

Causal Record Linkage

Critical compromises with false discoveries

Causal Record Linkage

Link observations X, T, Y collected on different occasions without unique identifier, to obtain 'complete' profiles and perform causal inference.

Records are probabilistically linked based on partially identifying variables X
⇒ linked records carry similar X

- X are also involved in
- the outcome
 - as covariates
 - as effect modifiers ⇒ treatment heterogeneity
 - the treatment assignment

Motivations

- Integrate data to address causal questions without new trial
- Work with pseudonymised data
- Study long-term outcomes
- Explore secondary outcomes

Tools for causality

- Propensity score can be estimated on one file
- Non-adjusted method like stratification may be used
 - avoid direct adjustment of X (which depends on the linkage)
 - enforce conditional exchangeability

Impact on identifiability conditions

- Consistency** $Y(t)$ not well defined
the value we would have observed (had individuals taken the observed treatment) depends on the real latent treatment
linked data ✗ / stringent set of linked data ≈ ✓
- Overlap** potential lack of overlap in some areas when selecting a stringent set of linked records
linked data ✓ / stringent set of linked data ✗
- Exchangeability / No unobserved confounders**
with uncertainty quantification, observables should be rich enough to study causality with stratification
linked data ✓ / stringent set of linked data ≈ ✗

Parallel with defiers in non-compliance problems a *treatment opposite to the one producing the outcome* is observed though the linkage cannot be an instrument

→ Record Linkage as interference

More certainly linked data are outliers of the population, exhibiting rarer values of X → affect inference
→ get rid of the false discoveries, remove attenuation bias, avoid opposite outcome contributions
→ introduce rare values effect, potential lack of overlap and lack of exchangeability

2004	4236XF	0	2004	4236XF	y1(0)
1989	1106BX	1	1989	1106BX	y2(1)
1987	2042NZ	0	1987	2042NZ	y3(0)
2001	1102AB	1	2001	1102AB	y4(1)
1999	3272TC	0	1991	3272TC	y5
1991	3272TC	1	2002	3015CX	y6
1999	3272TC	0	2000	3015CX	y7
1997	3272TC	1	1997	3272TC	y8(0)
1995	3015CX	0	1997	3272TC	y9(1)
1991	3015CX	1	1999	3272TC	y10(1)
1995	3015CX	0	1991	3272TC	y11
1997	3015CX	0	1997	3272TC	y12
			1991	3015CX	y13(1)
			1995	3015CX	y14
			1997	3015CX	y15(0)
			1997	3015CX	y16

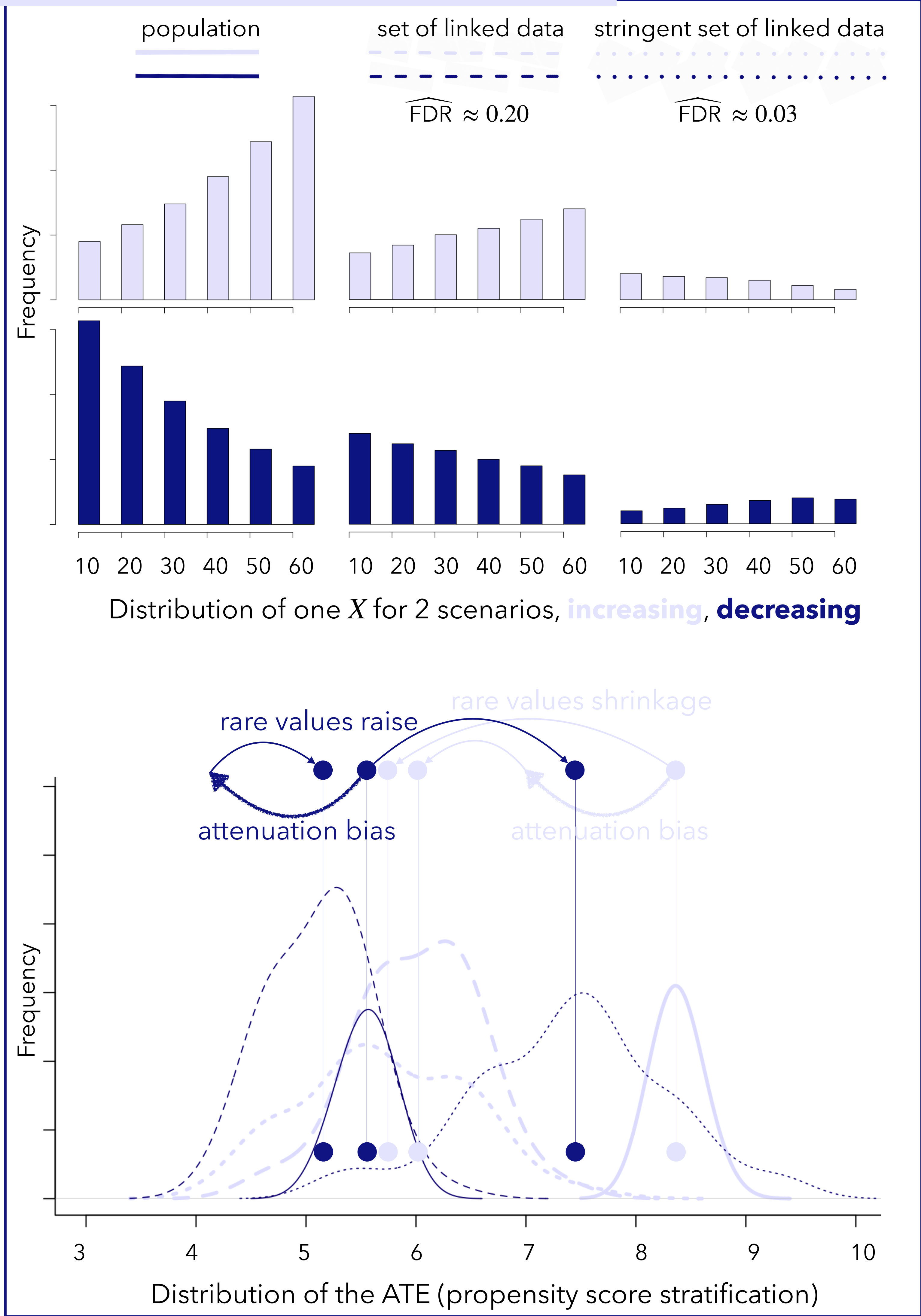
correctly linked records with rare values of X
⇒ high linkage probabilities

(in)correctly linked records with common values of X
⇒ lower linkage probabilities

Inference is conducted on

- True links
- Falsely linked records, with correct / wrong T and or Y
 $y(1) = f(ax + \beta x + \varepsilon)$ may be modelled with $\hat{ax}$
 $y(0) = f(ax + \varepsilon)$ may be modelled with $\hat{ax} + \hat{\beta}x$

Attenuation bias versus rare values effect



→ To be corrected with generalisability methods, from the stringent set of linked records towards the initial population

