Key Theme 3: Simulating and synthesizing data for biostatistical methods development, evaluation and dissemination

Georg Heinze, Sabine Hoffmann, Tim Morris

16 Sep 2024

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
  - ADEMP:
    Aims - Data generation - Estimands - Methods - Performance measures

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
  - ADEMP:
    Aims - Data generation - Estimands - Methods - Performance measures
- Neutral comparison studies [Boulesteix et al., 2024]

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
  - ADEMP:
    Aims - Data generation - Estimands - Methods - Performance measures
- Neutral comparison studies [Boulesteix et al., 2024]
- Phases of methodological development [Heinze et al., 2024]

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
    - ADEMP:
      Aims - Data generation - Estimands - Methods - Performance measures
- Neutral comparison studies [Boulesteix et al., 2024]
- Phases of methodological development [Heinze et al., 2024]
    - Phase I: Introduce new method, demonstrate validity through mathematical derivations

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
  - ADEMP:
    Aims - Data generation - Estimands - Methods - Performance measures
- Neutral comparison studies [Boulesteix et al., 2024]
- Phases of methodological development [Heinze et al., 2024]
  - Phase I: Introduce new method, demonstrate validity through mathematical derivations
  - Phase II: Demonstrate use of method with real data in a limited range of possible applications

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
  - ADEMP:
    Aims - Data generation - Estimands - Methods - Performance measures
- Neutral comparison studies [Boulesteix et al., 2024]
- Phases of methodological development [Heinze et al., 2024]
  - Phase I: Introduce new method, demonstrate validity through mathematical derivations
  - Phase II: Demonstrate use of method with real data in a limited range of possible applications
  - Phase III: Compare relatively new method with competitors and demonstrate use in practice for a wide range of applications

Background: Simulation studies

- Framework for reporting simulation studies [Morris et al., 2019]
    - ADEMP:
      Aims - Data generation - Estimands - Methods - Performance measures
- Neutral comparison studies [Boulesteix et al., 2024]
- Phases of methodological development [Heinze et al., 2024]
    - Phase I: Introduce new method, demonstrate validity through mathematical derivations
    - Phase II: Demonstrate use of method with real data in a limited range of possible applications
    - Phase III: Compare relatively new method with competitors and demonstrate use in practice for a wide range of applications
    - Phase IV: Summarise evidence about method, also in comparison with competing methods, uncover previously unknown behaviour of method in complex data analyses and consider extended range of applications

Background: Generation of synthetic data

- Increasing awareness that data sharing:
  - Improves transparency, credibility and reproducibility
  - Makes evidence synthesis more efficient
  - Increases reuse potential of scientific studies

Background: Generation of synthetic data

- Increasing awareness that data sharing:
    - Improves transparency, credibility and reproducibility
    - Makes evidence synthesis more efficient
    - Increases reuse potential of scientific studies
- ⇒ Journals and funders are increasingly incentivizing or even requiring data sharing practices

Background: Generation of synthetic data

- Increasing awareness that data sharing:
    - Improves transparency, credibility and reproducibility
    - Makes evidence synthesis more efficient
    - Increases reuse potential of scientific studies
- ⇒ Journals and funders are increasingly incentivizing or even requiring data sharing practices
- ⇒ Many researchers lack skills and knowledge to make their data publicly available while protecting confidentiality

Phase III neutral simulation study

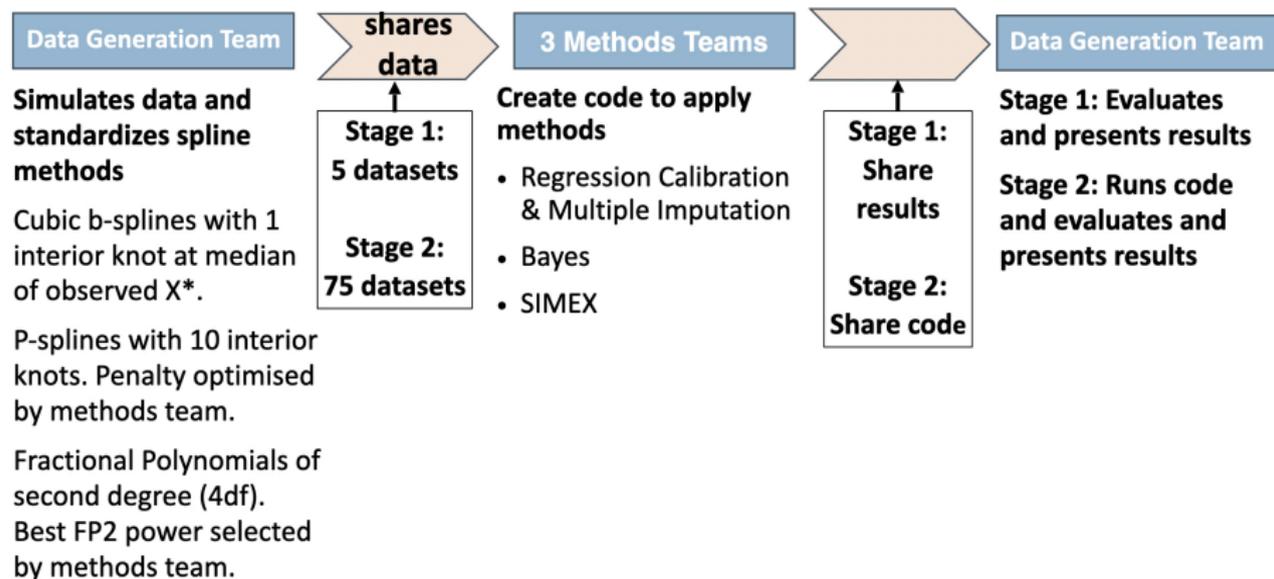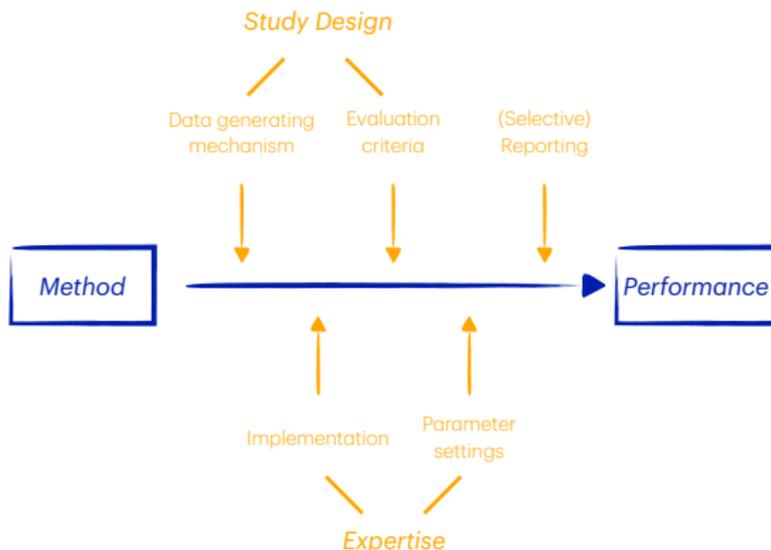Laurence Freedman, Monday 3:30 pm

Figure: TG2-TG4 collaboration on measurement error and splines

Phase IV neutral simulation study

Sabine Hoffmann, Monday 4:00 pm

- Correcting complex patterns of measurement error in occupational epidemiology
- Joint project with Anne-Laure Boulesteix, Nicole Ellenbach and Christina Sauer

Diagnostic for phase IV methodological research

Tim Morris, Tuesday 11:00 am

- Joint project with Ian White
- Background: Some trial statisticians prespecify identity-link binomial family GLMs for covariate-adjusted analysis of trials with a binary outcome.
- Aim: Develop and evaluate a diagnostic to guide the use of this method
- Tim will describe some of the ideas of this project in the hope that we can go from the specific to the general in discussing diagnostics

Plenary morning session

Wednesday, 9:00 am

- 9:00 am - Georg Heinze: Phases of methodological development in biostatistics

- 9:20 am - Tim Morris: Structured planning and reporting of simulation studies

- 9:40 am - Sabine Hoffmann: Making biomedical research data publicly available while protecting confidentiality

Chair: Willi Sauerbrei

Synthetic data generation

Sabine Hoffmann, Wednesday 11:00 am

- Joint work with Sarah Friedrich, Jan Kapar and Marvin Wright

- How to share biomedical research data?
- How to evaluate the quality of the shared data set in terms of disclosure risk and utility?
- What are the possible uses of shared data sets?

- Discuss desirable add-ons (missing data, functional forms, measurement error, counterfactuals etc.)

Further topics

- Presentation/visualisation of complex simulation results – discussion suggested by Daniela Dunkler with input from Theresa Ullmann, Michael Kammer and Tim Morris
- Simulation profiles – suggested by Willi Sauerbrei
  Short structured reporting of simulation studies
- Preregistration and peer review of simulation protocols – suggested by Willi Sauerbrei
  Example of STRATOS TG2 (Ullmann et al, PlosOne, 2024)

Overview

	Mon	Tue	Wed	Thu	Fri
9:00 – 10.00 .	9.00-9.45 walk-in 9.45-10.00 Welcome from Lorentz center 10.00-10.30 STRATOS news Aims for the week	Presentation + discussion key theme 2*: 'Estimands and time to event analysis'	Phases of methodological development *(Georg Heinze)* Structured planning and reporting of simulation studies *(Tim Morris)* Data sharing *(Sabine Hoffmann)*	Presentation + discussion key theme 4*: 'Curation plan for case study data '	Plenary Feedback parallel working sessions on key themes
Break 10:00-10.30					
10.30-12.30	10.30-11.30 Four pitches on key themes	10.30-11.00 STRATOS-wide topics	10.30-11.00 STRATOS-wide topics	10.30-11.00 STRATOS-wide topics	What have we learned / what are next steps / action points?
	11.30-12.30 Presentation + discussion key theme 1*: 'Machine Learning'	11.00-12.30 *Diagnostics for phase IV studies (Tim Morris)*	11.00-12.30 *Synthetic data generation (Sabine Hoffmann)*	11.00-12.30 *Visualisation of complex simulation results (Daniela Dunkler, Michael Kammer, Tim Morris, Theresa Ullmann)*	
Lunch 12:30-14:00					
14.00-15.00	TGs discuss and distribute their ideas for the key themes	Continue on key themes / Or work in TGs	Continue on key themes / Or work in TGs	Continue on key themes / Or work in TGs	END
Break 15:00-15,30					
15:30-16:30	*Phase III and IV neutral simulation studies with focus on measurement error correction (Laurence Freedman + Sabine Hoffmann)*	TG-specific progress and completion of existing projects	TG-specific progress and completion of existing projects	TG-specific progress and completion of existing projects	
16:30-17:30		Short presentations from each TG and some panels	TG-specific progress and completion of existing projects	TG-specific progress and completion of existing projects	
17:30-19:00			Debate:	TG-specific	

References

A. Boulesteix, M. Baillie, D. Edelmann, L. Held, T. P. Morris, and W. Sauerbrei. Editorial for the special collection 'Towards neutral comparison studies in methodological research'. *Biometrical Journal*, 66(2):2400031, 2024. doi: 10.1002/bimj.202400031.

G. Heinze, A.-L. Boulesteix, M. Kammer, T. P. Morris, I. R. White, and the Simulation Panel of the STRATOS initiative. Phases of methodological research in biostatistics – Building the evidence base for new methods. *Biometrical Journal*, 66(1):2200222, 2024. doi: 10.1002/bimj.202200222.

T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.