Advantages of using Data-driven Simulations in Real-World Epidemiology Research

Michal Abrahamowicz *[1] , **Anna-Laure Boulesteix** [2], **Tim Morris** [3] , **Willi Sauerbrei** [4] , **Jay Kaufman** [1] , **Marie-Eve Beauchamp** [1]

(for STRATOS Simulation Panel)

[1] McGill University, Montreal, Canada
[2] University of Munich, Germany
[3] MRC clinical trials unit at UCL, London, UK
[4] University of Freiburg, Germany

STRATOS INITIATIVE

McGill

Primary Objective

- Overall Goal:

to promote more wide-spread use of well designed

‘targetted’ * Data-driven ** Simulations in the context of

Real-World Epidemiological studies

- * ‘targetted’ = designed to gain insights re: a specific ‘analytical challenge’ relevant for (i.e. encountered in) a particular real-world empirical study (as opposed to a ‘generic’ methodological issue)
- ** ‘Data-driven’ = designed to account for the actual features of the particular dataset used in a given empirical study

Target Audience

- STRATOS "Level 2" paper
- Oriented toward **Practicing Epidemiologists, Graduate Students in Epidemiology and Applied Statisticians** (e.g. working as Research Assistants/Associates for Epidemiology Researchers) **with good understanding of Applied Statistics** and good Computational Skills in basic programming & Use of Statistical Software**

  (NO need to understand Statistical Theory)**
- possible **Target Journal(s):**

➢ *American Journal of Epidemiology* ('Practice of Epidemiology' section), OR

➢ *International Journal of Epidemiology* (Education corner)

Outline of the General Approach

- Reliance on 3-4 Concrete Empirical Examples related to Real-World Epidemiological Studies **
- Underlying assumption: ** **'Data have been already collected & are available**
- Use of a Mix of Real-World Data and 'Synthetic Data' (purposeful modifications of real data to better illustrate specific methodological issues, similar e.g. to the approach of TG7 *Simulation Learner*)
- Focus on Observational Studies of Association
- Each example focuses on a different analytical issue (e.g. assessing the expected impact of a specific '*data imperfection*')

General Principles of Simulations Design and Reporting

- **'*X matrix*' of Covariates & Exposure 'copied' from Empirical data**

(possibly with some purposeful modifications in 'sensitivity analyses')

- **Outcomes, conditional on *X*, simulated from the user-specified 'true model'**** (Challenge: to ensure Neutrality of the simulations [1])
- ** when relevant, the empirical distribution of Event Times may be also 'copied' and linked to *X* matrix, using Permutational Algorithm [2]
- **Empirical or Simulated Data & Codes for Simulations available for the readers** (through Supp. Mat.) to ensure **Reproducibility [1]**
- **Simulations will be reported using the ADEMP structure [3]**

[1] = Boulesteix et al, *Biom J* 2018; [2] = Sylvestre & Abrahamowicz, *StatMed* 2008;

[3] = Morris, White, Crowther, *StatMed* 2019