Using routinely collected data for research purposes: Challenges and mitigation strategies

Sabine Hoffmann

26.05.2026

1. The role of auxiliary assumptions in the analysis of routinely collected data

2. Challenges in the analysis of routinely collected data
   - Representativeness
   - Data quality
   - Time point alignment
   - Interventions and tests are not random
   - Multiplicity of possible analysis strategies

3. Mitigation strategies

4. Conclusion

The role of auxiliary assumptions in the analysis of routinely collected data

Auxiliary assumptions in research

Auxiliary assumptions in research

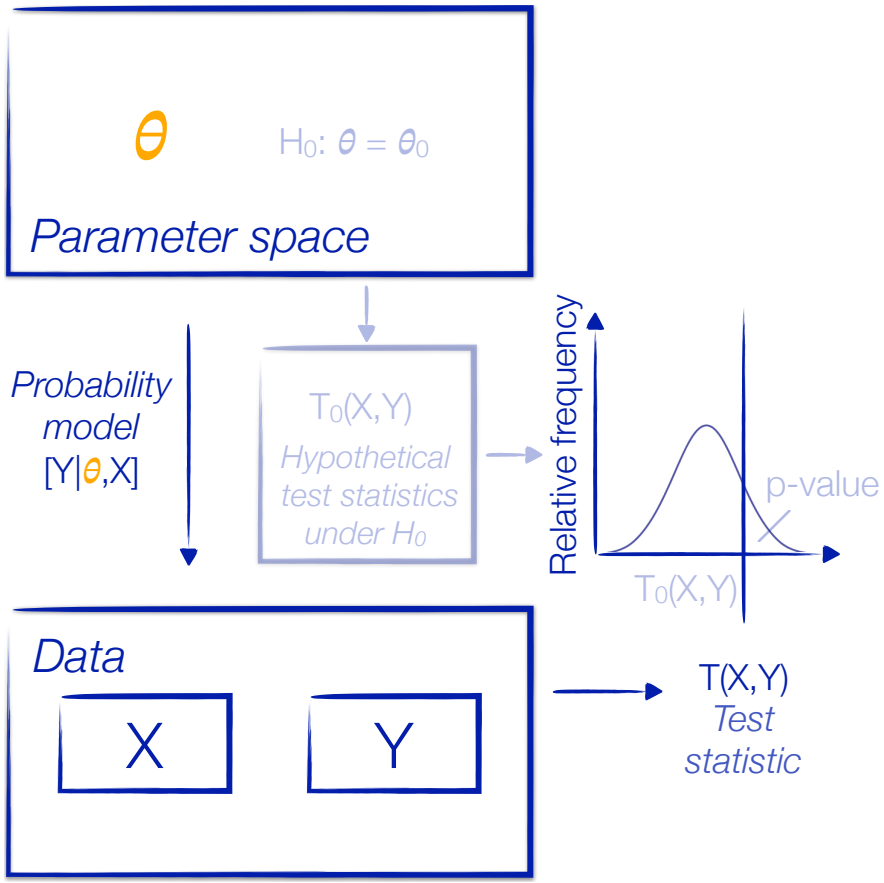
Underdetermination of scientific theory by evidence:

It is not possible to unambiguously falsify a theory, because theories are always tested in a bundle with various auxiliary assumptions.

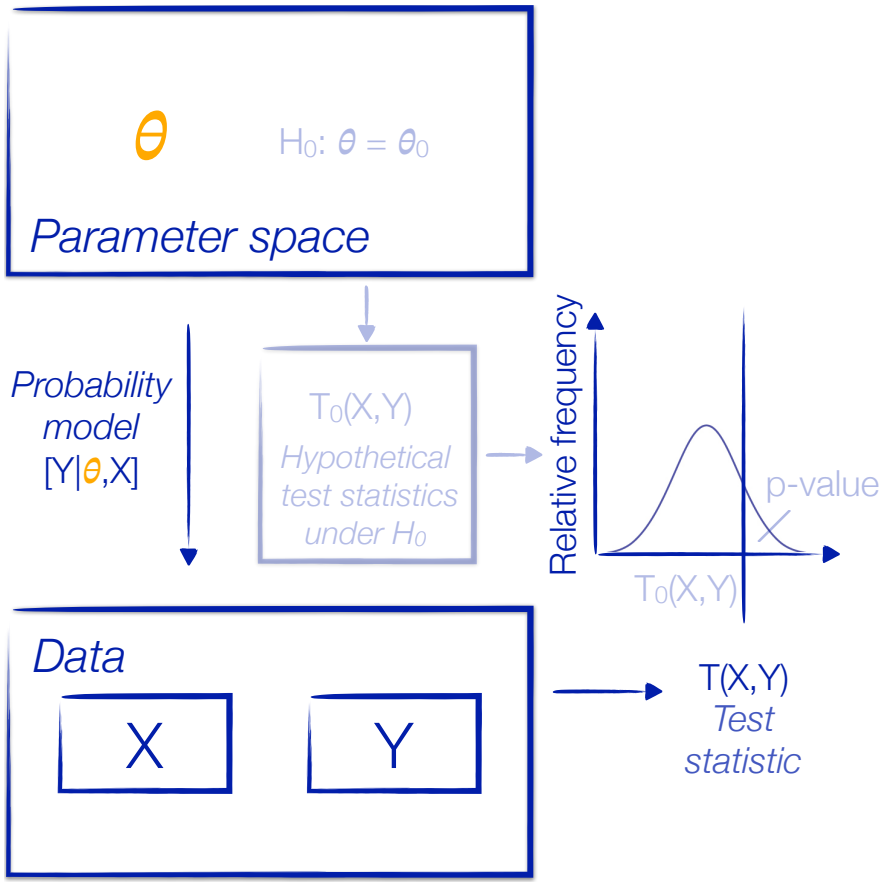
When a prediction fails, we never know whether we should blame the theory or one of the auxiliary assumptions.

(Duhem-Quine problem)

Auxiliary assumptions in statistical modelling in medicine

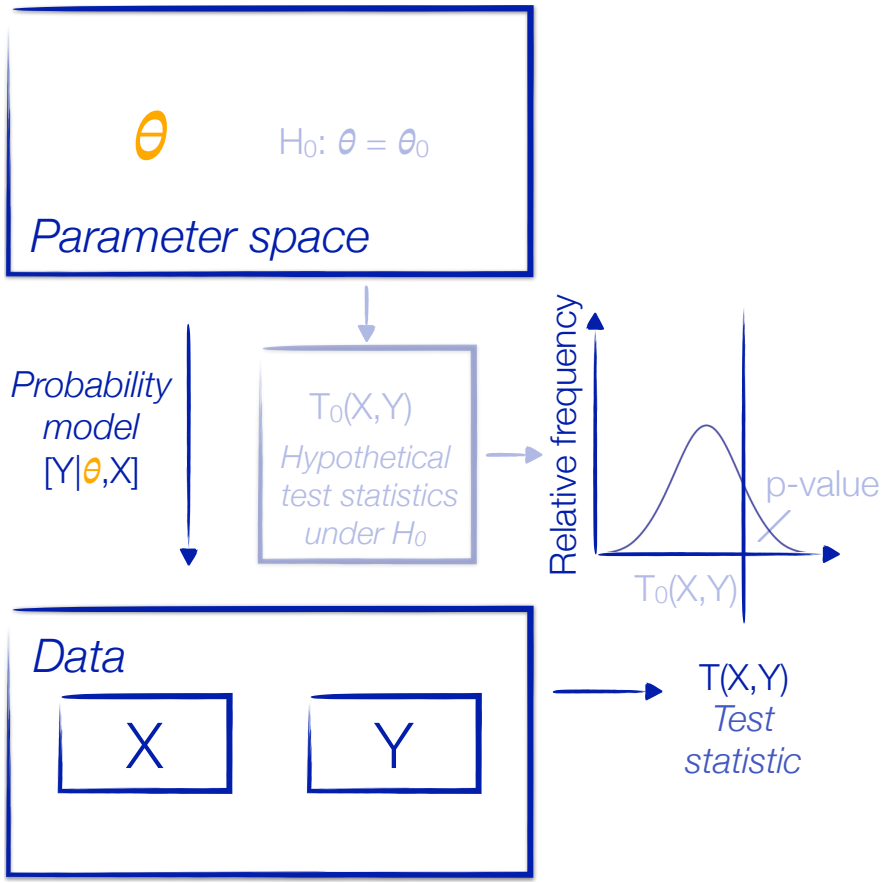


Auxiliary assumptions in statistical modelling in medicine



$\Rightarrow$ When results deviate from our expectations under H_0, should we reject H_0 or any of the auxiliary assumptions:

Auxiliary assumptions in statistical modelling in medicine



$\Rightarrow$ When results deviate from our expectations under H_0, should we reject H_0 or any of the auxiliary assumptions:

- *"The patients took the assigned treatment correctly"*
- *"All factors influencing treatment switching and the outcome were measured"*

Auxiliary assumptions play an even bigger role in the analysis of routinely collected data

- When data are prospectively collected for research purposes:
  - The research question informs the study design
  - In experimental studies, it even determines the data generating mechanism

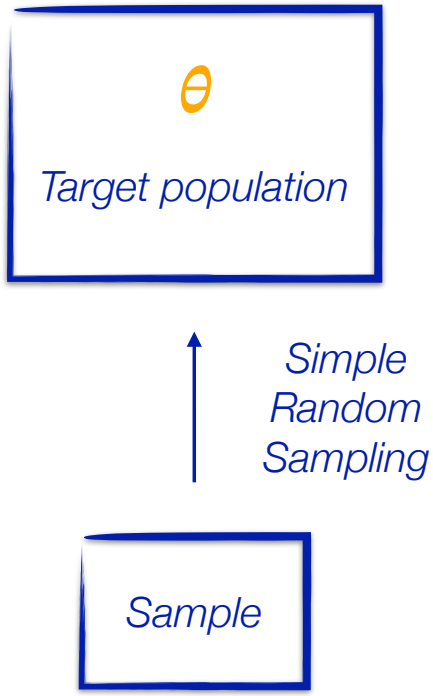
Auxiliary assumptions play an even bigger role in the analysis of routinely collected data

- When data are prospectively collected for research purposes:
  - The research question informs the study design
  - In experimental studies, it even determines the data generating mechanism
- In routinely collected data, there is
  - Little knowledge of how the data were generated
  - Little control over measurement procedures

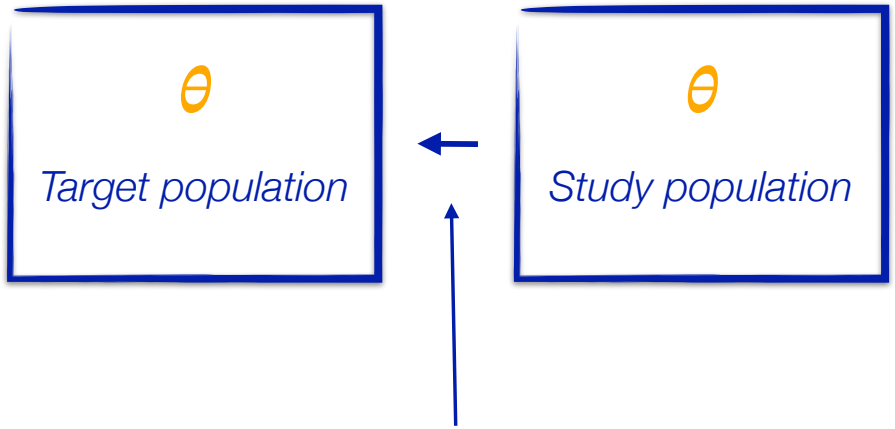
Auxiliary assumptions play an even bigger role in the analysis of routinely collected data

- When data are prospectively collected for research purposes:
  - The research question informs the study design
  - In experimental studies, it even determines the data generating mechanism
- In routinely collected data, there is
  - Little knowledge of how the data were generated
  - Little control over measurement procedures
- Consequences:
  - It is in general not possible to control who receives which treatments
  - It is unknown why certain patients received certain treatments
  - Sometimes one does not even know when and for how long patients received a certain treatment

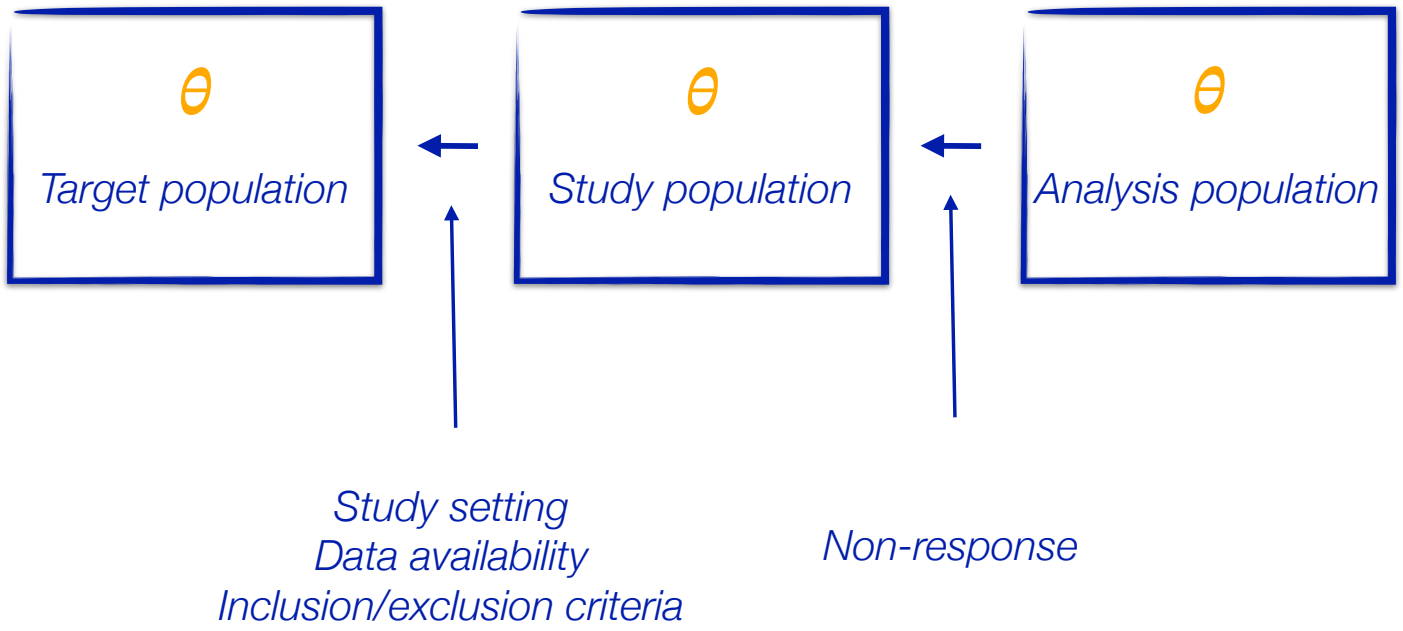
Auxiliary assumptions play an even bigger role in the analysis of routinely collected data



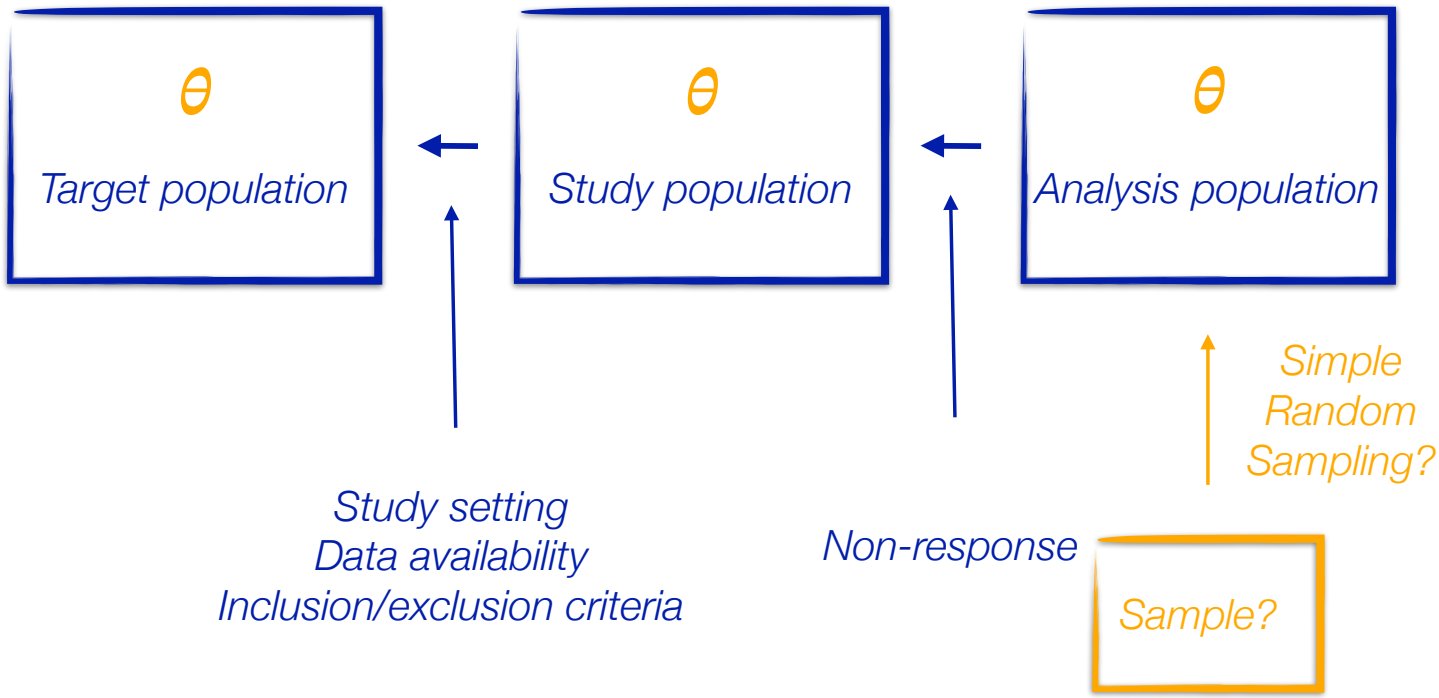
Auxiliary assumptions play an even bigger role in the analysis of routinely collected data



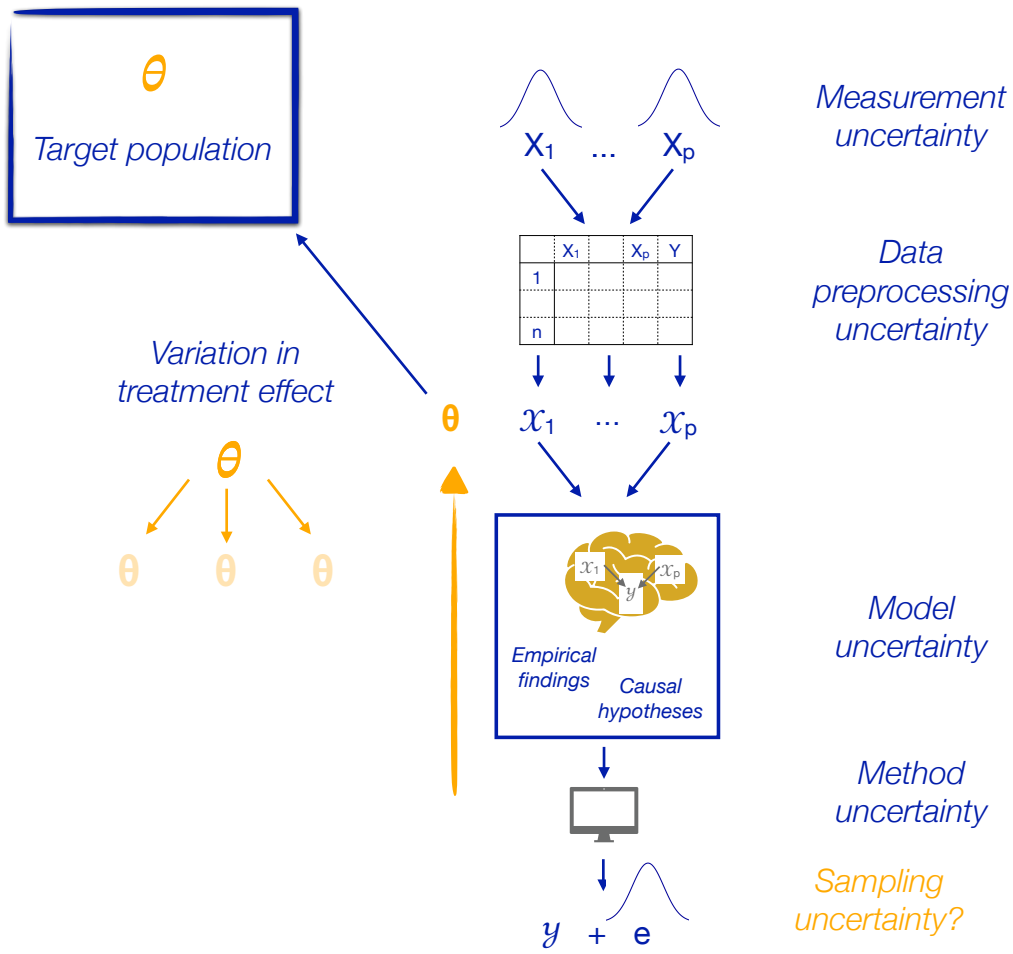
Auxiliary assumptions play an even bigger role in the analysis of routinely collected data



Auxiliary assumptions play an even bigger role in the analysis of routinely collected data

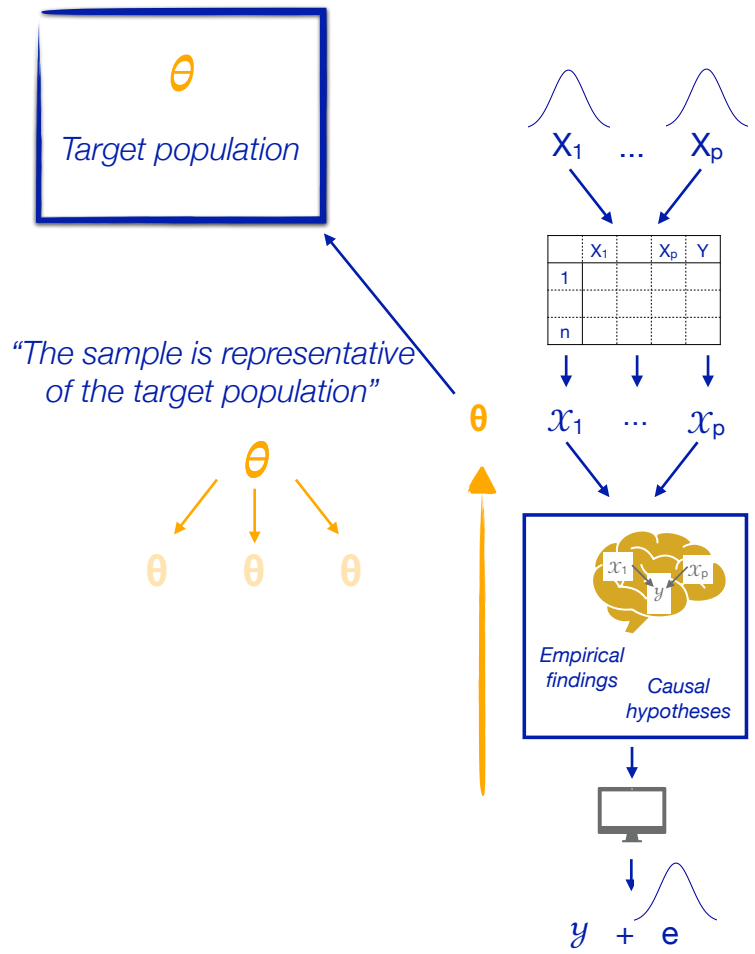


Existing and nonexistant sources of uncertainty



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

Existing and nonexistant sources of uncertainty



"There is neither measurement error nor informative missingness"

"There is a correct and unique pre-processing pipeline"

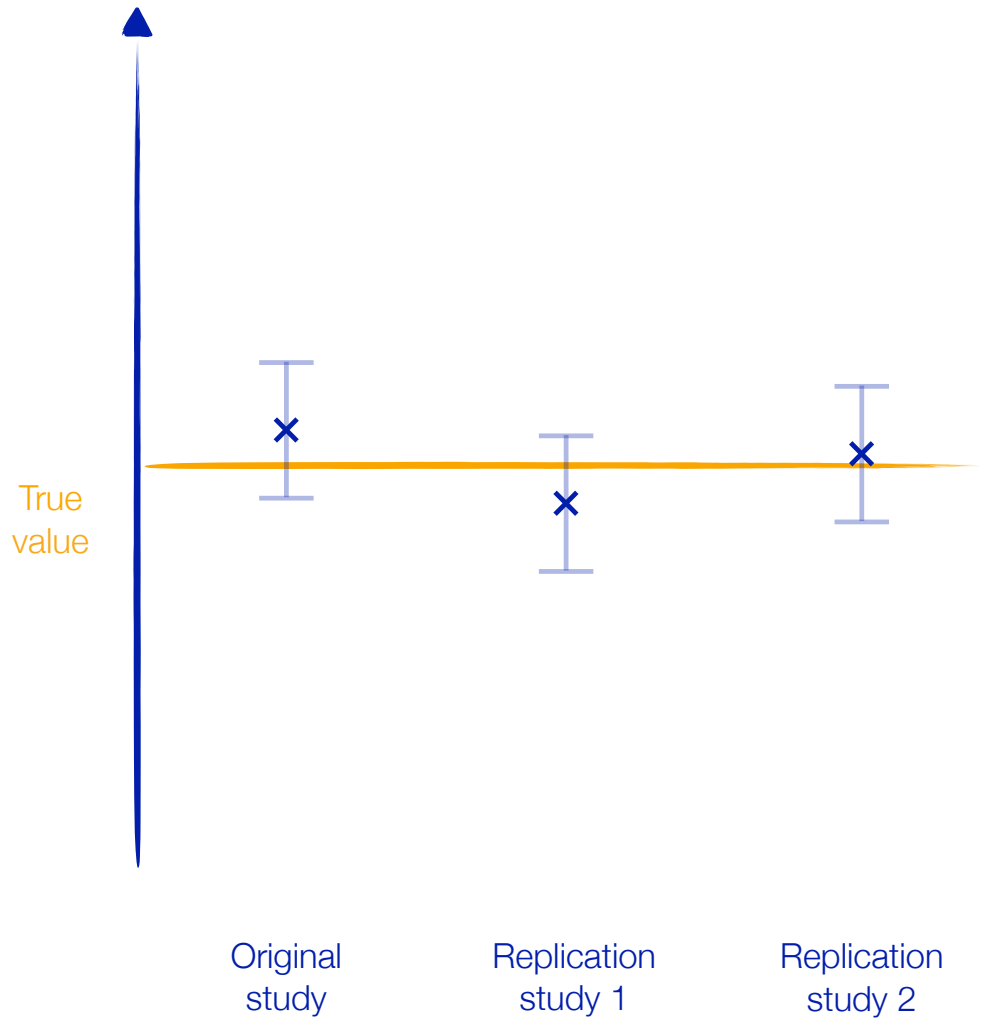
"The data generation mechanism is known"

"All confounders were measured without error"

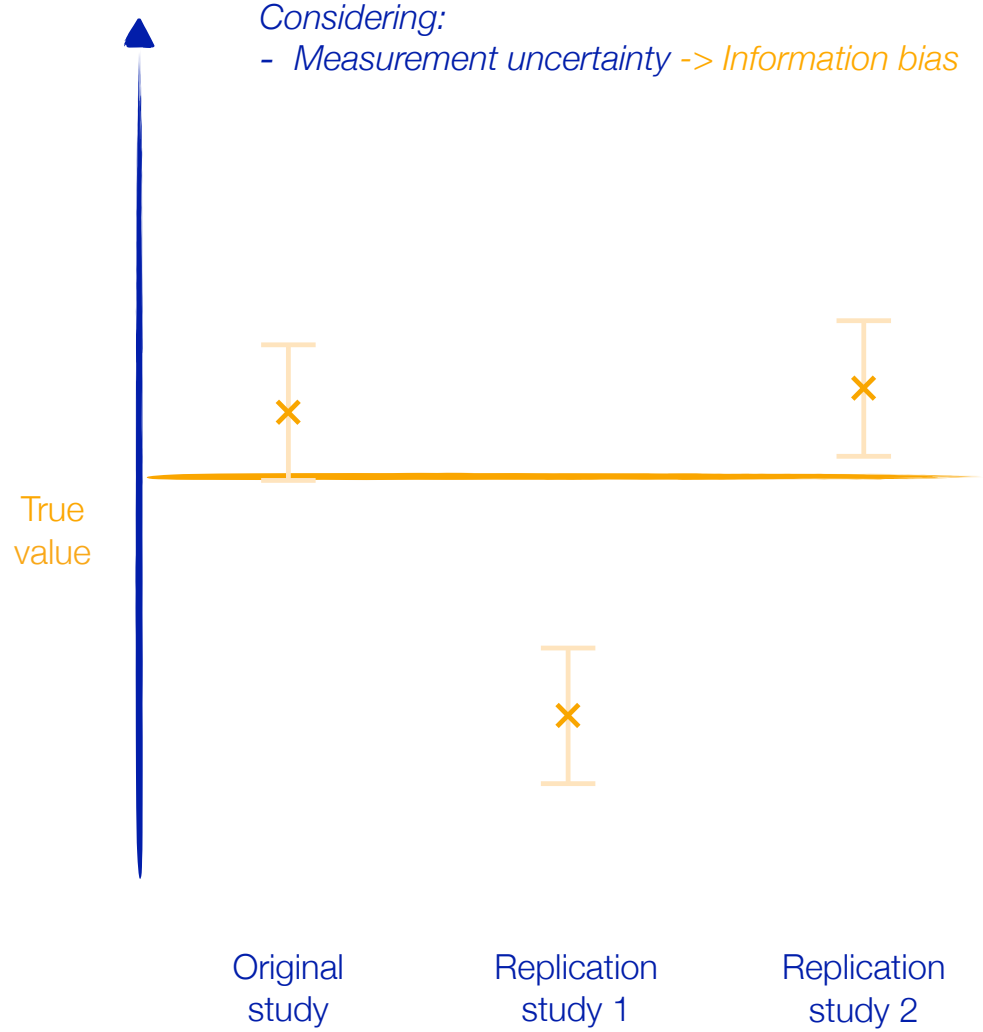
"There is a correct and unique method to analyse the data"

Random sampling from the target population

If we ignore uncertainty, it leads to bias

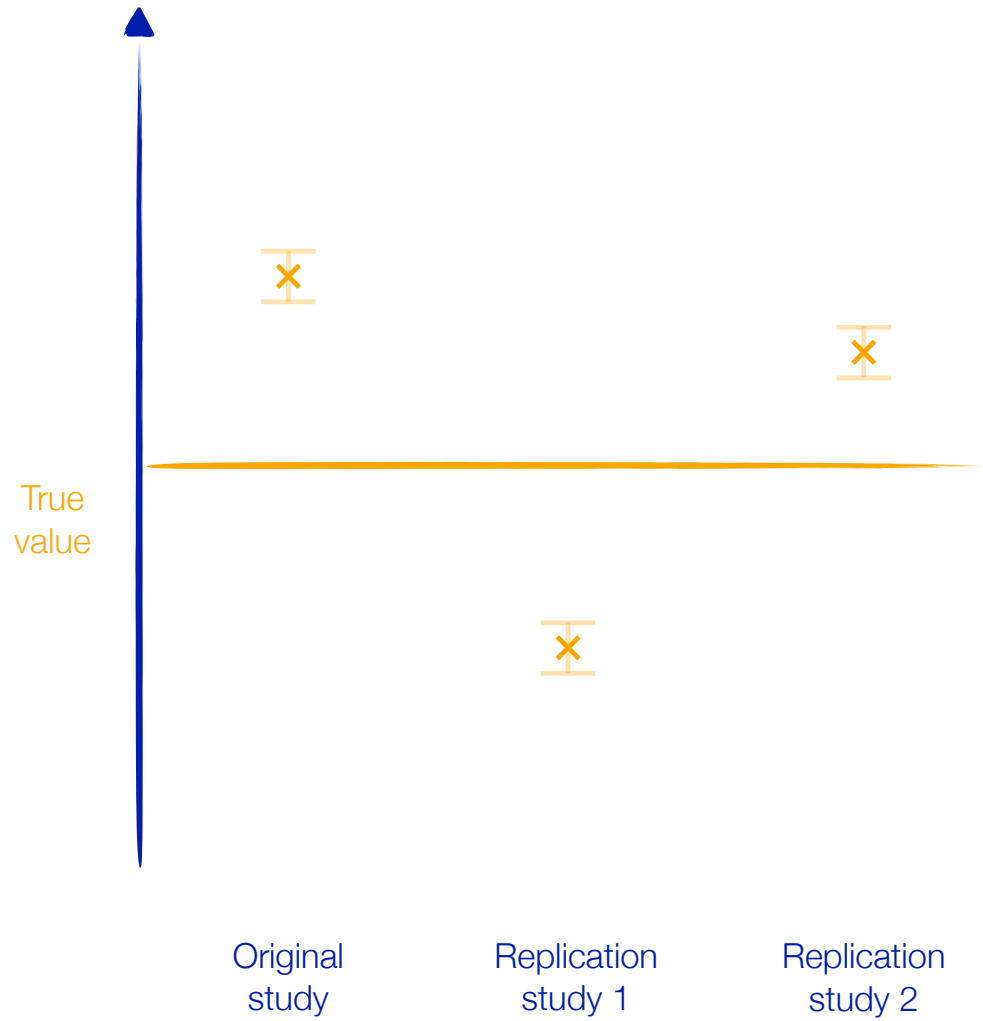


If we ignore uncertainty, it leads to bias



If we ignore uncertainty, it leads to bias

Big data paradoxes (Meng (2018))



"The more the data, the surer we fool ourselves"

Challenges in the analysis of routinely collected data

"The sample is representative of the target population"

- Due to disparities in the access and use of health-care services, routinely collected data may under- or mis-represent certain subgroups [1, 2, 3, 4, 5], including
  - ethnic minorities
  - patients without medical coverage
  - low-income and rural populations

"The sample is representative of the target population"

- Due to disparities in the access and use of health-care services, routinely collected data may under- or mis-represent certain subgroups [1, 2, 3, 4, 5], including
  - ethnic minorities
  - patients without medical coverage
  - low-income and rural populations
- If patients from under-represented groups are present in the data, there is a risk that they may be mis-represented, because they are
  - more likely to visit multiple institutions [6, 7, 8]
  - they receive fewer diagnostics tests and interventions [2]
  - they are more likely to be lost during data linkage

"The sample is representative of the target population"

- Due to disparities in the access and use of health-care services, routinely collected data may under- or mis-represent certain subgroups [1, 2, 3, 4, 5], including
  - ethnic minorities
  - patients without medical coverage
  - low-income and rural populations
- If patients from under-represented groups are present in the data, there is a risk that they may be mis-represented, because they are
  - more likely to visit multiple institutions [6, 7, 8]
  - they receive fewer diagnostics tests and interventions [2]
  - they are more likely to be lost during data linkage
- There may be substantial variation in patient characteristics and in disease occurrence across sites in multisite studies

"There is no measurement error or informative missingness"

- Routinely collected data are not recorded for a specific research purpose at hand

"There is no measurement error or informative missingness"

- Routinely collected data are not recorded for a specific research purpose at hand

$\Rightarrow$ Documentation practices may vary between different clinical settings, as a function of incentives and of the overall workload of the personnel collecting the data [9, 10]

"There is no measurement error or informative missingness"

- Routinely collected data are not recorded for a specific research purpose at hand

$\Rightarrow$ Documentation practices may vary between different clinical settings, as a function of incentives and of the overall workload of the personnel collecting the data [9, 10]

- Examples:
  - On an emergency call, vital parameters may not be measured if they are irrelevant to the clinical question at hand
  - During a busy period, nurses may not find the time to record clinical events or changes in medication, or they may only find time at the end of their shift
  - Prescription orders may not be filled or consumed by the patient
  - Temporal changes in the recording of data may produce systematic differences over time

"There is no measurement error or informative missingness"

- Routinely collected data are not recorded for a specific research purpose at hand

$\Rightarrow$ Documentation practices may vary between different clinical settings, as a function of incentives and of the overall workload of the personnel collecting the data [9, 10]

- Examples:
  - On an emergency call, vital parameters may not be measured if they are irrelevant to the clinical question at hand
  - During a busy period, nurses may not find the time to record clinical events or changes in medication, or they may only find time at the end of their shift
  - Prescription orders may not be filled or consumed by the patient
  - Temporal changes in the recording of data may produce systematic differences over time

$\Rightarrow$ Parameter estimates can be biased in any direction

"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"

- In routinely collected data, the first entry for a patient typically has no clear clinical meaning and it is often unknown which diagnoses or treatments the patient received before this time point

"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"

- In routinely collected data, the first entry for a patient typically has no clear clinical meaning and it is often unknown which diagnoses or treatments the patient received before this time point
- In a representative sample of reports of comparative non-randomised studies that assessed the effectiveness and/or safety of drug treatments, Yaacoub et al. (2024) [11] found that in 72% of studies eligibility, treatment assignment, and start of follow-up were not aligned.

"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"

- In routinely collected data, the first entry for a patient typically has no clear clinical meaning and it is often unknown which diagnoses or treatments the patient received before this time point
- In a representative sample of reports of comparative non-randomised studies that assessed the effectiveness and/or safety of drug treatments, Yaacoub et al. (2024) [11] found that in 72% of studies eligibility, treatment assignment, and start of follow-up were not aligned.
- In a data audit on the quality of observational study data in an international HIV research network, treatment regimens and associated dates and the timings of laboratory measurements were especially prone to error with error rates of up to 56% and 42% [12]

"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"

Intensive Care Med
https://doi.org/10.1007/s00134-025-07805-4

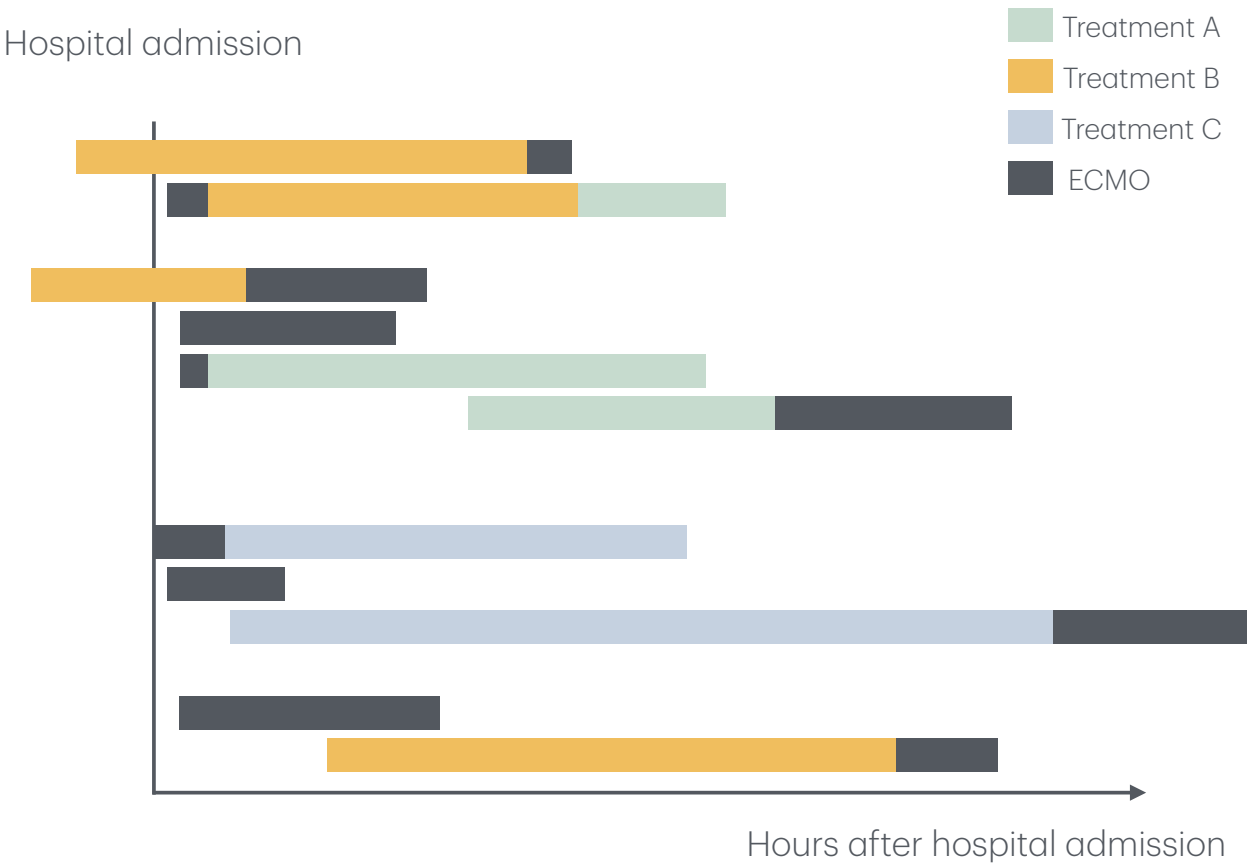
ORIGINAL

Management of high-risk acute pulmonary embolism: an emulated target trial analysis

Andrea Stadlbauer[1], Tom Verbelen[2], Leonhard Binzenhöfer[3], Tomaz Goslar[4], Alexander Supady[5], Peter M. Spieth[6], Marko Noc[4], Andreas Verstraete[2], Sabine Hoffmann[7], Michael Schomaker[41], Julia Höpler[7], Marie Kraft[7], Esther Tautz[5], Daniel Hoyer[8], Jörn Tongers[8], Franz Haertel[9], Aschraf El-Essawi[10], Mostafa Salem[11], Rafael Henrique Rangel[11], Carsten Hullermann[12], Marvin Kriz[13], Benedikt Schrage[13], Jorge Moisés[14], Manel Sabate[14], Federico Pappalardo[15], Lisa Crusius[16], Norman Mangner[16], Christoph Adler[17], Tobias Tichelbäcker[17], Carsten Skurk[18], Christian Jung[19], Sebastian Kufner[20], Tobias Graf[21], Clemens Scherer[3], Laura Villegas Sierra[3], Hannah Billig[22], Nicolas Majunke[23], Walter S. Speidl[24], Robert Zilberszac[24], Luis Chiscano-Camón[25], Aitor Uribarri[26], Jordi Riera[25], Roberto Roncon-Albuquerque Jr[27], Elizabete Terauda[28], Andrejs Erglis[28], Guido Tavazzi[29], Uwe Zeymer[30], Maike Knorr[31], Juliane Kilo[32], Sven Möbius-Winkler[9], Robert H. G. Schwinger[33], Derk Frank[11], Oliver Borst[34], Helene Häberle[35], Frederic De Roeck[36], Christiaan Vrints[36], Christof Schmid[1], Georg Nickenig[22], Christian Hagl[37], Steffen Massberg[3], Andreas Schäfer[38], Dirk Westermann[39], Sebastian Zimmer[22], Alain Combes[40], Daniele Camboni[1*], Holger Thiele[23] and Enzo Lüsebrink[22*] for the High-risk P. E. Investigator Group

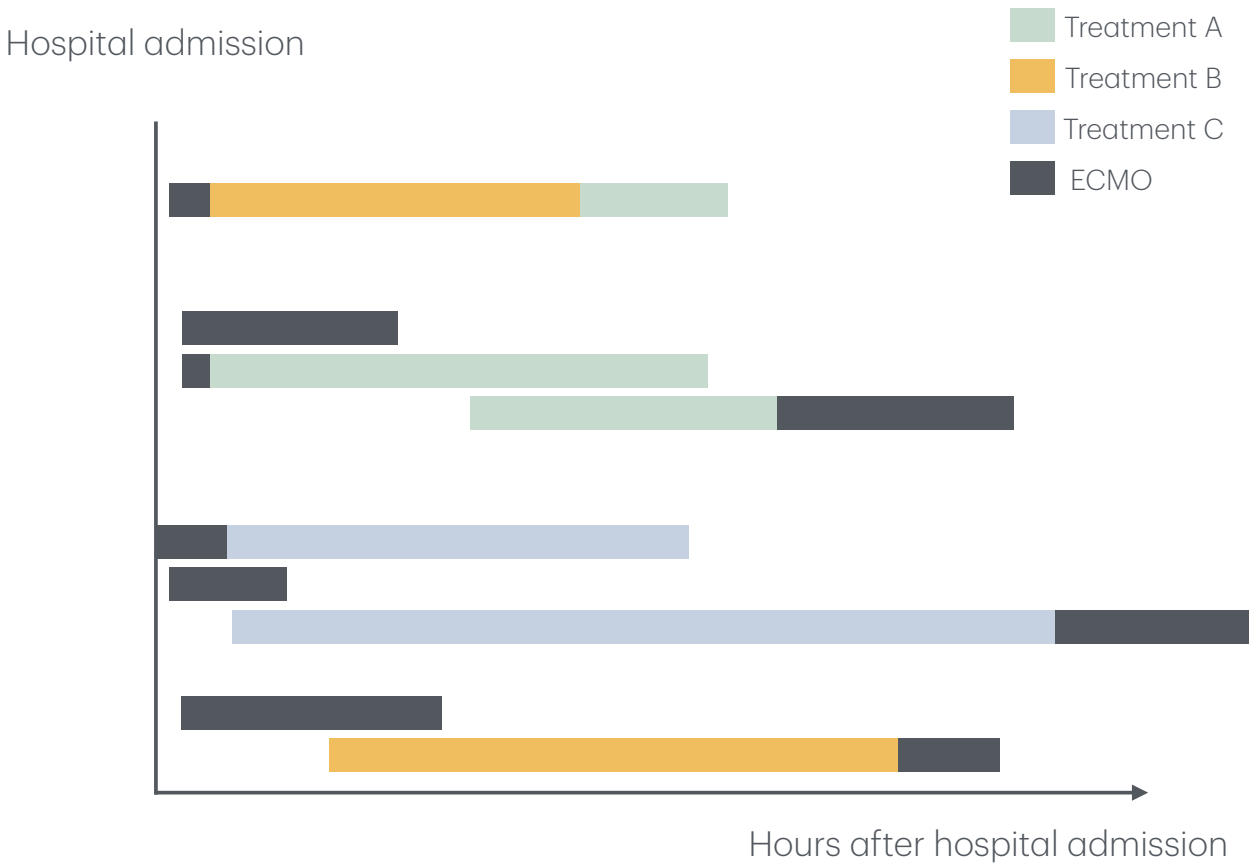
ID	A	B	C	ECMO	Mortality
1	0	1	0	1	1
2	1	1	0	1	0
3	1	0	0	0	1
4	0	1	0	1	1
5	0	0	0	1	1
6	1	0	0	1	0
7	1	0	0	1	1
8	0	0	1	0	1
9	0	1	0	0	0
10	0	0	1	1	0
11	0	0	0	1	1
12	0	0	1	1	1
13	0	0	1	0	0
14	0	0	0	1	1
15	0	1	0	1	1

"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"



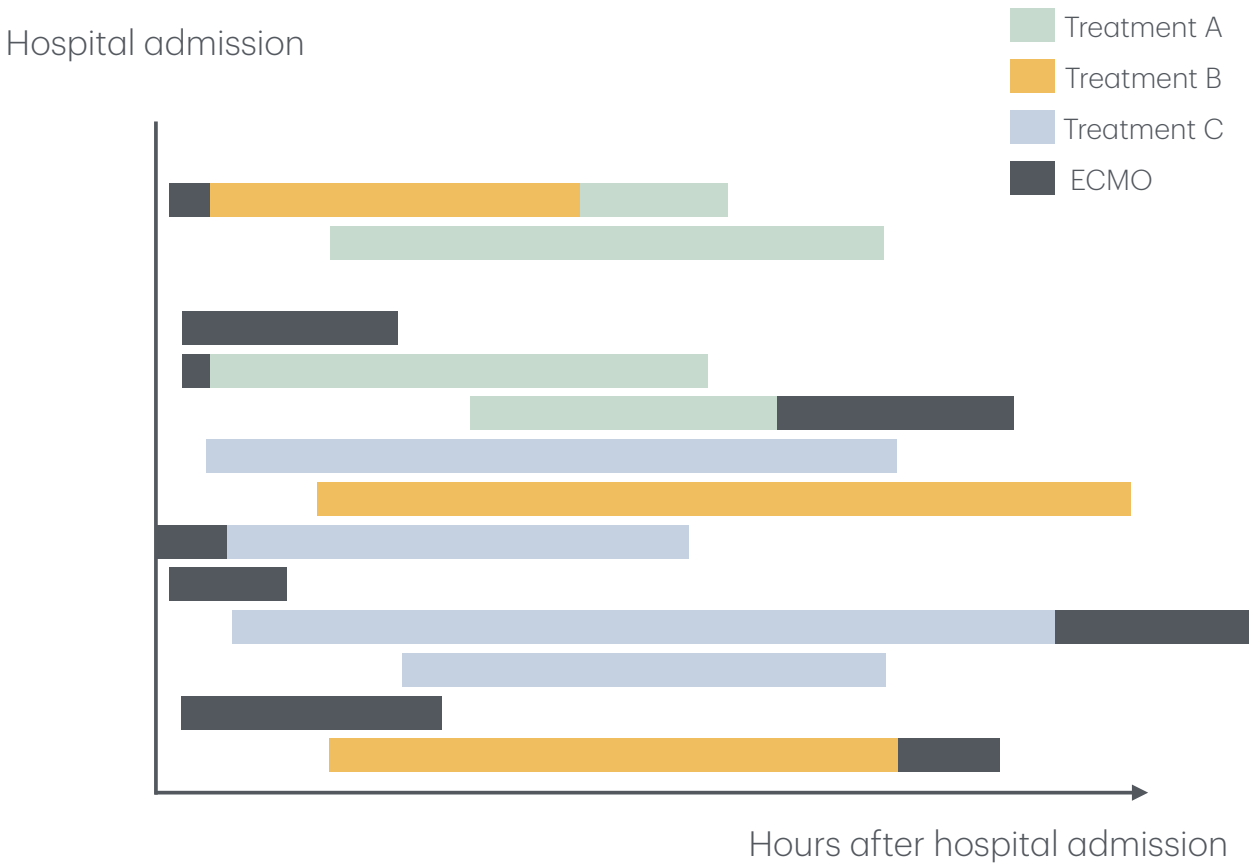
ID	A	B	C	ECMO	Mortality
1	0	1	0	1	1
2	1	1	0	1	0
3	1	0	0	0	1
4	0	1	0	1	1
5	0	0	0	1	1
6	1	0	0	1	0
7	1	0	0	1	1
8	0	0	1	0	1
9	0	1	0	0	0
10	0	0	1	1	0
11	0	0	0	1	1
12	0	0	1	1	1
13	0	0	1	0	0
14	0	0	0	1	1
15	0	1	0	1	1

"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"



ID	A	B	C	ECMO	Mortality
1	0	1	0	1	1
2	1	1	0	1	0
3	1	0	0	0	1
4	0	1	0	1	1
5	0	0	0	1	1
6	1	0	0	1	0
7	1	0	0	1	1
8	0	0	1	0	1
9	0	1	0	0	0
10	0	0	1	1	0
11	0	0	0	1	1
12	0	0	1	1	1
13	0	0	1	0	0
14	0	0	0	1	1
15	0	1	0	1	1

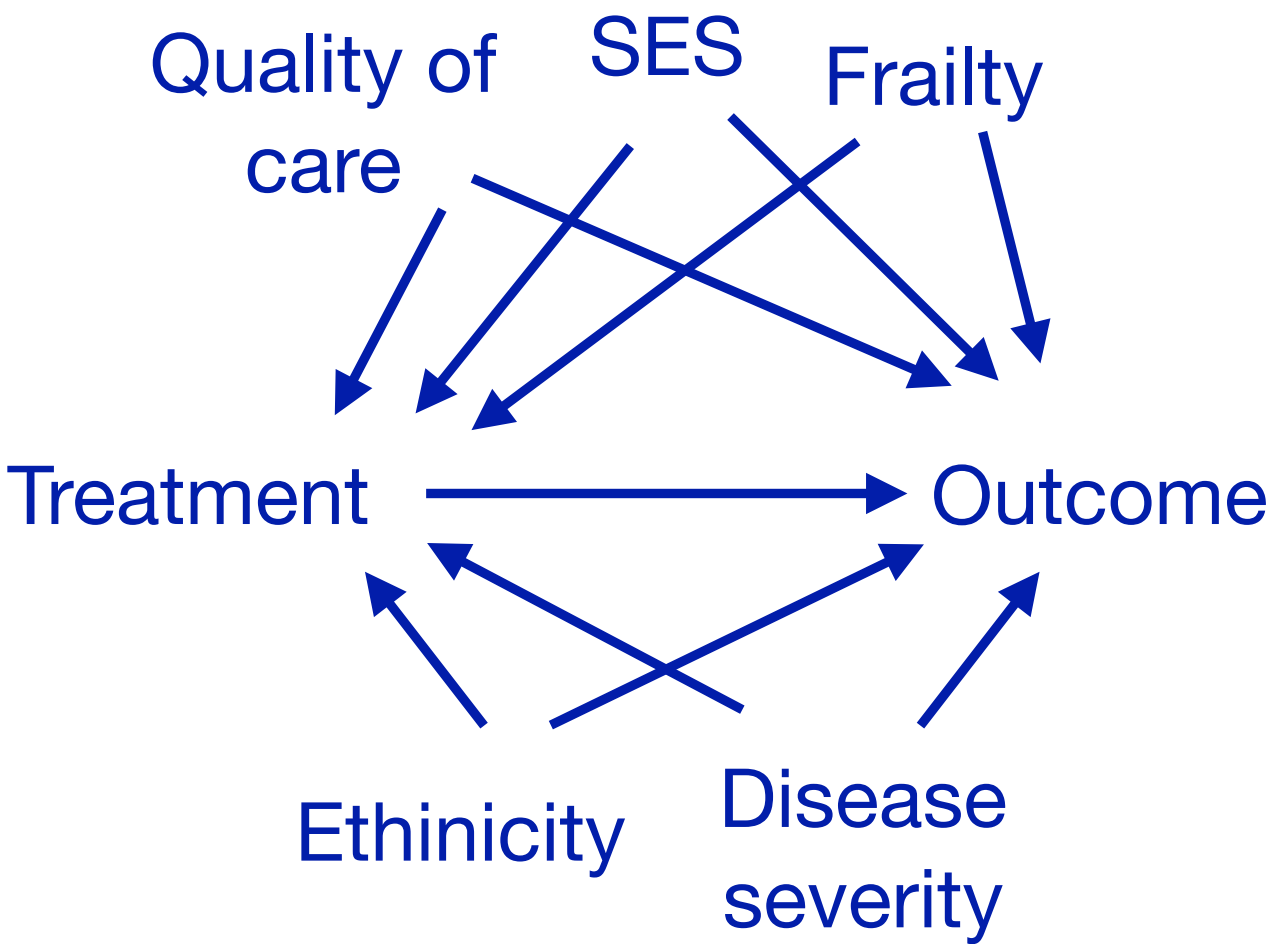
"It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"

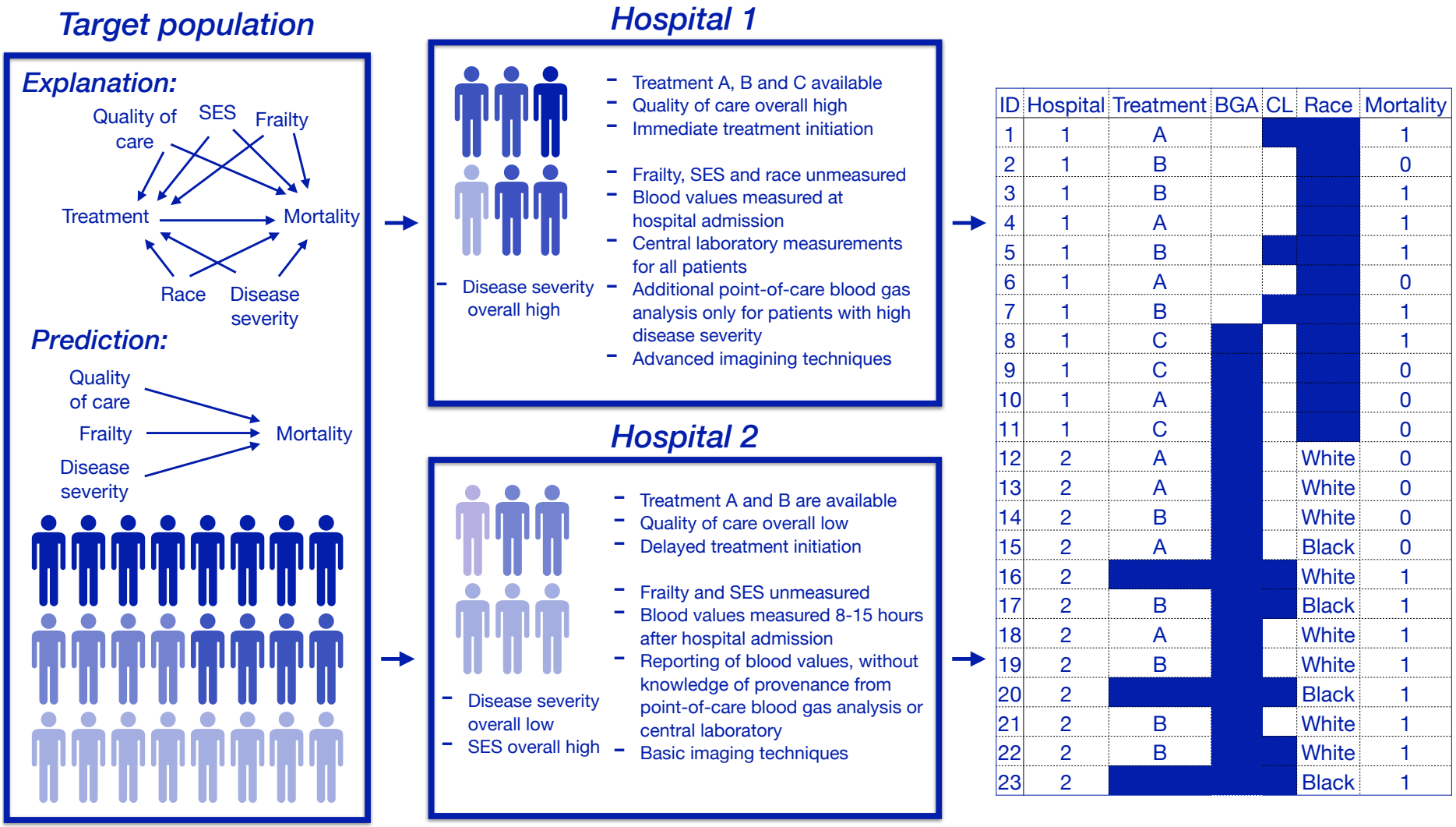


ID	A	B	C	ECMO	Mortality
1	0	1	0	1	1
2	1	1	0	1	0
3	1	0	0	0	1
4	0	1	0	1	1
5	0	0	0	1	1
6	1	0	0	1	0
7	1	0	0	1	1
8	0	0	1	0	1
9	0	1	0	0	0
10	0	0	1	1	0
11	0	0	0	1	1
12	0	0	1	1	1
13	0	0	1	0	0
14	0	0	0	1	1
15	0	1	0	1	1

"All relevant confounders were measured without error"

"All relevant confounders were measured without error"





ID	Hospital	Treatment	BGA	CL	Race	Mortality
1	1	A				1
2	1	B				0
3	1	B				1
4	1	A				1
5	1	B				1
6	1	A				0
7	1	B				1
8	1	C				1
9	1	C				0
10	1	A				0
11	1	C				0
12	2	A			White	0
13	2	A			White	0
14	2	B			White	0
15	2	A			Black	0
16	2				White	1
17	2	B			Black	1
18	2	A			White	1
19	2	B			White	1
20	2				Black	1
21	2	B			White	1
22	2	B			White	1
23	2				Black	1

Hoffmann, S., Morris, T., Herrmann, M., Heinze, G., Wynants, L., Van Calster, B., Bischl, B., Schmid, M., Shaw, P.A., Mathes, T., Naudet, F., Harrell, F.E. (...), Thiele, H., Lüsebrink, E. (2026). Using routinely collected data for research purposes: Challenges and mitigation strategies. The BMJ *In print*

Relying on auxiliary assumptions leads to a multiplicity of possible analysis strategies

- "There is no measurement error or informative missingness"
- "It is possible to find an alignment between eligibility, treatment assignment and start of follow-up"
- "All relevant confounders were measured without error"

Relying on auxiliary assumptions leads to a multiplicity of possible analysis strategies

Underdetermination of scientific theory by evidence:

It is not possible to unambiguously falsify a theory, because theories are always tested in a bundle with various auxiliary assumptions.

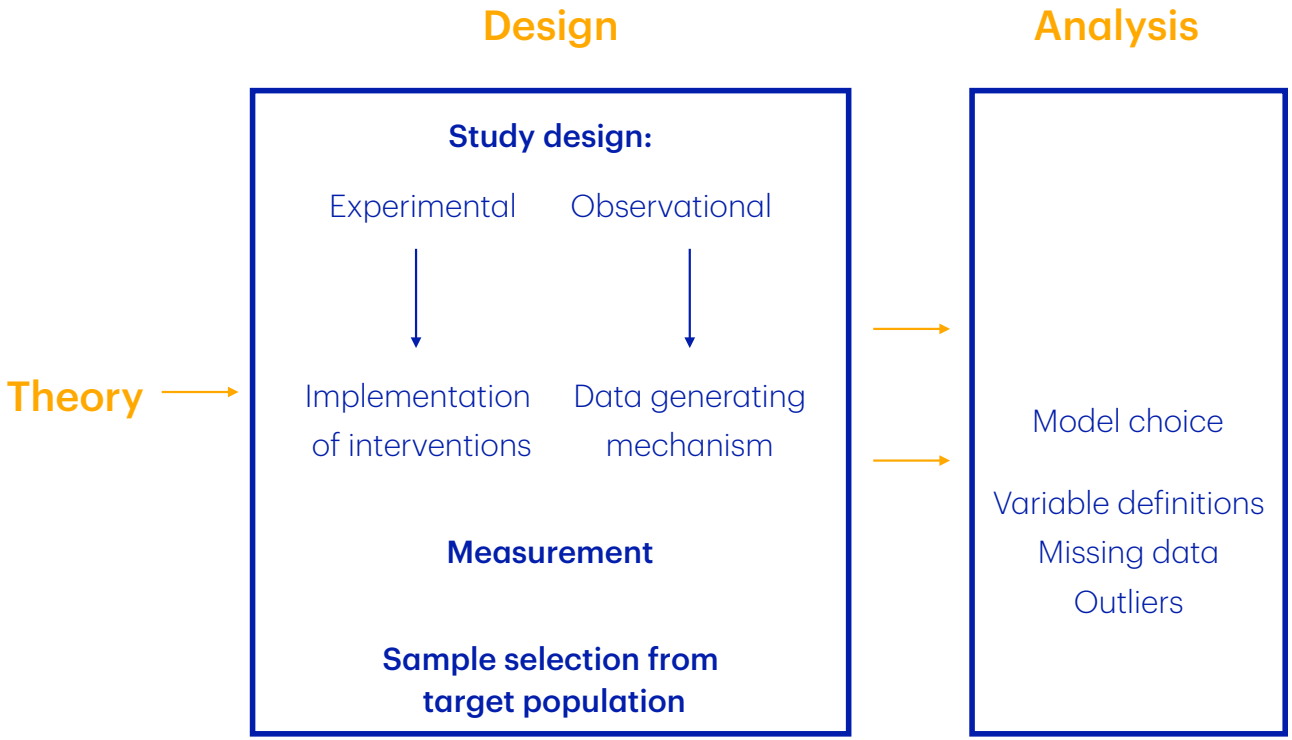
When a prediction fails, we never know whether we should blame the theory or one of the auxiliary assumptions.

(Duhem-Quine problem)

Relying on auxiliary assumptions leads to a multiplicity of possible analysis strategies

Underdetermination of scientific theory by evidence

Type 2: Underdetermination of evidence by scientific theory



Landy et al. (2020, Psychological Bulletin)
Almaatouq et al. (2022, Behavioural and Brain Sciences)
Huber et al. (2023, PNAS)

Silberzahn et al. (2015, Nature)
Botvinik-Nezer et al. (2020, Nature)
Aczel et al. (2026, Nature)

Relying on auxiliary assumptions leads to a multiplicity of possible analysis strategies

nature

Explore content | About the journal | Publish with us

nature > articles > article

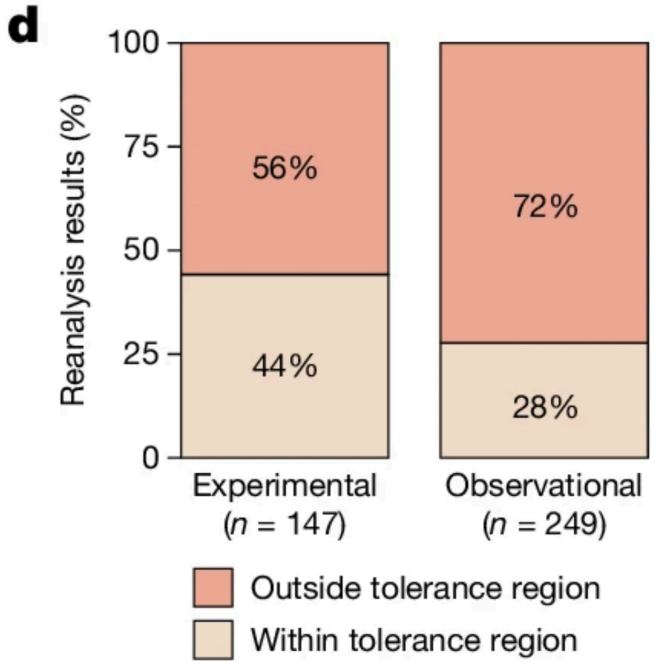
Article | Published: 01 April 2026

Investigating the analytical robustness of the social and behavioural sciences

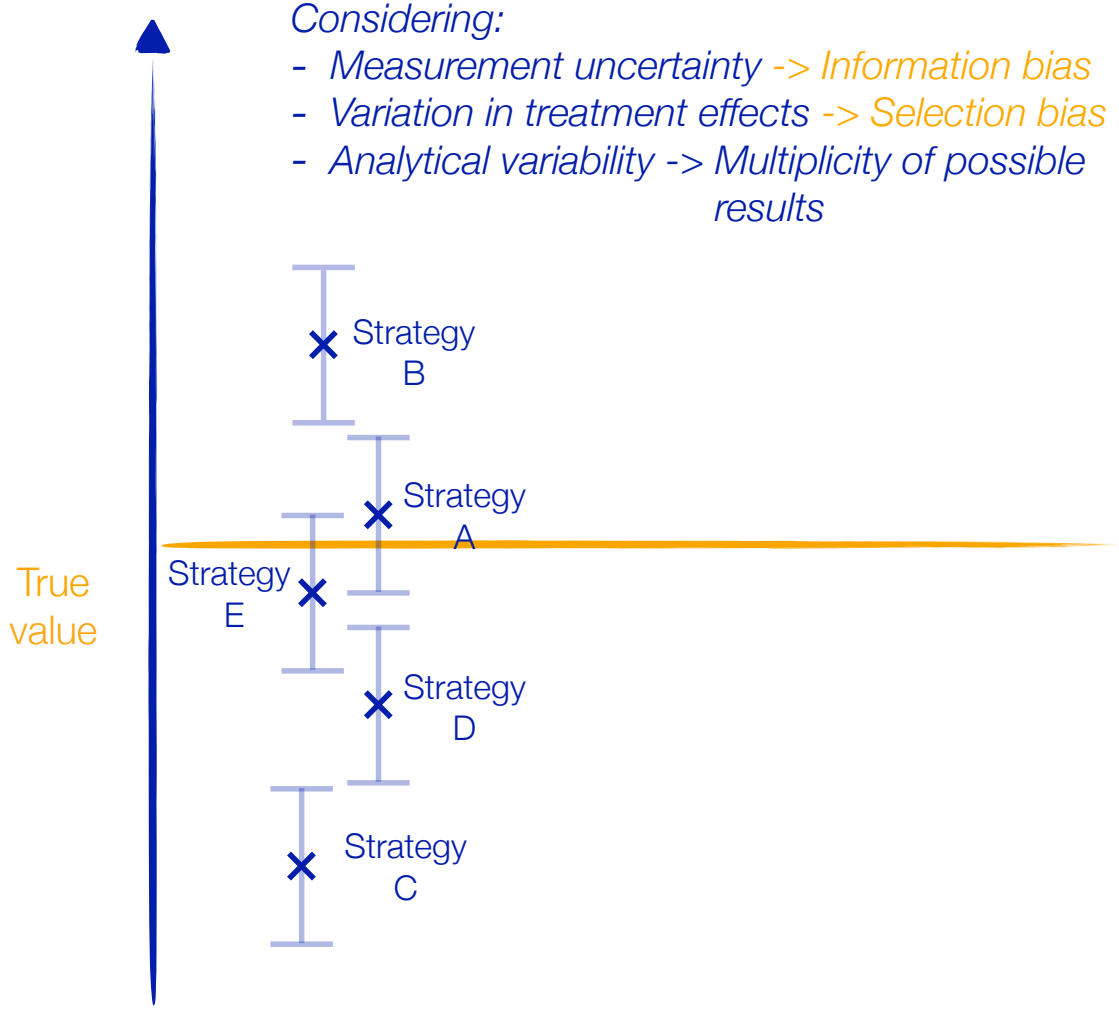
Balazs Aczel, Barnabas Szaszi, Harry T. Clelland, Marton Kovacs, Felix Holzmeister, Don van Ravenzwaaij, Hannah Schulz-Kümpel, Sabine Hoffmann, Gustav Nilsonne, Livia Kosa, Zoltan A. Torma, Yousuf Abdelfatah, Christopher L. Aberson, Oguz A. Acar, Ensar Acem, Matus Adamkovic, Timofey Adamovich, Krisna Adiasto, Love Ahnström, Atakan M. Akil, Adil S. Al-Busaidi, Ali H. Al-Hoorie, Casper J. Albers, Peter J. Allen, ... Brian A. Nosek + Show authors

Nature **652**, 135–142 (2026) | Cite this article

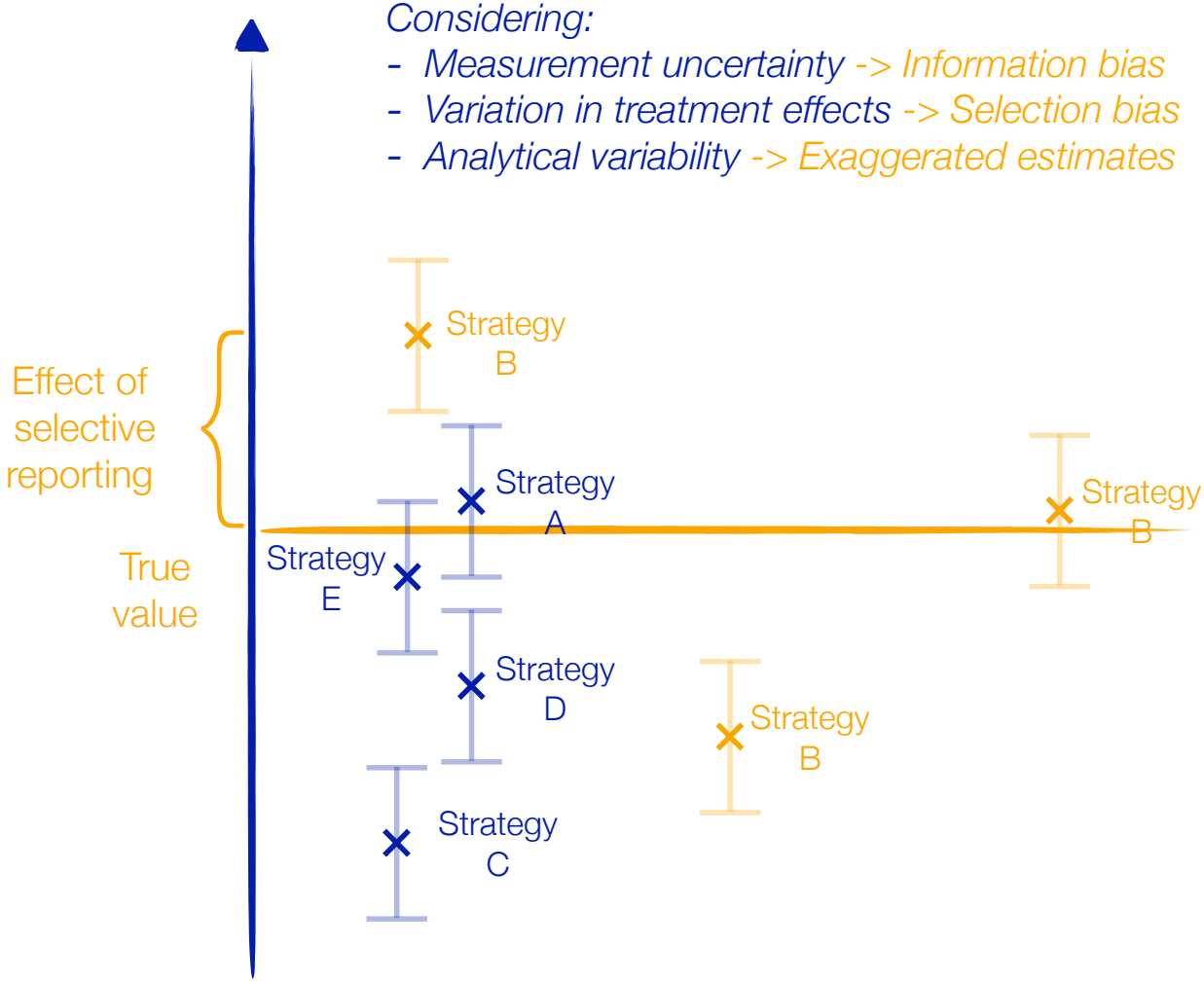
17k Accesses | **5** Citations | **614** Altmetric | Metrics



Selective reporting among a multiplicity of possible analysis strategies leads to non-replicability and overconfidence



Selective reporting among a multiplicity of possible analysis strategies leads to non-replicability and overconfidence



Selective reporting among a multiplicity of possible analysis strategies leads to conflicting findings [13, 14, 15]

Surgery 165 (2019) 953–957

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

Colon/Rectum

Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection?

Adam C. Fields, MD[a,*], Pamela Lu, MD[a,b], Deanna L. Palenzuela, BS[a], Ronald Bleday, MD[a], Joel E. Goldberg, MD, MPH[a], Jennifer Irani, MD[a], Jennifer S. Davids, MD[c], Nelya Melnitchouk, MD, MSc[a,b,*]

[a] Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA
[b] Center for Surgery and Public Health, Department of Surgery, Brigha
[c] Department of Surgery, University of Massachusetts Memorial Medic

Surgery 165 (2019) 1199–1202

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

Presented at the Academic Surgical Congress 2019

Utilization of a specimen retrieval bag during laparoscopic appendectomy for both uncomplicated and complicated appendicitis is not associated with a decrease in postoperative surgical site infection rates

Scott A. Turner, MD*, Hee Soo Jung, MD, FACS, John E. Scarborough, MD, FACS

adison, WI

Table. Comparison of 2 Studies on the Association of a Specimen Retrieval Bag With Surgical Site Infection Rates in Laparoscopic Appendectomy

	Source	
Criteria	Fields et al,[3] 2019	Turner et al,[4] 2019
Inclusion criteria	*CPT* code 44970, not missing information on intra-abdominal abscess	Laparoscopic appendectomy, pathology with acute appendicitis, no additional major procedure
Analytic sample reported, No.	11 475	10 357
Primary outcome	Postoperative intra-abdominal abscess	Any SSI (superficial, deep, organ space)
Primary predictor	Use of retrieval bag	Use of retrieval bag
Covariates included, with operationalization	Age (continuous)	Age (dichotomized at 65 y)
	Sex (dichotomized)	Sex (dichotomized)
	BMI (continuous)	Obesity (categorical: not obese, class I/II/III obesity, missing)
	Race (categorized as: White, Black, Asian, other)	Not included
	Diabetes (dichotomized)	Diabetes (dichotomized)
	Hypertension (dichotomized)	Not included
	COPD (dichotomized)	Not included
	Smoker (dichotomized)	Not included
	Functional status (dichotomized)	Not included
	Steroid use (dichotomized)	Steroid use (dichotomized)
	Weight loss (dichotomized)	Not included
	Preoperative sepsis (dichotomized)	Unclear if included
	Wound class 3/4 (dichotomized)	Not included
	Complicated appendicitis (dichotomized)	2 Indicator variables: presence of abscess and presence of perforation
	ASA class 3/4 (dichotomized)	Not included
	Operative time (continuous)	Operative time dichotomized at 75th percentile
	White blood cell count (continuous)	Not included
Coefficient on primary predictor	OR (95% CI): 0.6 (0.42-0.95) *P* value: .03	OR (95% CI): 1.15 (0.78-1.69) *P* value: .49

Mitigation strategies to address challenges in the analysis of routinely collected data

Transparency is both more important and more difficult in the analysis of routinely collected data

ANALYSIS

1 CHU Rennes, Inserm, Institut de Recherche en Santé, Environnement et Travail-UMR_S 1085, University of Rennes, Rennes, France

2 Institut Universitaire de France, Paris, France

3 Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

4 Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

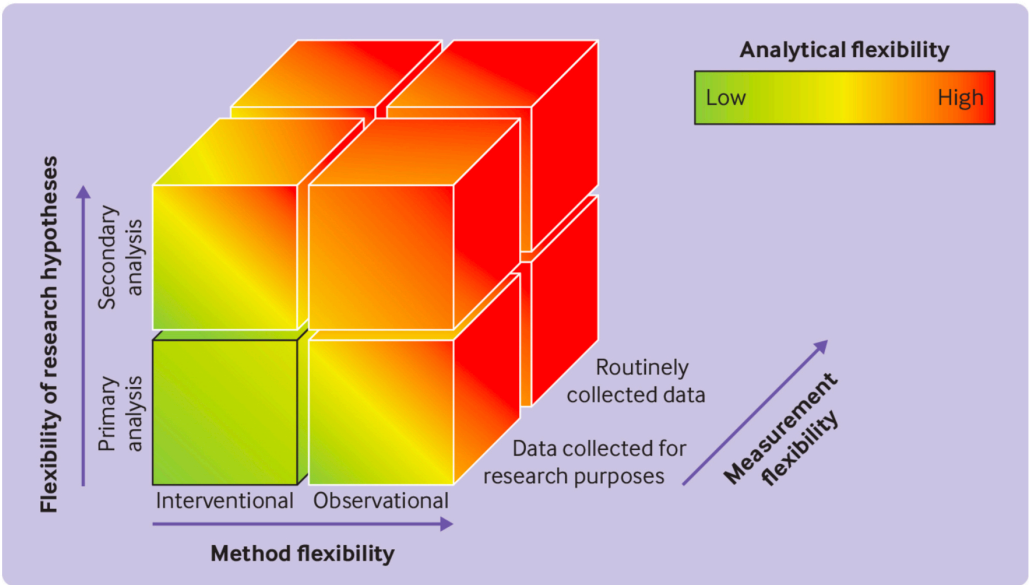
Improving the transparency and reliability of observational studies through registration

Florian Naudet and colleagues argue that routine registration of observational research is needed and suggest how current processes can be adapted to facilitate it

Florian Naudet, [1,2] Chirag J Patel, [3] Nicholas J DeVito, [4] Gérard Le Goff, [5] Ioana A Cristea, [6] Alain Braillon, [7] Sabine Hoffmann [8,9]

From the use of booster doses against covid-19[1] to approval of novel cancer therapies [2] health

This already happens for certain studies, such as non-interventional post-authorisation safety studies

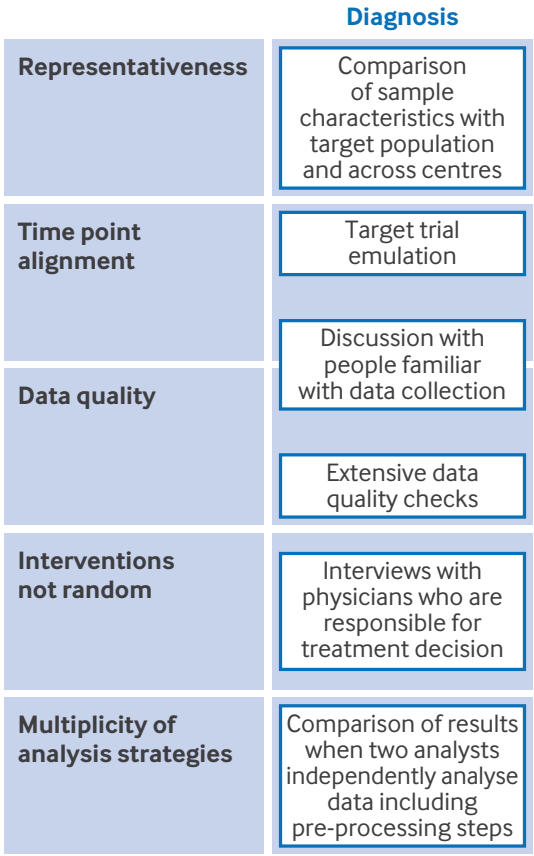


BMJ: first published as 10.1136/bmj-2023-076123

Mitigation strategies

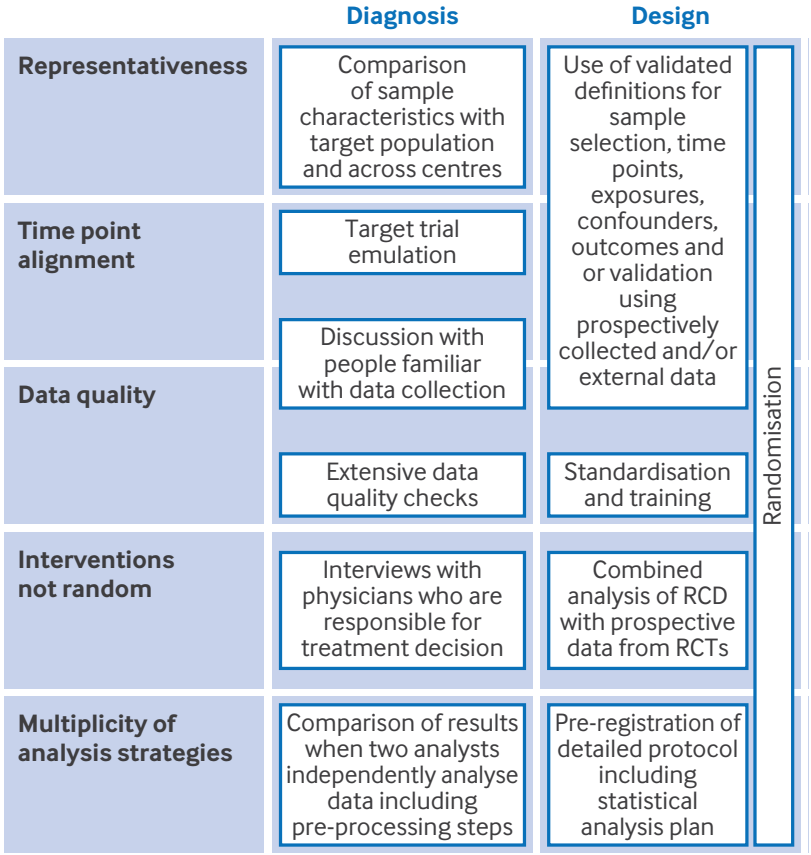
Hoffmann, S., Morris, T., Herrmann, M., Heinze, G., Wynants, L., Van Calster, B., Bischl, B., Schmid, M., Shaw, P.A., Mathes, T., Naudet, F., Harrell, F.E. (...), Thiele, H., Lüsebrink, E. (2026). Using routinely collected data for research purposes: Challenges and mitigation strategies. The BMJ *In print*

Mitigation strategies



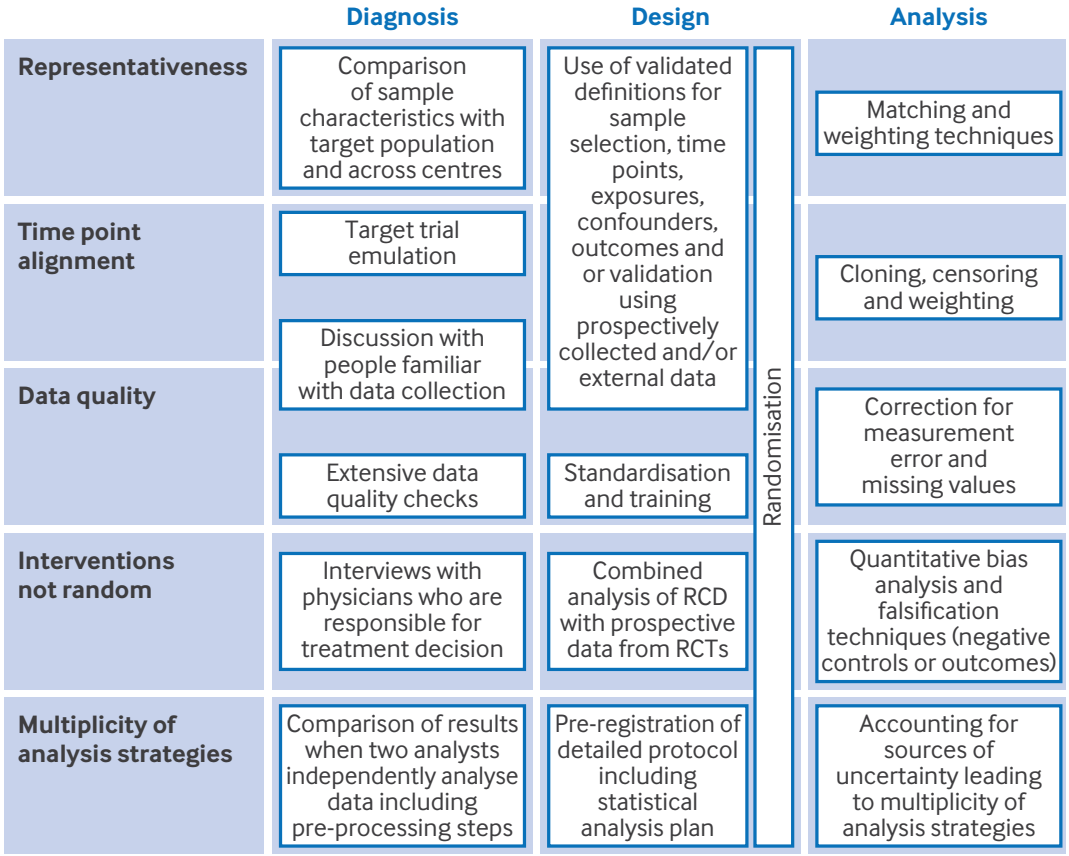
Hoffmann, S., Morris, T., Herrmann, M., Heinze, G., Wynants, L., Van Calster, B., Bischl, B., Schmid, M., Shaw, P.A., Mathes, T., Naudet, F., Harrell, F.E. (...), Thiele, H., Lüsebrink, E. (2026). Using routinely collected data for research purposes: Challenges and mitigation strategies. The BMJ *In print*

Mitigation strategies



Hoffmann, S., Morris, T., Herrmann, M., Heinze, G., Wynants, L., Van Calster, B., Bischl, B., Schmid, M., Shaw, P.A., Mathes, T., Naudet, F., Harrell, F.E. (...), Thiele, H., Lüsebrink, E. (2026). Using routinely collected data for research purposes: Challenges and mitigation strategies. The BMJ *In print*

Mitigation strategies



Hoffmann, S., Morris, T., Herrmann, M., Heinze, G., Wynants, L., Van Calster, B., Bischl, B., Schmid, M., Shaw, P.A., Mathes, T., Naudet, F., Harrell, F.E. (...), Thiele, H., Lüsebrink, E. (2026). Using routinely collected data for research purposes: Challenges and mitigation strategies. The BMJ *In print*

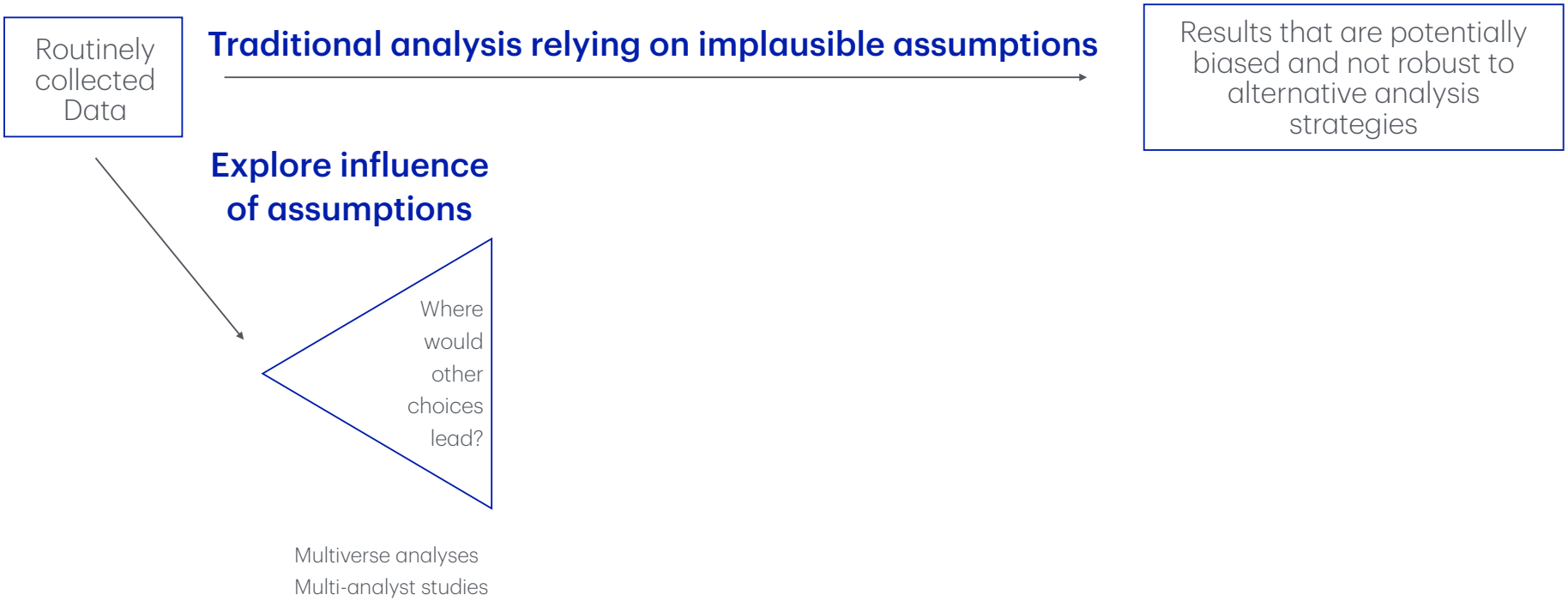
Steps to produce reliable results on routinely collected data

Routinely collected Data

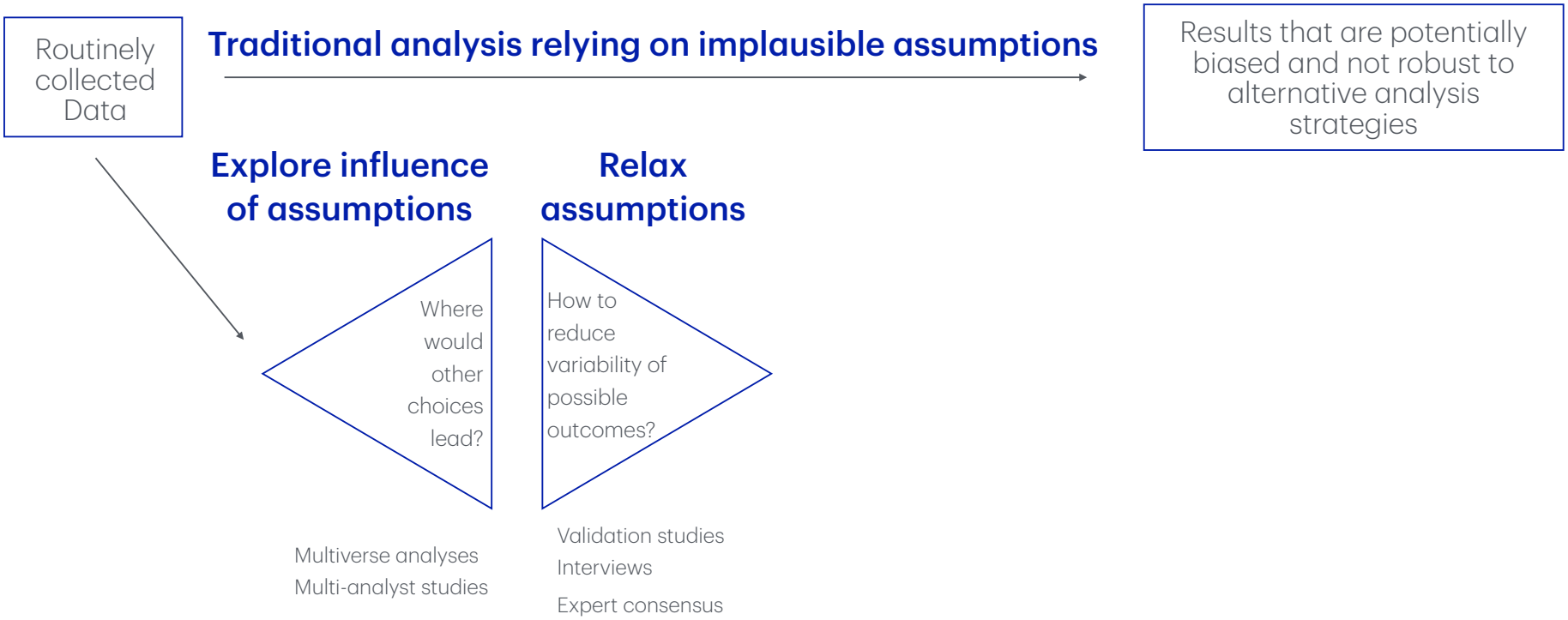
Traditional analysis relying on implausible assumptions

Results that are potentially biased and not robust to alternative analysis strategies

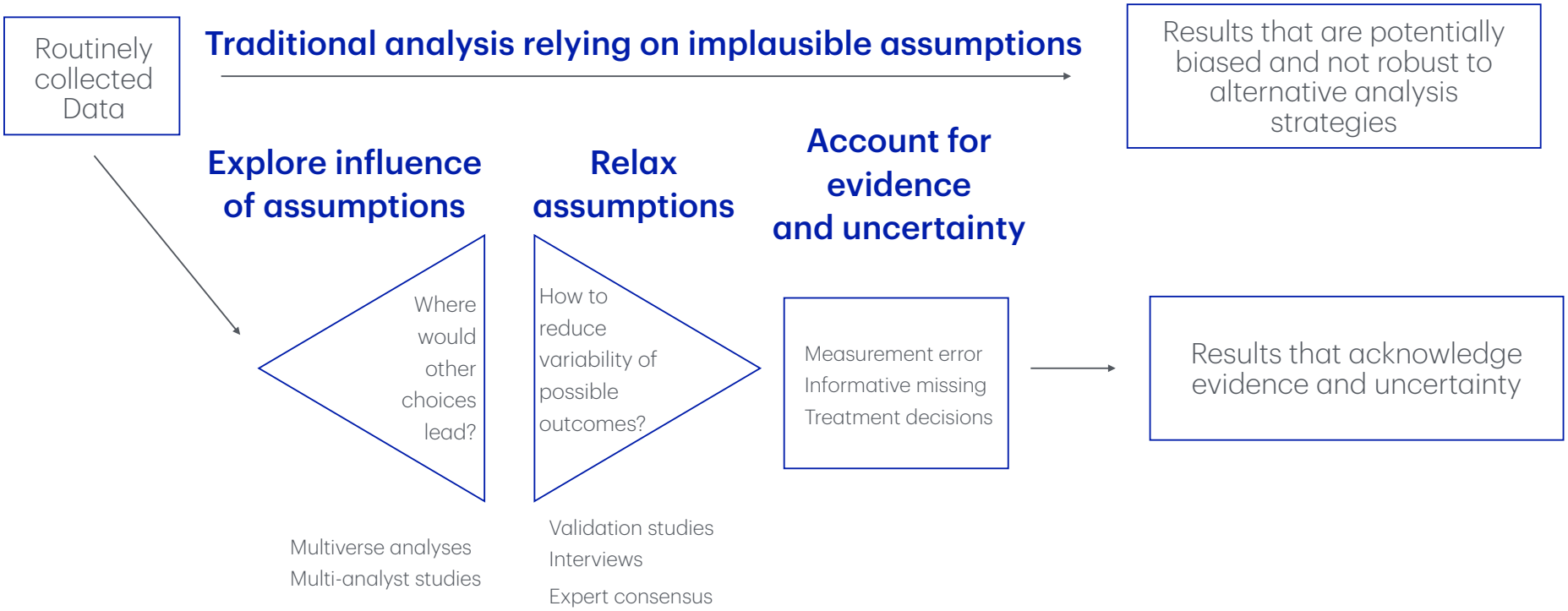
Steps to produce reliable results on routinely collected data



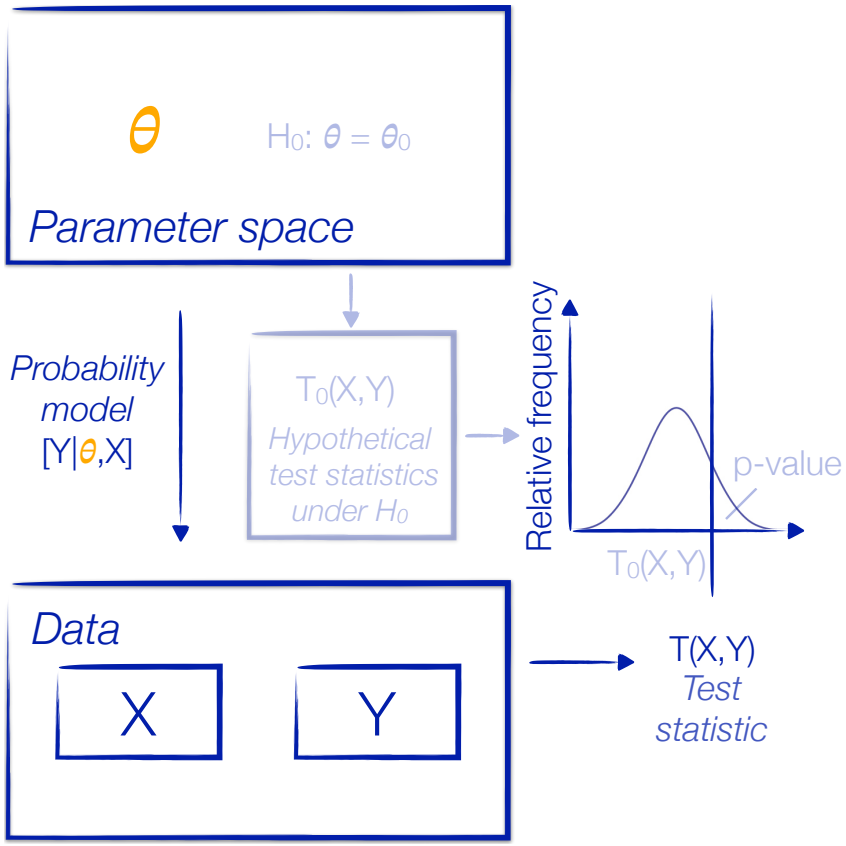
Steps to produce reliable results on routinely collected data



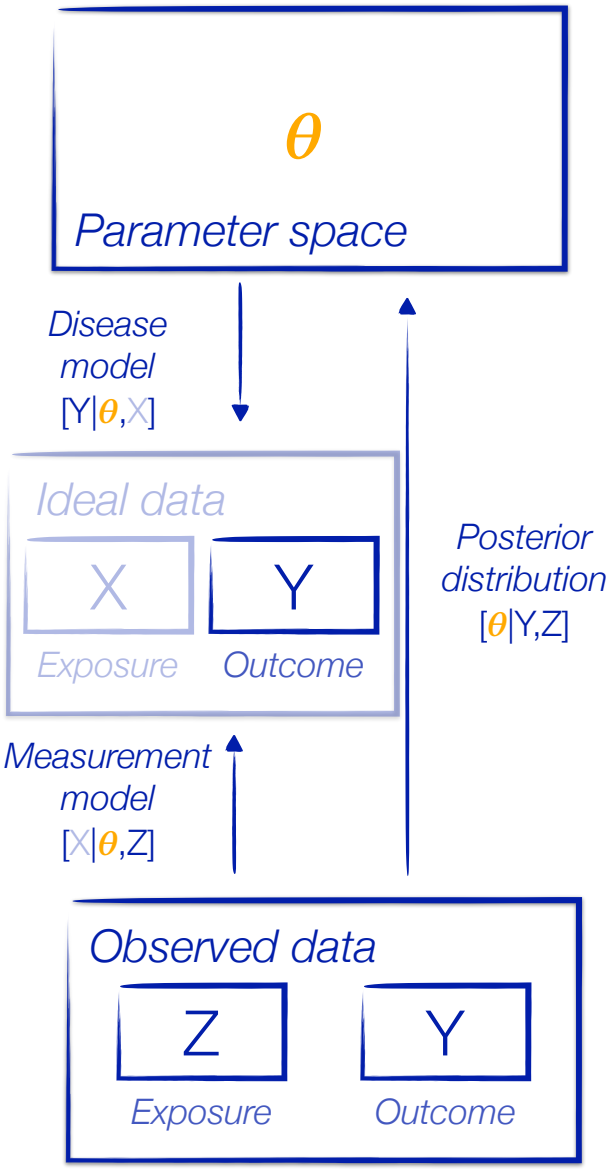
Steps to produce reliable results on routinely collected data



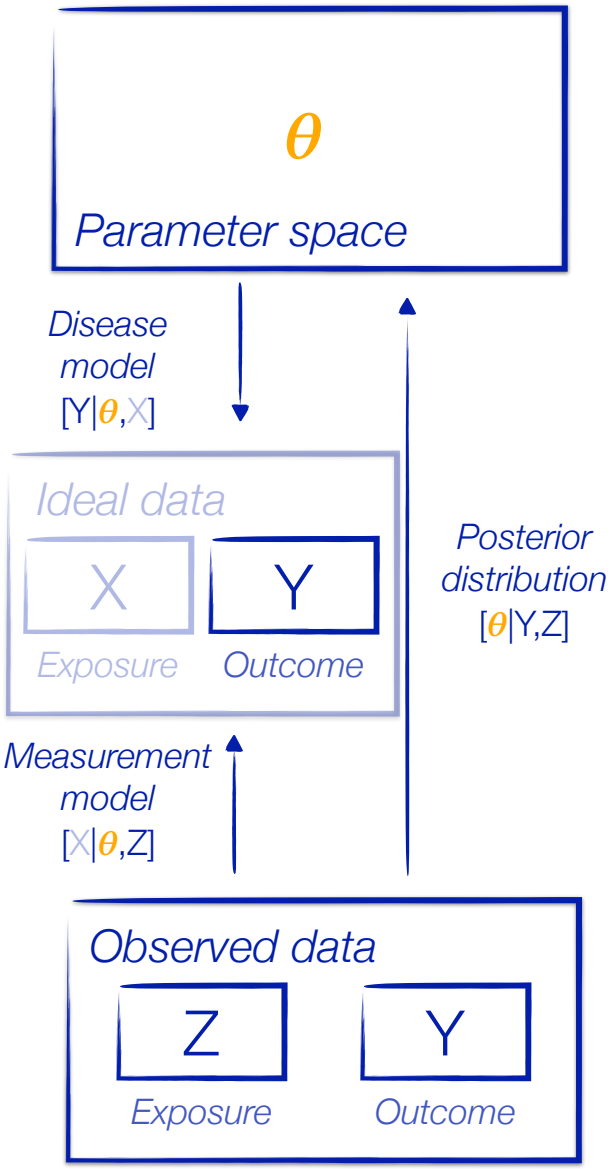
Accounting for evidence and uncertainty



Accounting for evidence and uncertainty

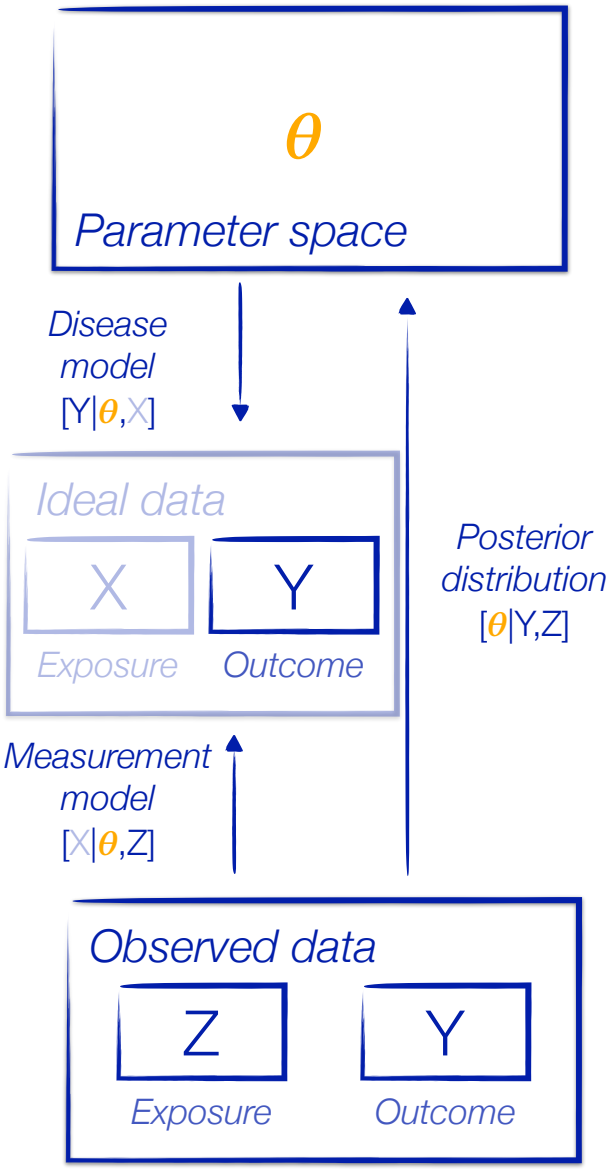


Accounting for evidence and uncertainty



Suggestions on how to relax auxiliary assumptions:

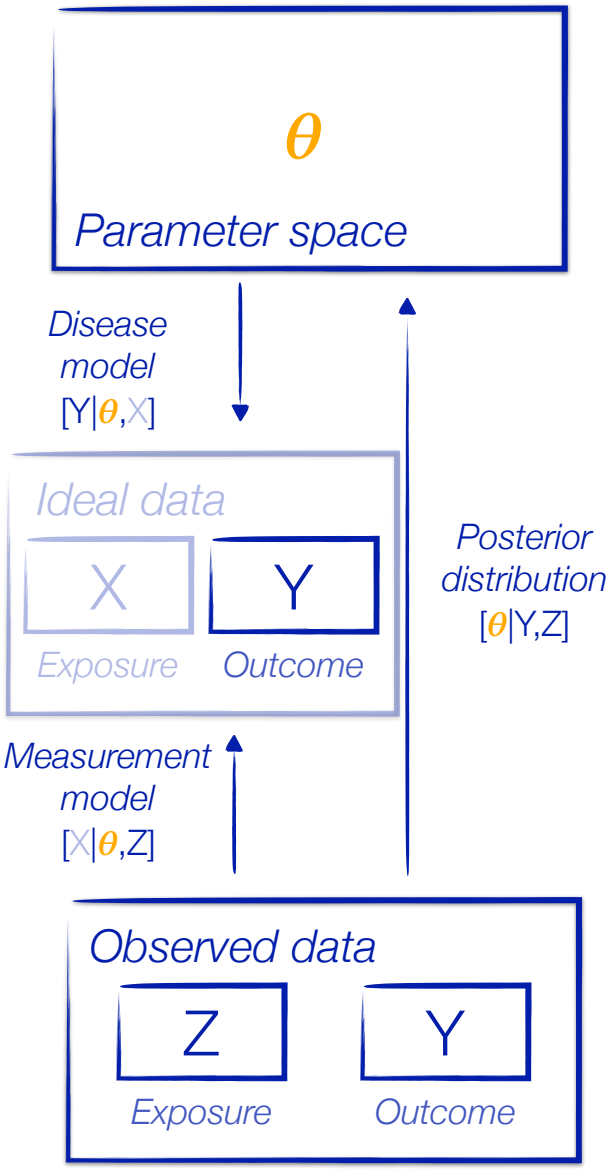
Accounting for evidence and uncertainty



Suggestions on how to relax auxiliary assumptions:

- Think (very hard) about the data generating mechanism

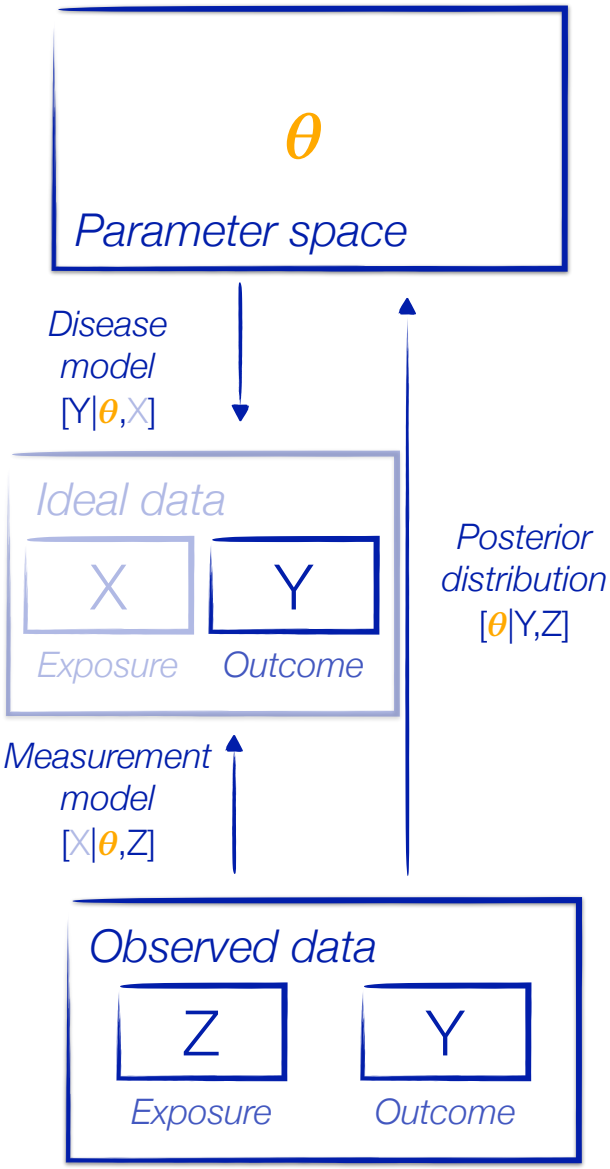
Accounting for evidence and uncertainty



Suggestions on how to relax auxiliary assumptions:

- Think (very hard) about the data generating mechanism
- Conduct validation studies and interviews with medical personnel to relax auxiliary assumptions

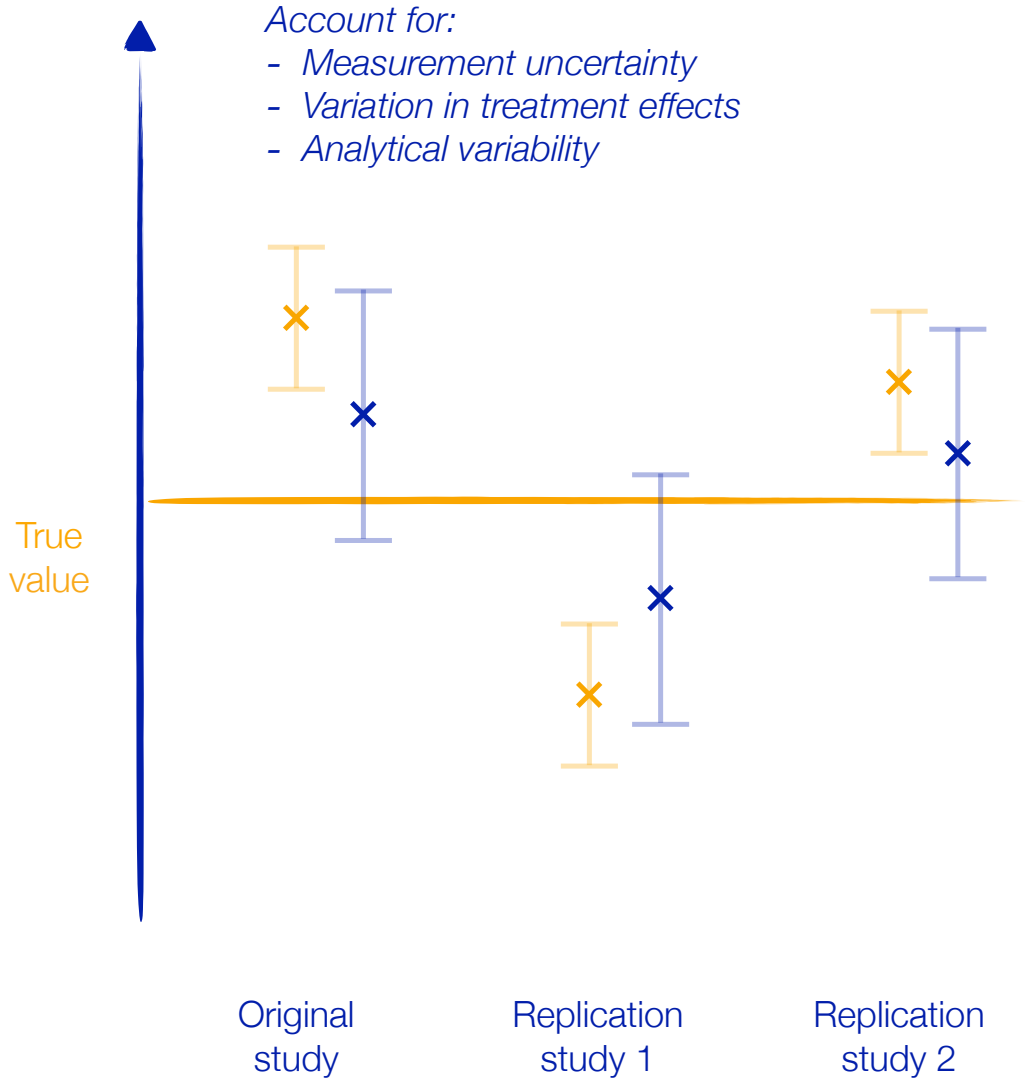
Accounting for evidence and uncertainty



Suggestions on how to relax auxiliary assumptions:

- Think (very hard) about the data generating mechanism
- Conduct validation studies and interviews with medical personnel to relax auxiliary assumptions
- Account for all sources of evidence and all sources of uncertainty

Relax auxiliary assumptions to reduce overconfidence and improve replicability



Conclusion

Conclusion

- Auxiliary assumptions play a very important role in the analysis of routinely collected data because researchers usually lack knowledge of the data generating mechanism and they have no control over measurement procedures

Conclusion

- Auxiliary assumptions play a very important role in the analysis of routinely collected data because researchers usually lack knowledge of the data generating mechanism and they have no control over measurement procedures
- To produce reliable and credible results, it is important to verify and to relax auxiliary assumptions and to account for all evidence and uncertainty in the analysis of routinely collected data

Conclusion

- Auxiliary assumptions play a very important role in the analysis of routinely collected data because researchers usually lack knowledge of the data generating mechanism and they have no control over measurement procedures
- To produce reliable and credible results, it is important to verify and to relax auxiliary assumptions and to account for all evidence and uncertainty in the analysis of routinely collected data
- Qualitatively discussing biases arising from the the violation of auxiliary assumptions makes scientific uncertainty essentially inaccessible to readers who have limited knowledge and understanding of methodological weaknesses and challenges

Thank you for your attention!


Benjamin A Goldstein, Nrupen A Bhavsar, Matthew Phelan, and Michael J Pencina.
Controlling for informed presence bias due to the number of health encounters in an electronic health record.
American journal of epidemiology, 184(11):847–855, 2016.


Milena A Gianfrancesco and Neal D Goldstein.
A narrative review on the validity of electronic health record-based research in epidemiology.
BMC medical research methodology, 21(1):234, 2021.


Ariana Mihan and Harriette GC Van Spall.
Interventions to enhance digital health equity in cardiovascular care.
Nature Medicine, 30(3):628–630, 2024.


Ariana Mihan, Ambarish Pandey, and Harriette GC Van Spall.
Mitigating the risk of artificial intelligence bias in cardiovascular care.
The Lancet Digital Health, 6(10):e749–e754, 2024.


Ban Al-Sahab, Alan Leviton, Tobias Loddenkemper, Nigel Paneth, and Bo Zhang.
Biases in electronic health records data for generating real-world evidence: An overview.
Journal of Healthcare Informatics Research, 8(1):121–139, 2024.


Nicholas C Arpey, Anne H Gaglioti, and Marcy E Rosenbaum.
How socioeconomic status affects patient perceptions of health care: a qualitative study.
Journal of primary care & community health, 8(3):169–175, 2017.


Judy H Ng, Faye Ye, Lauren M Ward, Samuel C,ÄúChris,Äù Haffer, and Sarah Hudson Scholle.
Data on race, ethnicity, and language largely incomplete for managed care plan members.
Health Affairs, 36(3):548–552, 2017.


Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk.

Potential biases in machine learning algorithms using electronic health record data.
JAMA internal medicine, 178(11):1544–1547, 2018.


Kaiwen Ni, Hongling Chu, Lin Zeng, Nan Li, and Yiming Zhao.
Barriers and facilitators to data quality of electronic health records used for clinical research in china: a qualitative study.
BMJ open, 9(7):e029314, 2019.


Robert A Verheij, Vasa Curcin, Brendan C Delaney, and Mark M McGilchrist.
Possible sources of bias in primary care electronic health record data use and reuse.
Journal of medical Internet research, 20(5):e185, 2018.


Sally Yaacoub, Raphael Porcher, Anna Pellat, Hillary Bonnet, Viet-Thi Tran, Philippe Ravaud, and Isabelle Boutron.
Characteristics of non-randomised studies of drug treatments: cross sectional study.
BMJ medicine, 3(1):e000932, 2024.


Stephany N Duda, Bryan E Shepherd, Cynthia S Gadd, Daniel R Masys, and Catherine C McGowan.
Measuring the quality of observational study data in an international hiv research network.
PLoS one, 7(4):e33908, 2012.


Adam C Fields, Pamela Lu, Deanna L Palenzuela, Ronald Bleday, Joel E Goldberg, Jennifer Irani, Jennifer S Davids, and Nelya Melnitchouk.
Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection?
Surgery, 165(5):953–957, 2019.


Scott A Turner, Hee Soo Jung, and John E Scarborough.
Utilization of a specimen retrieval bag during laparoscopic appendectomy for both uncomplicated and complicated appendicitis is not associated with a decrease in postoperative surgical site infection rates.
Surgery, 165(6):1199–1202, 2019.


Christopher P Childers and Melinda Maggard-Gibbons.
Same data, opposite results?: A call to improve surgical database research.
JAMA surgery, 156(3):219–220, 2021.


Yeganeh Khazaei, Helmut Küchenhoff, Sabine Hoffmann, Diella Syliqi, and Raphael Rehms.
Using a bayesian hierarchical approach to study the association between non-pharmaceutical interventions and the spread of covid-19 in germany.
Scientific Reports, 13(1):18900, 2023.


Raphael Rehms, Nicole Ellenbach, Eva Rehfuess, Jacob Burns, Ulrich Mansmann, and **Hoffmann, Sabine**.
A Bayesian hierarchical approach to account for evidence and uncertainty in the modeling of infectious diseases: An application to COVID-19.
Biometrical Journal, 66(1):2200341, 2024.

Criteria to assure the quality of studies on routinely collected data

	Minimal requirements	**Acceptable**	**Ideal approach**
Eligibility	Detailed comparison of patient characteristics with target population and across centres		
Data quality	Extensive data quality checks and reporting of data pre-processing and missing patterns		
Time point alignment	Report exact timing of all measurements and of treatment trajectories		
Interventions and tests not random	Directed acyclic graph including unmeasured and mismeasured confounders		
Multiplicity of analysis	Pre-registration of		

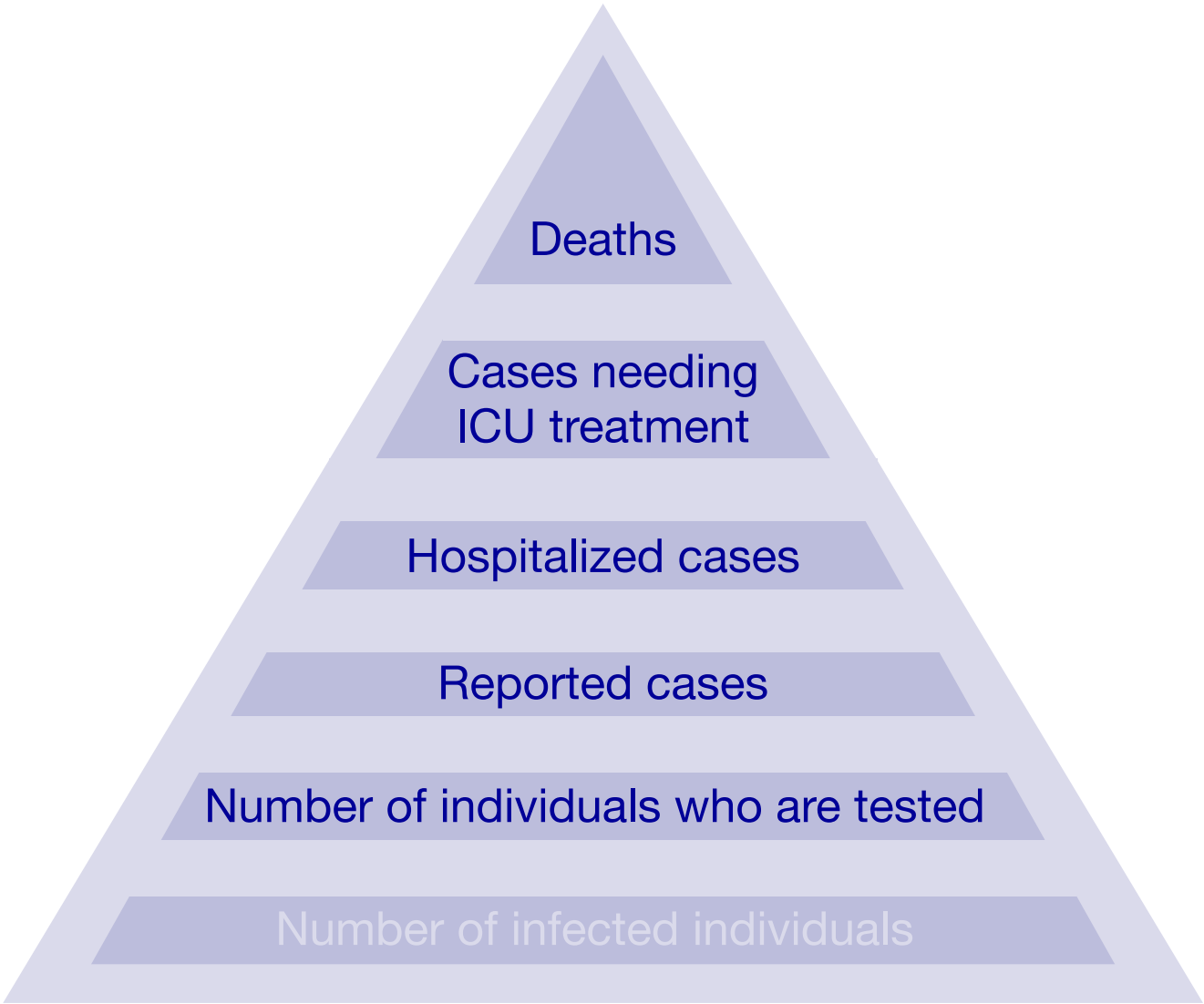
Criteria to assure the quality of studies on routinely collected data

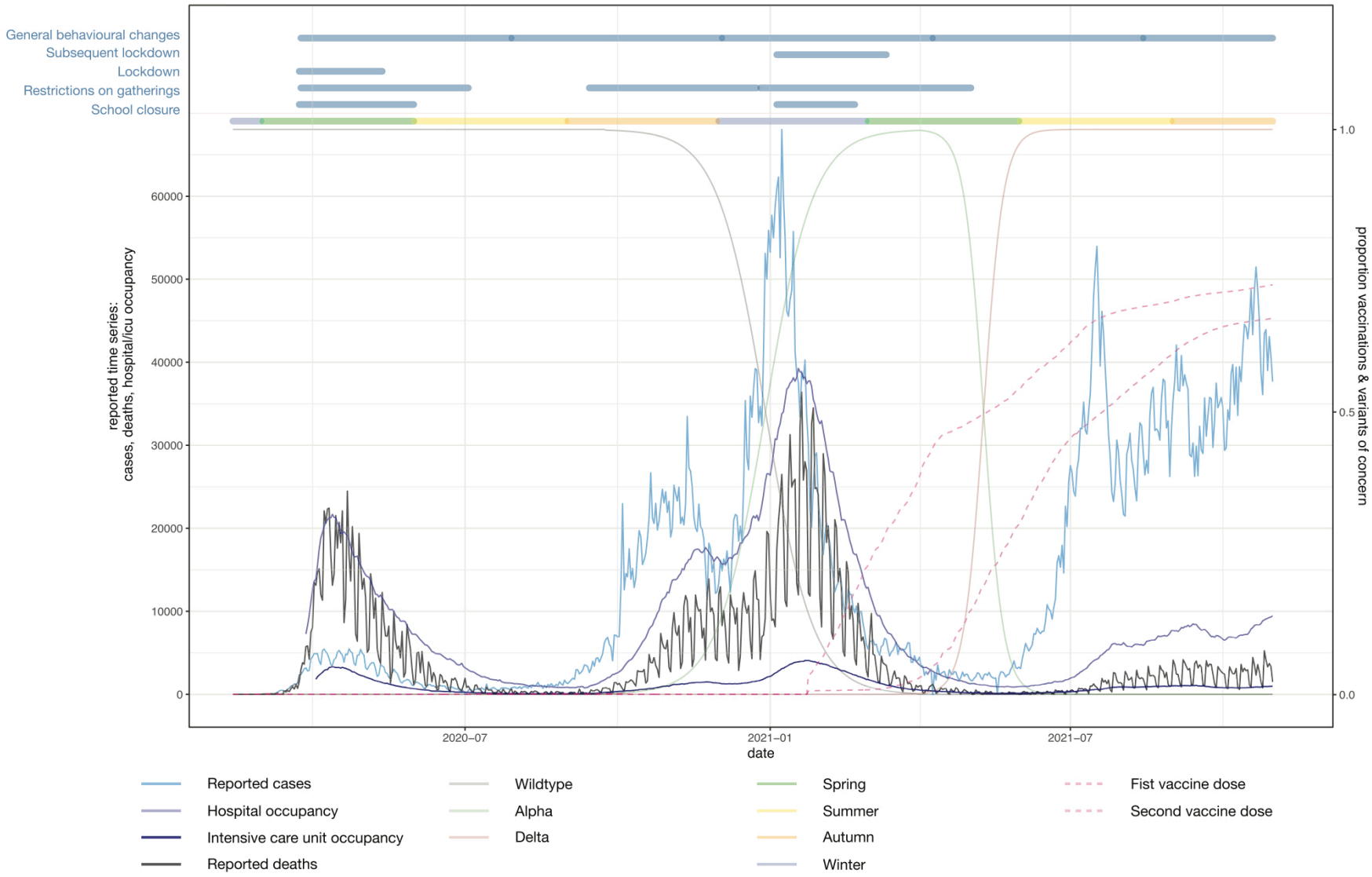
	Minimal requirements	**Acceptable**	**Ideal approach**
Eligibility	Detailed comparison of patient characteristics with target population and across centres	Inverse probability weighting or multilevel regression modelling with post-stratification	
Data quality	Extensive data quality checks and reporting of data pre-processing and missing patterns	Audits, standardisation and training in data collection, validation data to quantify errors	
Time point alignment	Report exact timing of all measurements and of treatment trajectories	Target trial emulation	
Interventions and tests not random	Directed acyclic graph including unmeasured and mismeasured confounders	Quantitative bias analysis and falsification endpoints of negative controls	
Multiplicity of analysis	Pre-registration of	Multi-analyst studies or extensive multiverse	

Criteria to assure the quality of studies on routinely collected data

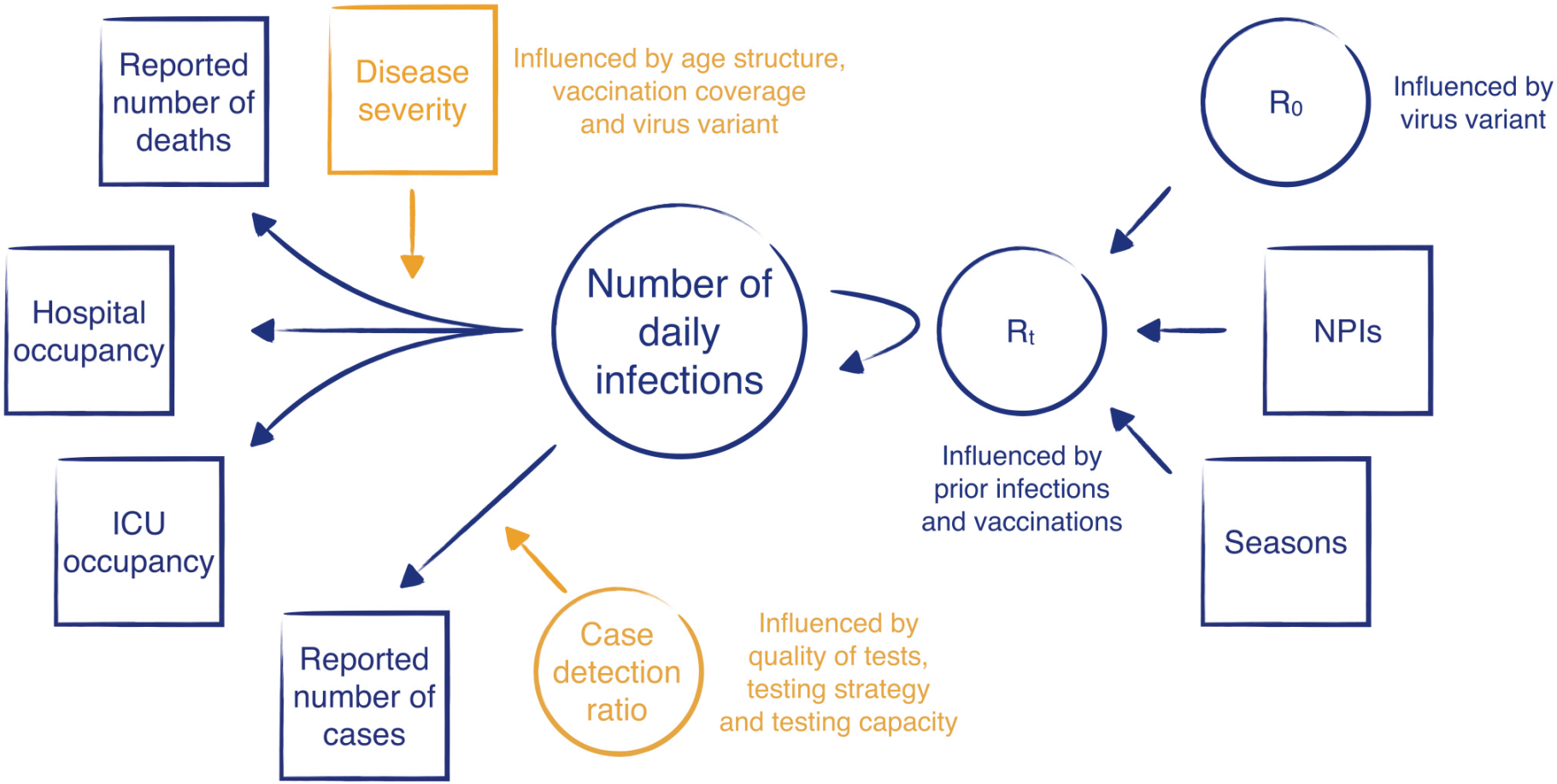
	Minimal requirements	**Acceptable**	**Ideal approach**
Eligibility	Detailed comparison of patient characteristics with target population and across centres	Inverse probability weighting or multilevel regression modelling with post-stratification	Combination of data with more representative data sources
Data quality	Extensive data quality checks and reporting of data pre-processing and missing patterns	Audits, standardisation and training in data collection, validation data to quantify errors	Account for complex measurement error and informative missing data
Time point alignment	Report exact timing of all measurements and of treatment trajectories	Target trial emulation	Validate accuracy of time stamps, report reasons for treatment switches
Interventions and tests not random	Directed acyclic graph including unmeasured and mismeasured confounders	Quantitative bias analysis and falsification endpoints of negative controls	Combined analysis with RCT, high quality documentation of treatment decisions
Multiplicity of analysis	Pre-registration of	Multi-analyst studies or extensive multiverse	Uncertainty intervals that account for

Sources of information in the modelling of COVID-19




General behavioural changes
Subsequent lockdown
Lockdown
Restrictions on gatherings
School closure
reported time series:
cases, deaths, hospital/icu occupancy
proportion vaccinations & variants of concern
0
10000
20000
30000
40000
50000
60000
0.0
0.5
1.0
2020−07
2021−01
2021−07
date
Reported cases
Hospital occupancy
Intensive care unit occupancy
Reported deaths
Wildtype
Alpha
Delta
Spring
Summer
Autumn
Winter
Fist vaccine dose
Second vaccine dose

[16, 17]


Reported number of deaths
Disease severity
Influenced by age structure, vaccination coverage and virus variant
R_0
Influenced by virus variant
Hospital occupancy
Number of daily infections
R_t
NPIs
ICU occupancy
Influenced by prior infections and vaccinations
Seasons
Reported number of cases
Case detection ratio
Influenced by quality of tests, testing strategy and testing capacity

[16, 17]

A hierarchical model of COVID-19 propagation

- The disease model:

$$C_{t,m} = \sum_{u<t} I_{u,m}(F_{\xi^C}(t-u+1) - F_{\xi^C}(t-u))$$

A hierarchical model of COVID-19 propagation

- The disease model:

$$C_{t,m} = \sum_{u<t} I_{u,m}(F_{\xi^C}(t-u+1) - F_{\xi^C}(t-u))$$

- The death model:
  $D_{t,m} \sim$
  Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi^D}(t-u+1) - F_{\xi^D}(t-u)), \phi_d\right)$

A hierarchical model of COVID-19 propagation

- The disease model:

$$C_{t,m} = \sum_{u<t} I_{u,m}(F_{\xi^C}(t-u+1) - F_{\xi^C}(t-u))$$

- The death model:

$D_{t,m} \sim$
Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi^D}(t-u+1) - F_{\xi^D}(t-u)), \phi_d\right)$

- The reporting model:

$C_{t,m}^R \sim$
Negative Binomial $\left(\rho_{t,m} \sum_{u<t} C_{u,m}(F_{\xi^R}(t-u+1) - F_{\xi^R}(t-u)), \phi_c\right)$

A hierarchical model of COVID-19 propagation

- The disease model:

$$C_{t,m} = \sum_{u<t} I_{u,m}(F_{\xi^C}(t-u+1) - F_{\xi^C}(t-u))$$

- The death model:
  $D_{t,m} \sim$
  Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi_m^D}(t-u+1) - F_{\xi_m^D}(t-u)), \phi_d\right)$
- The reporting model:
  $C_{t,m}^R \sim$
  Negative Binomial $\left(\rho_{t,m} \sum_{u<t} C_{u,m}(F_{\xi_m^R}(t-u+1) - F_{\xi_m^R}(t-u)), \phi_c\right)$

A hierarchical model of COVID-19 propagation

- The disease model:

$$C_{t,m} = \sum_{u<t} I_{u,m}(F_{\xi^C}(t-u+1) - F_{\xi^C}(t-u))$$

- The death model:

$D_{t,m} \sim$

Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi_m^D}(t-u+1) - F_{\xi_m^D}(t-u)), \phi_d\right)$

- The reporting model:

$C_{t,m}^R \sim$

Negative Binomial $\left(\rho_{t,m} \sum_{u<t} C_{u,m}(F_{\xi_m^R}(t-u+1) - F_{\xi_m^R}(t-u)), \phi_c\right)$

- The hospitalization model:

$H_{t,m} \sim$

Negative Binomial $\left(\pi_m^H \sum_{u<t} C_{u,m}(F_{\xi^H}(t-u+1) - F_{\xi^H}(t-u)), \phi_h\right)$

A hierarchical model of COVID-19 propagation

- The death model:
  $D_{t,m} \sim$
  Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi_m^D}(t-u+1) - F_{\xi_m^D}(t-u)), \phi_d\right)$
- The reporting model:
  $C_{t,m}^R \sim$
  Negative Binomial $\left(\rho_{t,m} \sum_{u<t} C_{u,m}(F_{\xi_m^R}(t-u+1) - F_{\xi_m^R}(t-u)), \phi_c\right)$
- The hospitalization model:
  $H_{t,m} \sim$
  Negative Binomial $\left(\pi_m^H \sum_{u<t} C_{u,m}(F_{\xi^H}(t-u+1) - F_{\xi^H}(t-u)), \phi_h\right)$
- The renewal model:
  $I_{t,m} \sim$
  Negative Binomial $\left(R_{t,m} \sum_{u<t} I_{u,m}(F_{\gamma}(t-u+1) - F_{\gamma}(t-u)), \phi_i\right)$

A hierarchical model of COVID-19 propagation

- The death model:

  $D_{t,m} \sim$

  Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi_m^D}(t-u+1) - F_{\xi_m^D}(t-u)), \phi_d\right)$

- The renewal model:

  $I_{t,m} \sim$

  Negative Binomial $\left(R_{t,m} \sum_{u<t} I_{u,m}(F_\gamma(t-u+1) - F_\gamma(t-u)), \phi_i\right)$

$$R_{t,m} = R_m^0 \cdot \exp\left(-\sum_{k=1}^{K} \alpha_{k,m} In_{k,t,m}\right)$$

$$R_m^0 \sim \mathcal{N}(R^0, \sigma_R)$$

$$\alpha_{k,m} \sim \mathcal{N}(\alpha_k, \sigma_{\alpha_k})$$

A hierarchical model of COVID-19 propagation

- The death model:
  $D_{t,m} \sim$
  Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi_m^D}(t-u+1) - F_{\xi_m^D}(t-u)), \phi_d\right)$

- The renewal model:
  $I_{t,m} \sim$ Negative Binomial (τ_m, ϕ_i) for t = 1
  $I_{t,m} \sim$
  Negative Binomial $\left(R_{t,m} \sum_{u<t} I_{u,m}(F_\gamma(t-u+1) - F_\gamma(t-u)), \phi_i\right)$

$$R_{t,m} = R_m^0 \cdot \exp\left(-\sum_{k=1}^{K} \alpha_{k,m} In_{k,t,m}\right)$$

A hierarchical model of COVID-19 propagation

- The death model:
  $D_{t,m} \sim$
  Negative Binomial $\left(\pi_m^D \sum_{u<t} C_{u,m}(F_{\xi_m^D}(t-u+1) - F_{\xi_m^D}(t-u)), \phi_d\right)$

- The renewal model:
  $I_{t,m} \sim$
  Negative Binomial $\left(R_{t,m} \sum_{u<t} I_{u,m}(F_\gamma(t-u+1) - F_\gamma(t-u)), \phi_i\right)$

$$R_{t,m} = R^0_{t,m} \cdot \exp\left(-\sum_{k=1}^{K} \alpha_{k,m} In_{k,t,m}\right)$$

$$R^0_{t,m} = R^0_m \cdot (1 - p^\alpha_{t,m} - p^\delta_{t,m}) + (1+\beta^\alpha) \cdot R^0_m \cdot p^\alpha_{t,m} + (1+\beta^\delta) \cdot R^0_m \cdot p^\delta_{t,m}$$

A hierarchical model of COVID-19 propagation

- The death model:

  $D_{t,m} \sim$

  Negative Binomial $\left(\pi^{D}_{m,t} \sum_{u<t} C_{u,m}(F_{\xi^{D}_{m}}(t-u+1) - F_{\xi^{D}_{m}}(t-u)), \phi_d\right)$

- The renewal model:

  $I_{t,m} \sim$

  Negative Binomial $\left(R_{t,m} \sum_{u<t} I_{u,m}(F_{\gamma}(t-u+1) - F_{\gamma}(t-u)), \phi_i\right)$

$$R_{t,m} = R^{0}_{t,m} \cdot \exp\left(-\sum_{k=1}^{K} \alpha_{k,m} In_{k,t,m}\right) \cdot (1 - c^{1}_{t,m} - c^{2}_{t,m} \cdot (1 - c^{1}_{t,m}))$$

$$c^{1}_{t,m} = \frac{\sum_{u<t} I_{u,m}}{N_m} \cdot (1 - \beta^{reinf}) \text{ and}$$

$$c^{2}_{t,m} = \frac{\sum_{u<t} Vacc^{1}_{u,m} \cdot \beta^{v1} + Vacc^{2}_{u,m} \cdot \beta^{v2}}{N_m}$$

A hierarchical model of COVID-19 propagation

- The death model:

  $D_{t,m} \sim$

  Negative Binomial $\left(\pi^{D}_{m,t} \sum_{u<t} C_{u,m}(F_{\xi^{D}_{m}}(t-u+1) - F_{\xi^{D}_{m}}(t-u)), \phi_d\right)$

- The renewal model:

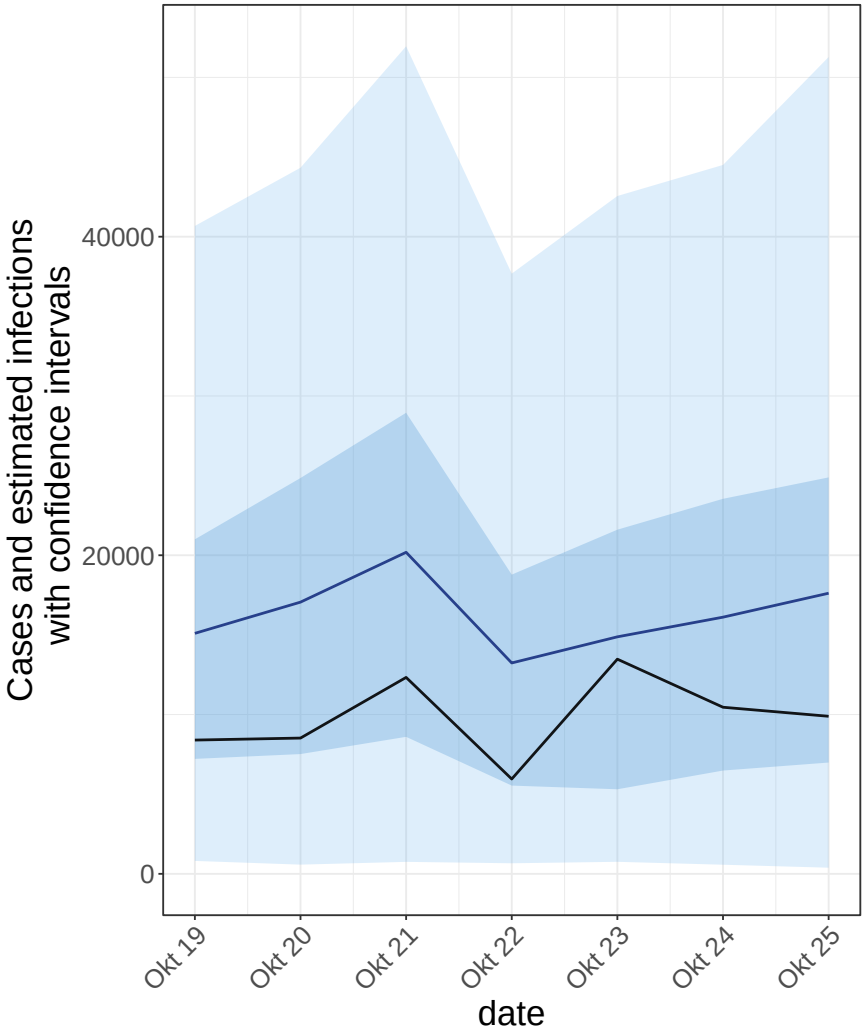
  $I_{t,m} \sim$

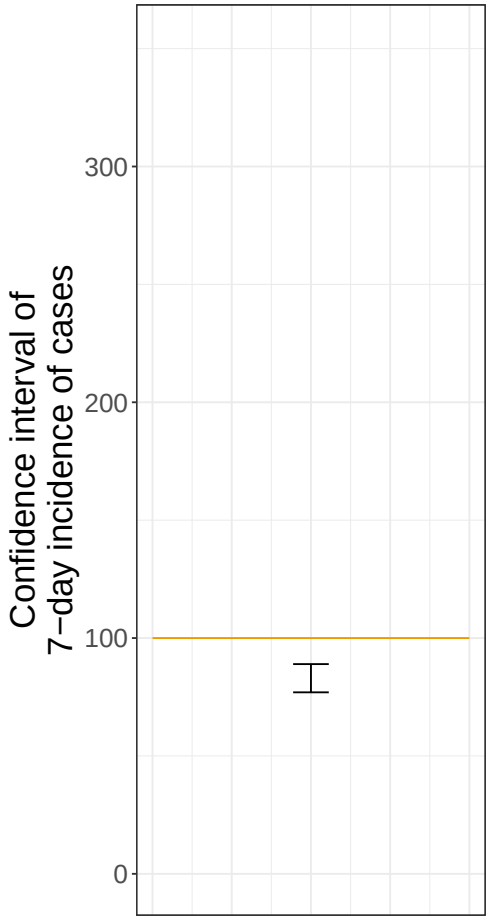
  Negative Binomial $\left(R_{t,m} \sum_{u<t} I_{u,m}(F_{\gamma}(t-u+1) - F_{\gamma}(t-u)), \phi_i\right)$

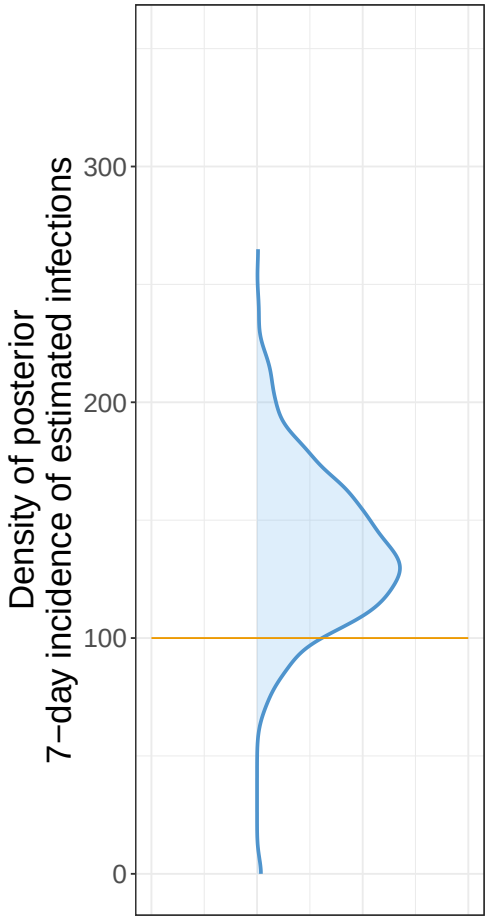
$$R_{t,m} = R^{0}_{t,m} \cdot \exp\left(-\sum_{k=1}^{K} \alpha_{k,m} In_{k,t,m}\right) \cdot (1 - c^{1}_{t,m} - c^{2}_{t,m} \cdot (1 - c^{1}_{t,m}))$$

$$c^{1}_{t,m} = \frac{\sum_{u<t} I_{u,m}}{N_m} \cdot (1 - \beta^{reinf}) \text{ and}$$

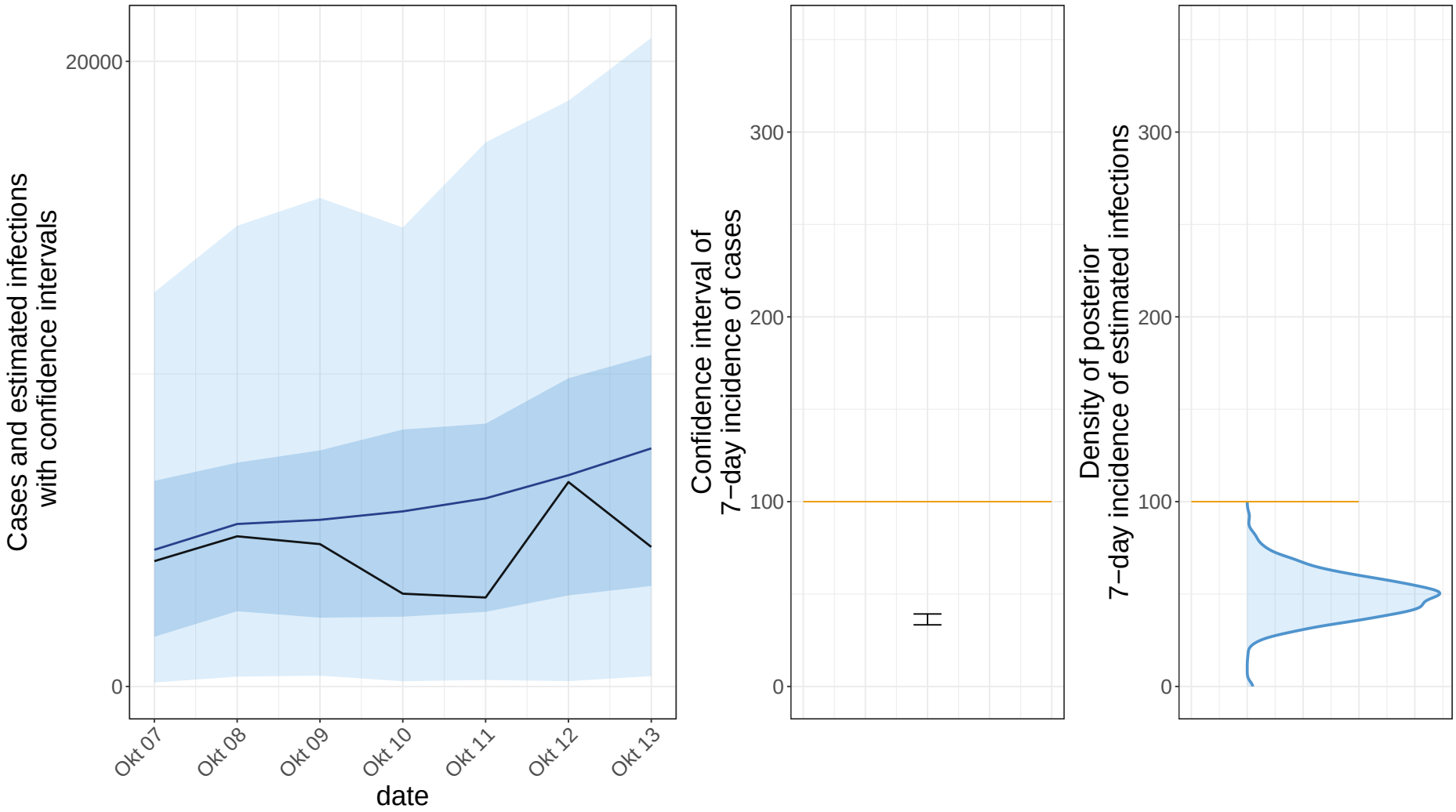
$$c^{2}_{t,m} = \frac{\sum_{u<t} Vacc^{1}_{u,m} \cdot \beta^{v1} + Vacc^{2}_{u,m} \cdot \beta^{v2}}{N_m}$$


Cases and estimated infections
with confidence intervals
40000
20000
0
Okt 19
Okt 20
Okt 21
Okt 22
Okt 23
Okt 24
Okt 25
date

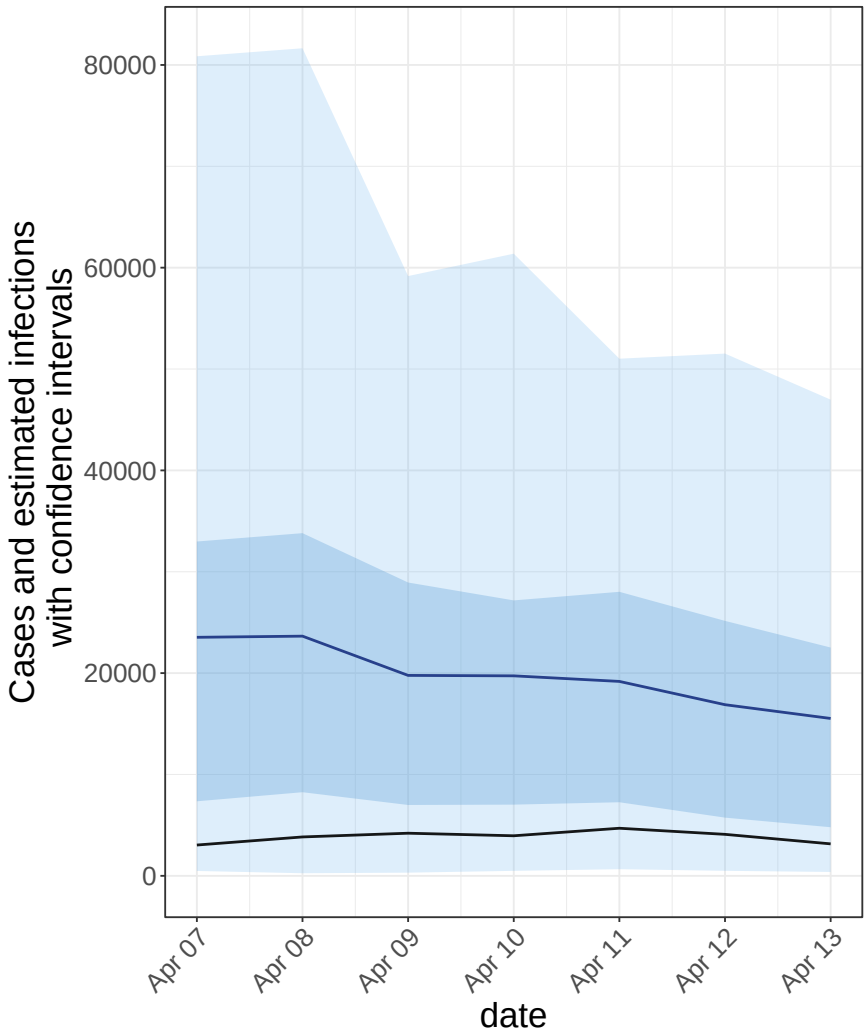

Confidence interval of
7−day incidence of cases
300
200
100
0

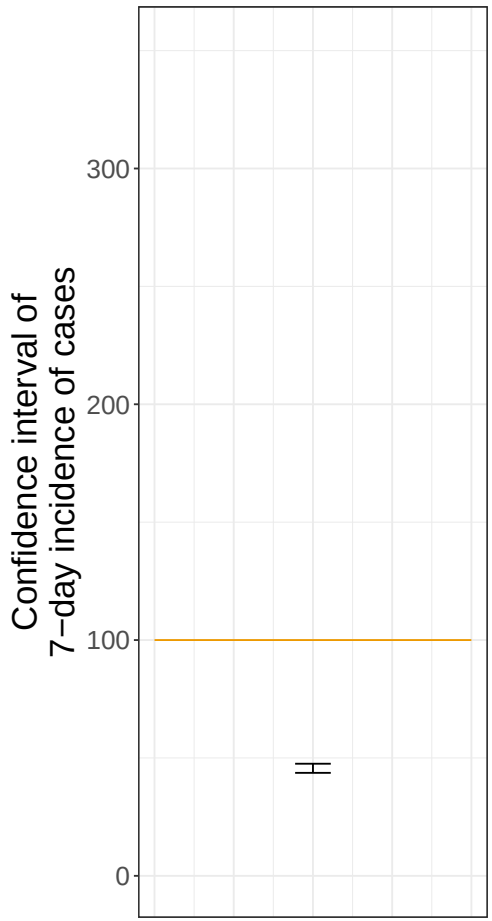

Density of posterior
7−day incidence of estimated infections
300
200
100
0

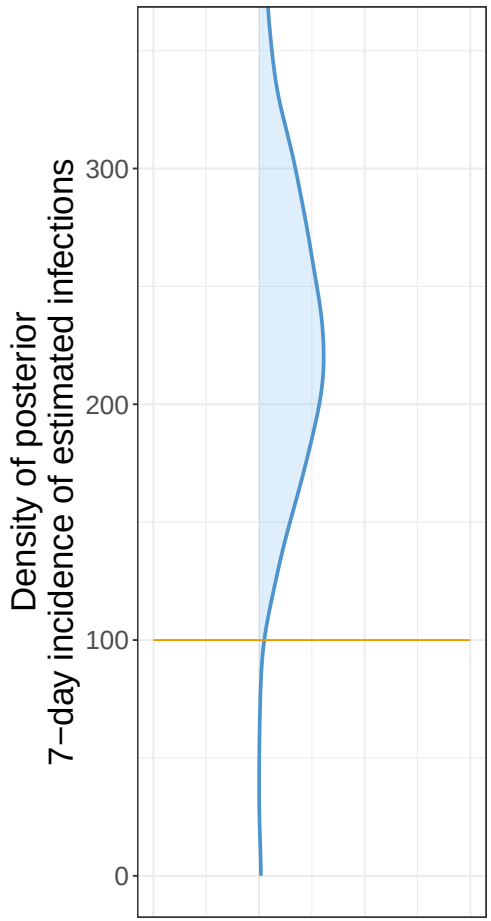
[16, 17]


Cases and estimated infections
with confidence intervals
20000
0
Okt 07
Okt 08
Okt 09
Okt 10
Okt 11
Okt 12
Okt 13
date
Confidence interval of
7−day incidence of cases
300
200
100
0
Density of posterior
7−day incidence of estimated infections
300
200
100
0

[16, 17]


Cases and estimated infections with confidence intervals
80000
60000
40000
20000
0
Apr 07
Apr 08
Apr 09
Apr 10
Apr 11
Apr 12
Apr 13
date


Confidence interval of 7−day incidence of cases
300
200
100
0
H


Density of posterior 7−day incidence of estimated infections
300
200
100
0

[16, 17]