time in days, while in the text, we use 1 year as the unit of time, so that the values on the axes should be divided by 365.

From Sylvestre and Abrahamowicz (2009, Statistics in Medicine, Figure 2)

"**Tip of the Iceberg?**": exploring RARE but DISASTROUS

Results (MSM Cox in **[Xiao, Moodie & Abrahamowicz, *Epi Methods 2013*]**)

- **Original Reasons for "Alert"** (re Cox Marginal Structural Models (MSM)):
  In simulations, **Cox MSM-based Hazard Ratios (HR), with stabilized IPT weights,** were (on average) almost un-biased but **had surprisingly high variance** and, thus, **worse RMSE than conventional Cox HR's**
- This triggered *Additional Exploration of individual estimates*, revealing that Variance Inflation was due to a few samples with very biased estimates
- *In-deep assessment* of these "outlying' simulated samples showed that each contained a single 'mis-fit observation' with a Huge IPT weight (IPTW>500)
- Consistent with Survival Analysis theory, the impact of high IPTW's was especially 'dramatic' if they were associated with an Event (rather than a censored observation)

- **To better understand the issue, we undertook Bootstrap Stability investigation of these outlying samples (based on 100 resamples)**

Bootstrap stability analyses for 2 “extreme” samples:

Huge Bias IFF subject with extreme IPTW & event (Y=1) included (**dark dots**)!

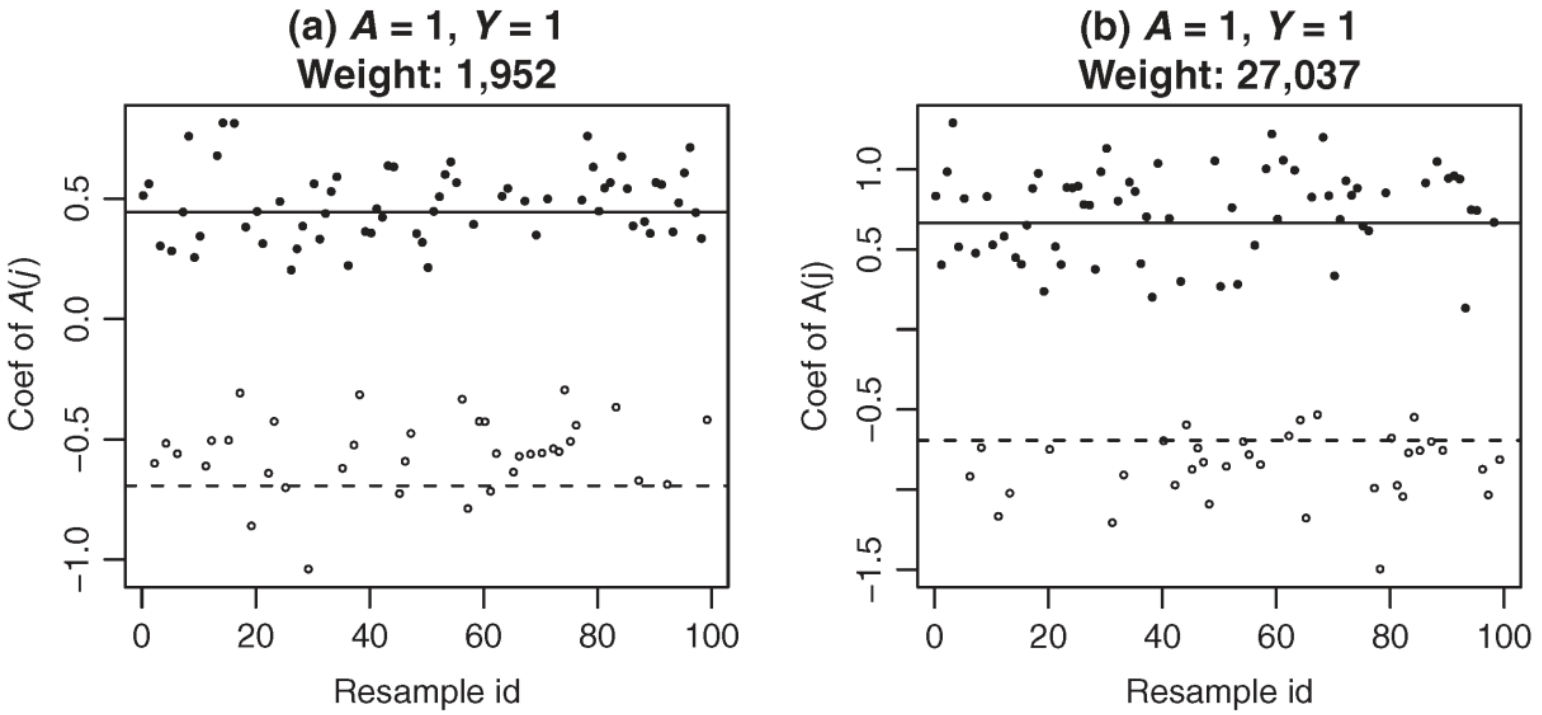


Figure 1. Investigation of the impact of the extreme weights on the IPTW estimates of A(j) using resampling. Dark dots indicate that the subject with the highest weight in the original sample was included in a resample. The solid line represents the estimated effect of A(j) in the original sample and the dashed line represents the true effect. The treatment (A), the event (Y), and the weight for the observation with the highest weight are indicated in the title of each panel.

Selected "Recommendations"

In Preliminary Simulations:

- **Test Data-Generation procedures using "Ideal Case" scenario** with *a priori* "Known" (Theory-based) results
- **Use "Bracketing" (1 or 2 parameters at a time) to Identify "crucial parameters'** that need to be Varied in "Main Sims"

In "Main Simulations":

- **Consider Alternative (Plausible) Assumptions** (about Data Structure/True model etc.) including those where:

(a) assumptions **"favor" different** among the "**Competing Methods**" **
[** IF you propose a "new" Complex Method (e.g. Non-linear estimate) include scenario where "true model' is Simple (e.g. Linear effect), so "Complexity" is NOT necessary ...]

(b1) **your Proposed Method is expected to Fail,**

and/or (b2) **None of the Methods considered is Expected "to work"**

Selected "Recommendations" (continued ...)

In "Main Simulations":

- **Assess** Not only Bias but also **Variance & Mean Squared Error**
  (Bias/Variance trade-off may vary dramatically with N !)
- **If your estimands are Functions, show Not only the Mean** Estimate **but also the estimates from Individual Simulated Samples**
- **Do NOT Ignore rare but Strange results** (Tip of the Iceberg?)
- If relevant, **Design New *Post-hoc* Simulations to explore issues revealed by Unexpected results of pre-planned simulations**
- **Avoid "Publication Bias" & Report Results of All Simulations you've performed** (use *Supplementary Materials* if needed)

Selected References

- **Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine* 2013; 32: 2262-2277.**
- **Boulesteix AL, Binder H, Abrahamowicz M, Sauerbrei W. Simulation Panel of the STRATOS Initiative. On the necessity and design of studies comparing statistical methods. *Biometrical Journal* 2018; 60(1): 216-18.**
- **Ionescu-Ittu R, Delaney JAC, Abrahamowicz M. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. *Pharmacoepidemiology and Drug Safety* 2009; 18(7): 562-571.**
- **Kyle R, Moodie E, Klein M, Abrahamowicz M. Correcting for measurement error in time-varying covariates in marginal structural models. *Am J Epidemiology*. 2016 Aug;184(3):249-58.**
- **Morris TP, White IR, Crowther MJ. 2019. Using simulation studies to evaluate statistical methods. Stati Med 38:2074-2102.**
- Sylvestre MP, Abrahamowicz M. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine* 2008; 27(14): 2618-2634.
- **Sylvestre M-P, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine* 2009; 28(27): 3437-3453.**
- **Xiao Y, Moodie EEM, Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods* 2013; 2(1): 1-20.**

THANK YOU

- michal.abrahamowicz@mcgill.ca