An overview and recent developments of the STRATOS Open Science panel

Sabine Hoffmann on behalf of the Open Science panel

28.08.2025

Overview

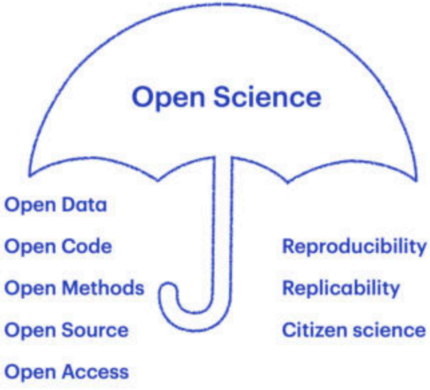
1. What is open science and why do we need it?
2. Synthetic data generation to make biomedical research publicly available while protecting confidentiality
3. Guidance on how to deal with research degrees of freedom in the analysis of observational data

What is open science and why do we need it?

What is open science?

What is open science?

"Umbrella term that reflects the idea that scientific knowledge of all kinds, where appropriate, should be openly accessible, transparent, rigorous, reproducible, replicable, accumulative and inclusive"



nature human behaviour

Explore content | About the journal | Publish with us | Subscribe

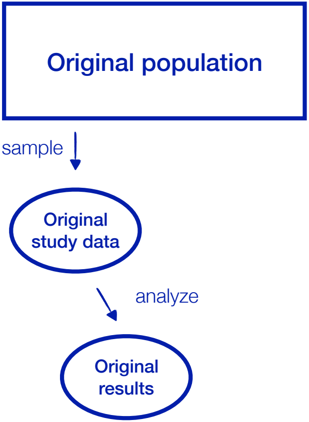
nature > nature human behaviour > comment > article

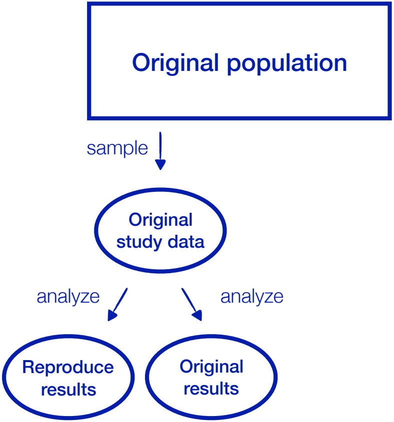
Comment | Published: 21 February 2022

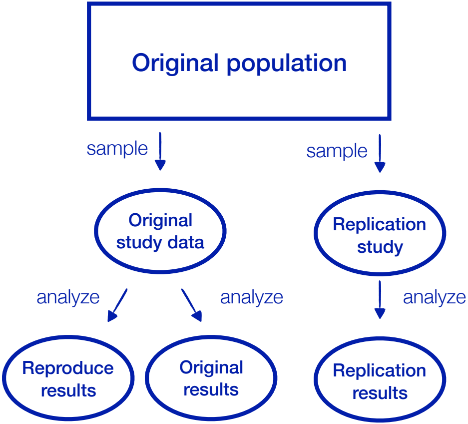
A community-sourced glossary of open scholarship terms

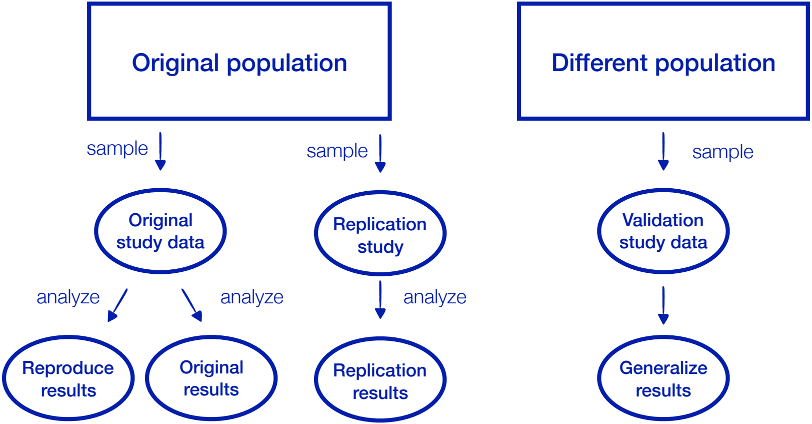
Sam Parsons, Flávio Azevedo, Mahmoud M. Elsherif, Samuel Guay, Owen N. Shahim, Gisela H. Govaart, Emma Norris, Aoife O'Mahony, Adam J. Parker, Ana Todorovic, Charlotte R. Pennington, Elias Garcia-Pelegrin, Aleksandra Lazić, Olly Robertson, Sara L. Middleton, Beatrice Valentini, Joanne McCuaig, Bradley J. Baker, Elizabeth Collins, Adrien A. Fillon, Tina B. Lonsdorf, Michele C. Lim, Norbert Vanek, Marton Kovacs, ... Balazs Aczel + Show authors

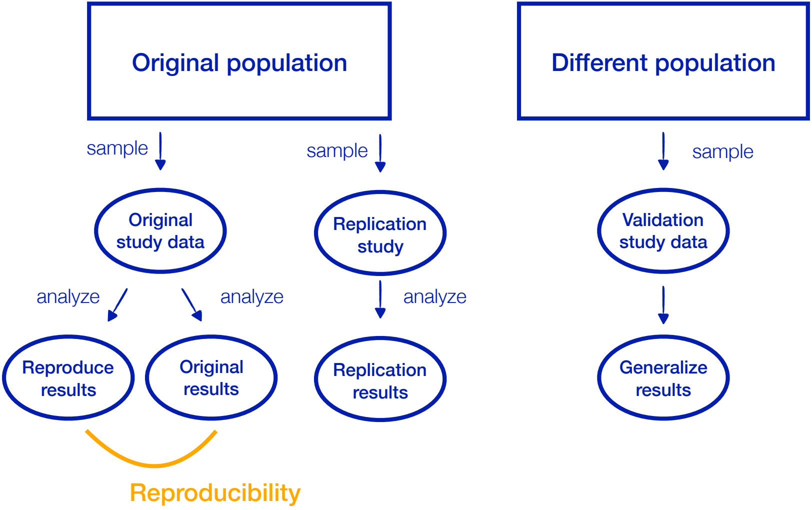
Nature Human Behaviour 6, 312–318 (2022) | Cite this article

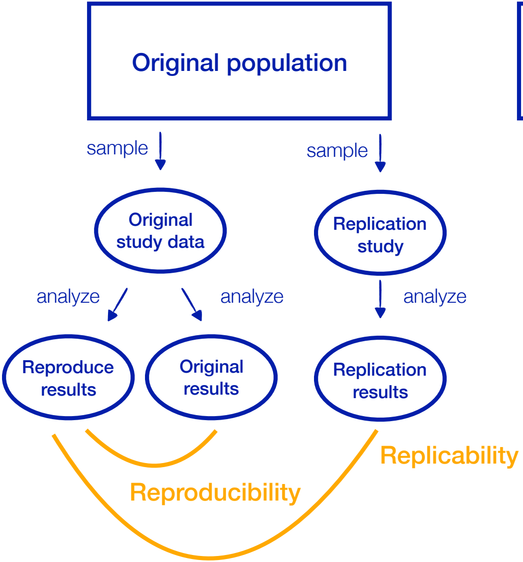

Original population
sample
Original study data
analyze
Original results

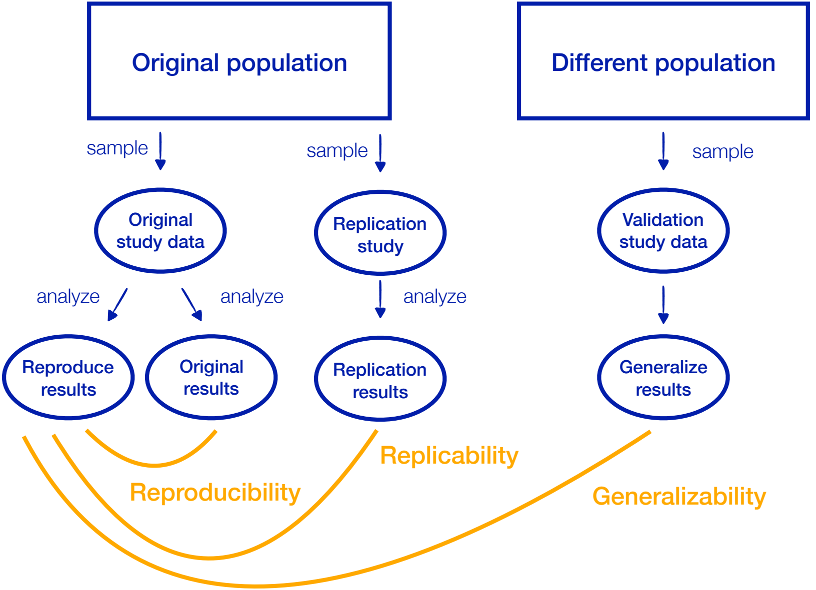

Original population
sample
Original study data
analyze
analyze
Reproduce results
Original results


Original population
sample
sample
Original study data
Replication study
analyze
analyze
analyze
Reproduce results
Original results
Replication results


Original population
Different population
sample
sample
sample
Original study data
Replication study
Validation study data
analyze
analyze
analyze
Reproduce results
Original results
Replication results
Generalize results


Original population
Different population
sample
sample
sample
Original study data
Replication study
Validation study data
analyze
analyze
analyze
Reproduce results
Original results
Replication results
Generalize results
Reproducibility


Original population
Different population
sample
sample
sample
Original study data
Replication study
Validation study data
analyze
analyze
analyze
Reproduce results
Original results
Replication results
Generalize results
Replicability
Reproducibility


Original population
Different population
sample
sample
sample
Original study data
Replication study
Validation study data
analyze
analyze
analyze
Reproduce results
Original results
Replication results
Generalize results
Reproducibility
Replicability
Generalizability

Why do we need open science in clinical research?


Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

Modeling the Framework for False Positive Findings


To keep health as a unifying force, we must put resources into tackling health misinformation and disinformation

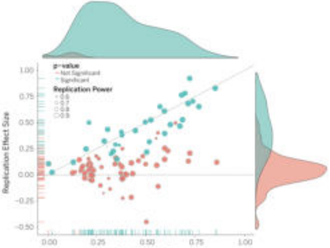
Published Online October 16, 2024

Health is political. This is what many practitioners of public and clinical health believe. Health and health policy are shaped by the political ideology of governments, whether that means more money to invest in health systems or less regulation on health-harming products. Health can also cut across political lines because health is a universally shared value. Everyone wants their loved ones to be healthy, so framing societal issues as health issues can draw people from across the political spectrum to advocate for change and policies. The health community has had successes using this strategy with, for example, the climate crisis and gun violence. Framing climate

themselves apart from opponents through differentiating policies. Yet because health often affects questions of bodily and individual autonomy, it is also vulnerable to weaponisation through the shaping of narratives.

As such, health can be used to ignite topics that are emblematic of broader societal conversations on the role of government and the state. In her book, Doppelganger: A Trip into the Mirror World,[1] Naomi Klein argues that this phenomenon is more potent now because of the rising numbers of people who feel left behind and abandoned from decades of free market economics that have prioritised profits over the wellbeing of individuals.

Replication crisis in psychology
Open Science Collaboration (2015)



Preclinical research
Freedman et al. (2015)

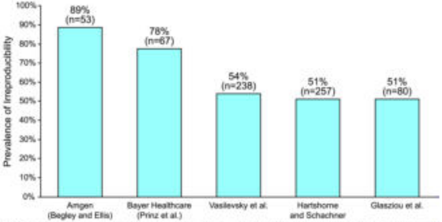


Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [5], and Glasziou et al. [9].

STRATOS and open science

	STRATOS members	Research community
Open access	Ideally, STRATOS publications should be open access	Underline importance of open access publications
Reproducibility	STRATOS papers should be reproducible (reproducibility checks: RH, FS)	Guidance reproducibility for level 1 audience
	STRATOS papers should use open access data sets	Guidance data sharing while protecting confidentiality
Transparency	Write study protocols (e.g. simulation protocol) and ask for community feedback	Open science practices to improve neutrality simulation studies
Replicability		Guidance dealing with uncertain choices for level 1 audience

Chairs: Sabine Hoffmann and Daniela Dunkler (since July 2025)
Members: Anne-Laure Boulesteix, Roman Hornung, Michael Kammer, Kim Luijken, Willi Sauerbrei, Fabian Scheipl, Pamela Shaw, Ewout Steyerberg

Synthetic data generation to make biomedical research publicly available while protecting confidentiality

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

Motivation

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

Motivation

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

$\Rightarrow$ Journals and funders are increasingly incentivizing or even requiring data sharing practices

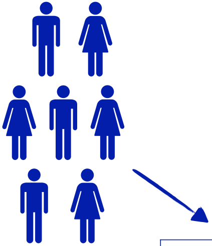
Motivation

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

$\Rightarrow$ Journals and funders are increasingly incentivizing or even requiring data sharing practices

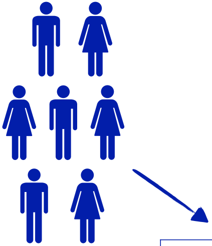
$\Rightarrow$ Many researchers lack skills and knowledge to make their data publicly available while protecting confidentiality

Challenges when making research data publicly available



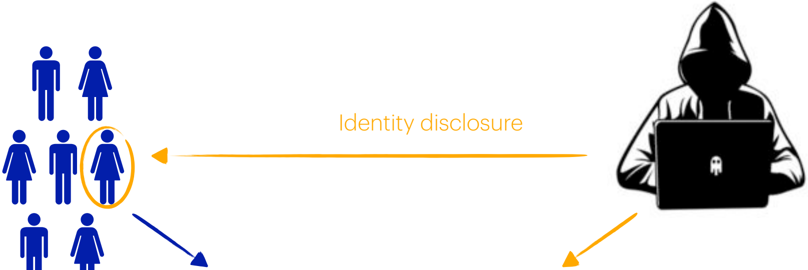
Name	Date of birth	Location	Height	Diagnosis	Prescription	Heart rate

Challenges when making research data publicly available



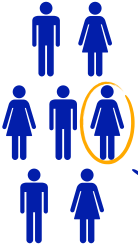
Id	Age	Location	Height	Diagnosis	Prescription	Heart rate

Challenges when making research data publicly available



Id	Age	Location	Height	Diagnosis	Prescription	Heart rate

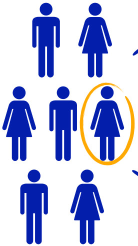
Challenges when making research data publicly available



Attribute disclosure

Id	Age	Location	Height	Diagnosis	Prescription	Heart rate

Challenges when making research data publicly available

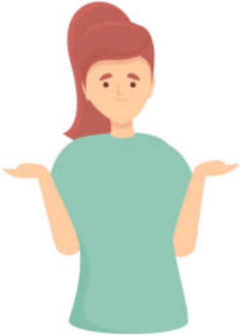


Identity disclosure

Attribute disclosure

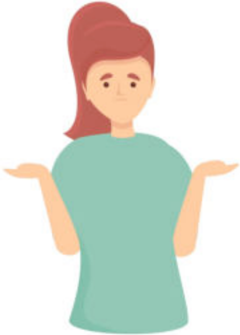
Id	Age	Location	Height	Diagnosis	Prescription	Heart rate

Challenges when making research data publicly available



- How to share biomedical research data?

Challenges when making research data publicly available



- How to share biomedical research data?
- How to evaluate the shared data set in terms of disclosure risk and utility?

Approaches to limit statistical disclosure

Reduce information

Id	Age	Location	Height	Diagnosis	Prescription	Heart rate

Approaches to limit statistical disclosure

Reduce information

Id	Age	Location	Height	Diagnosis	Prescription	Heart rate

- Dropping variables

Approaches to limit statistical disclosure

Reduce information

Id	Age	Location	Height	Diagnosis	Prescription	Heart rate
	20-25					
	55-60					
	90-95					

- Dropping variables
- Categorizing continuous variables or aggregating categories

Approaches to limit statistical disclosure

Reduce information

Id	Age	Location	Height	Diagnosis	Prescription	Heart rate
	20-25					
	55-60					
	90-95		> 2 meters			

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Approaches to limit statistical disclosure

Reduce information

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Data perturbation

- Data swapping
- Adding noise

Approaches to limit statistical disclosure

Reduce information

- Dropping variables
- Categorizing continuous variables or aggregating categories
- Censoring

Data perturbation

- Data swapping
- Adding noise

Generate synthetic data

Approaches to limit statistical disclosure

Reduce information

Data perturbation

- Data swapping
- Adding noise

Generate synthetic data

- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks

Approaches to limit statistical disclosure

Reduce information

Data perturbation

Generate synthetic data

- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks
  - Tree-based methods
    - Synthpop (Nowok et al., 2016)
    - Adversarial Random Forests (Watson et al., 2023)

Approaches to limit statistical disclosure

Reduce information

Data perturbation

Generate synthetic data

- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks
  - Tree-based methods
    - Synthpop (Nowok et al., 2016)
    - Adversarial Random Forests (Watson et al., 2023)
  - Bayesian networks

Approaches to limit statistical disclosure

Reduce information

Data perturbation

Generate synthetic data

- Methods:
  - Parametric methods
  - Deep learning:
    - Autoencoders
    - Generative Adversarial Networks
  - Tree-based methods
    - Synthpop (Nowok et al., 2016)
    - Adversarial Random Forests (Watson et al., 2023)
  - Bayesian networks
- Full or partial synthesis

Evaluating the quality of synthetic data

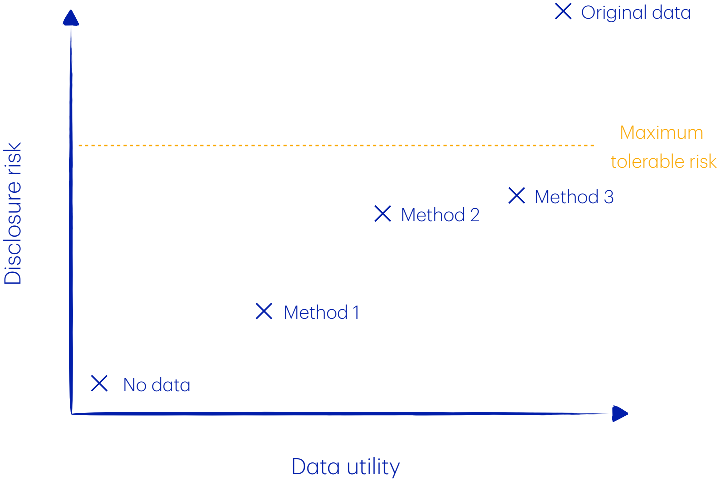


Illustration: Data utility

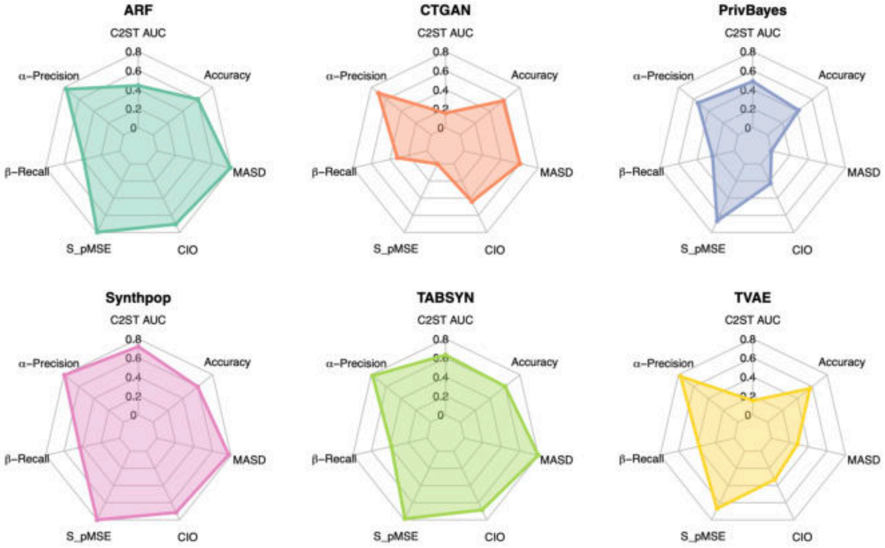
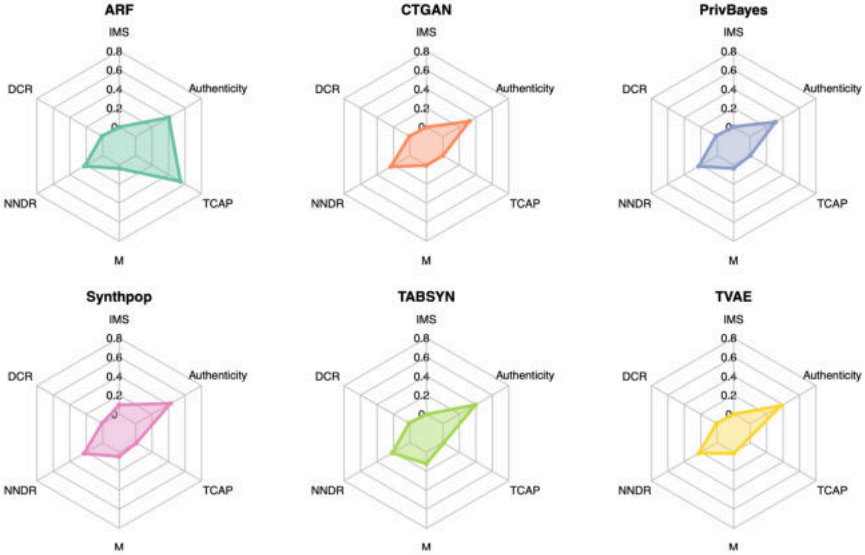


Illustration: Disclosure risk



Open questions

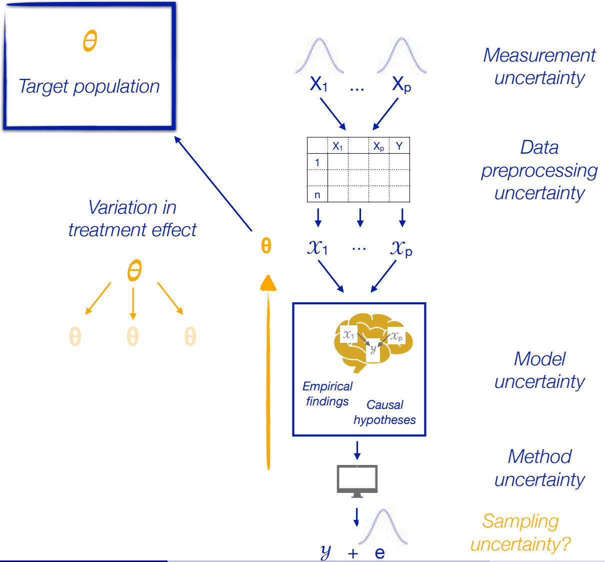
- Does sampling already provide some confidentiality?

Open questions

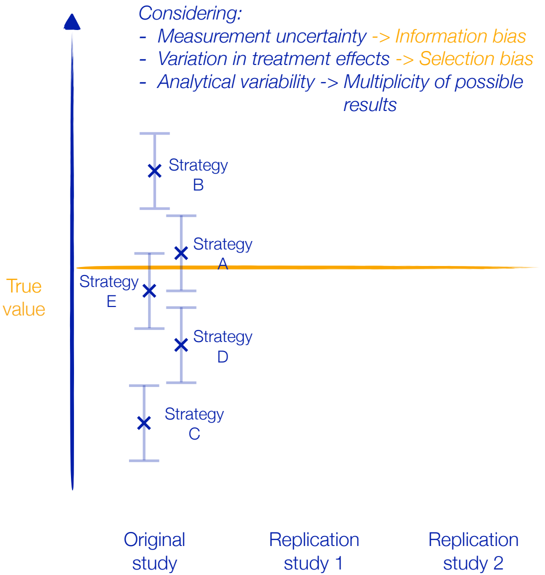
- Does sampling already provide some confidentiality?
- Challenges:
  - Logical constraints between variables
  - Missing data
  - Longitudinal data
  - High-dimensional data

Guidance on how to deal with research degrees of freedom in the analysis of observational data

Researcher degrees of freedom in observational studies



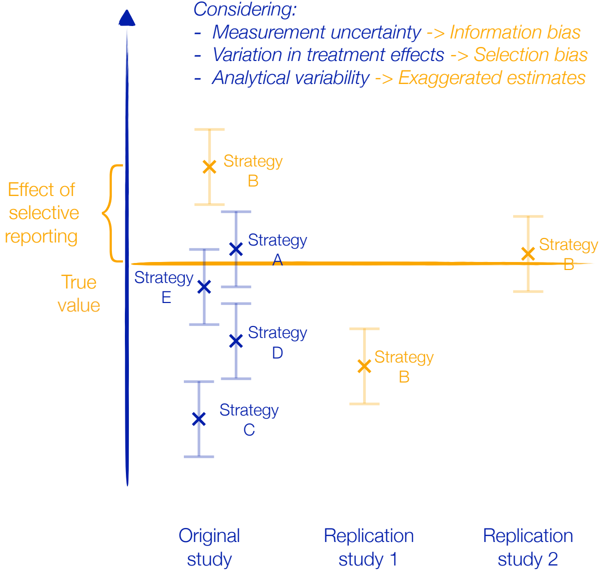
Consequences of selective reporting



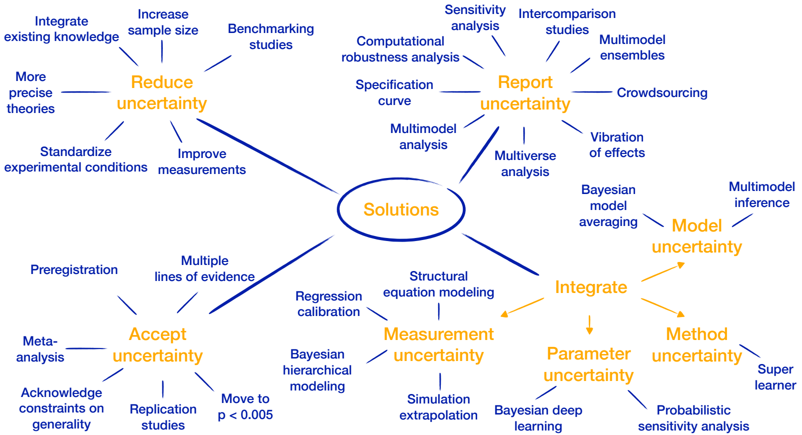
Consequences of selective reporting

Considering:

- *Measurement uncertainty -> Information bias*
- *Variation in treatment effects -> Selection bias*
- *Analytical variability -> Exaggerated estimates*



Dealing with researcher degrees of freedom



Hoffmann, S., F. Schönbrodt, R. Elsas, R. Wilson, U. Strasser, Boulesteix, A. L. (2021). The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. Royal Society Open Science 8 201925

Dealing with analytical choices in the analysis of observational studies (level 1/2 audience)

- Raise awareness of dangers of result-dependent selective reporting of analysis strategies

Dealing with analytical choices in the analysis of observational studies (level 1/2 audience)

- Raise awareness of dangers of result-dependent selective reporting of analysis strategies
- Illustrate solutions to deal with these analytical choices, tailored to observational studies in biomedical research:
  - Pre-registration
  - Increasing statistical power
  - Multiverse analyses, vibration of effects etc.
  - Validation on independent test data
  - Account for analytical variability through Bayesian hierarchical approaches

Association between body composition and cardiovascular disease

Literature review papers published in 2024

Joint work with Heiko Becher, Anne-Laure Boulesteix, Daniela Dunkler, Simon Lemster and Carsten Oliver Schmidt

Association between body composition and cardiovascular disease

Literature review papers published in 2024 → Illustrate solutions to deal with analytical choices on Study of Health in Pomerania (SHIP) data

Joint work with Heiko Becher, Anne-Laure Boulesteix, Daniela Dunkler, Simon Lemster and Carsten Oliver Schmidt

Association between body composition and cardiovascular disease

Literature review papers published in 2024 → Illustrate solutions to deal with analytical choices on Study of Health in Pomerania (SHIP) data → Assess replicability on other data sets, including the German National Cohort (NAKO)

Joint work with Heiko Becher, Anne-Laure Boulesteix, Daniela Dunkler, Simon Lemster and Carsten Oliver Schmidt

Literature review: Exposure and outcome definitions

Body composition

- Waist circumference
- Waist-to-hip ratio
- Percent body fat
- Relative fat mass
- Visceral adiposity index
- Body shape index
- Arm fat-to-lean mass ratio
- Leg fat-to-lean mass ratio
- BMI
- Weight-adjusted-waist index
- Body roundness index
- Chinese visceral adiposity index
- MRI measurement of abdominal subcutaneous adipose tissue

Cardiovascular disease

- **Self-report** of diagnosed cardiovascular disease or events
- Cardiovascular **mortality** (Death certificates)
- **Hospitalisation** for cardiovascular disease or events
- Prescription of **medication, surgery or other procedures** indicating cardiovascular disease
- **Physical examination**: pathological Q wave (ECG)

- Any combination of stroke/heart attack, coronary artery disease/angina/congestive heart failure/other heart problem/acute rheumatic fever/chronic rheumatic heart disease/arterial fibrillation or flutter

Literature review: Eligibility and confounders

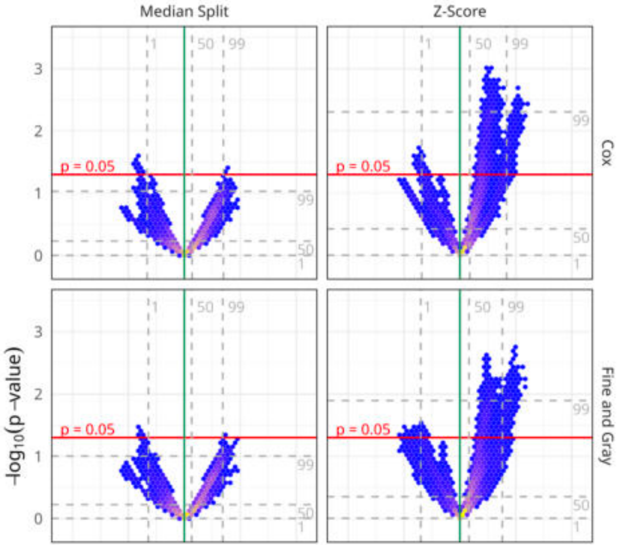
Table 6. Overview of the most frequently used variable groups for eligibility criteria

Variable group	Category	n (%)
Age	demographic	70 (82.4%)
Cardiovascular disease (CVD)	comorbidity	55 (64.7%)
Cancer	comorbidity	20 (23.5%)
Pregnancy	reproductive health	12 (14.1%)
Body mass index (BMI)	body composition	9 (10.6%)
Diabetes	comorbidity	7 (8.2%)
Early death/short follow-up	comorbidity	7 (8.2%)
Kidney disease	comorbidity	6 (7.1%)
General cardiometabolic/chronic disease	comorbidity	6 (7.1%)
Waist circumference (WC)	body composition	6 (7.1%)
Residence in region/community	study design related	5 (5.9%)
Medicine history	general treatment history	4 (4.7%)

Table 5. Overview of the most frequently used adjustment variable categories

Category	n (%)
Age*	79 (92.9%)
Smoking*	74 (87.1%)
Sex*	70 (82.4%)
Alcohol consumption*	62 (72.9%)
Blood pressure*	60 (70.6%)
Glucose metabolism*	59 (69.4%)
Blood lipids*	58 (68.2%)
Education*	51 (60.0%)
Physical activity*	42 (49.4%)
Body mass index (BMI)*	30 (35.3%)
Economic status*	28 (32.9%)
Ethnicity	27 (31.8%)
Kidney function*	27 (31.8%)
Civil status*	24 (28.2%)
Cardiovascular disease history*	22 (25.9%)
Inflammatory markers*	17 (20.0%)
Diet*	16 (18.8%)
Residence/registration	16 (18.8%)
Family history*	13 (15.3%)
Cancer history*	12 (14.1%)

Results of 16 × 2 × 6 × 2 × 3 × 470 = 541.440 analyses



Overview of other projects

- Guidance on reproducibility (Kim Luijken, Michael Kammer, Roman Hornung, Boris Hejblum)

Overview of other projects

- Guidance on reproducibility (Kim Luijken, Michael Kammer, Roman Hornung, Boris Hejblum)
- Improving the neutrality of simulation studies through open science practices

Overview of other projects

- Guidance on reproducibility (Kim Luijken, Michael Kammer, Roman Hornung, Boris Hejblum)
- Improving the neutrality of simulation studies through open science practices
- Adding measurement error to generate synthetic data

Overview of other projects

- Guidance on reproducibility (Kim Luijken, Michael Kammer, Roman Hornung, Boris Hejblum)
- Improving the neutrality of simulation studies through open science practices
- Adding measurement error to generate synthetic data
- Internal guidance for STRATOS projects

Thank you for your attention!

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biol*, 13(6):e1002165.

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11).

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Watson, D. S., Blesch, K., Kapar, J., and Wright, M. N. (2023). Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 5357–5375. PMLR.