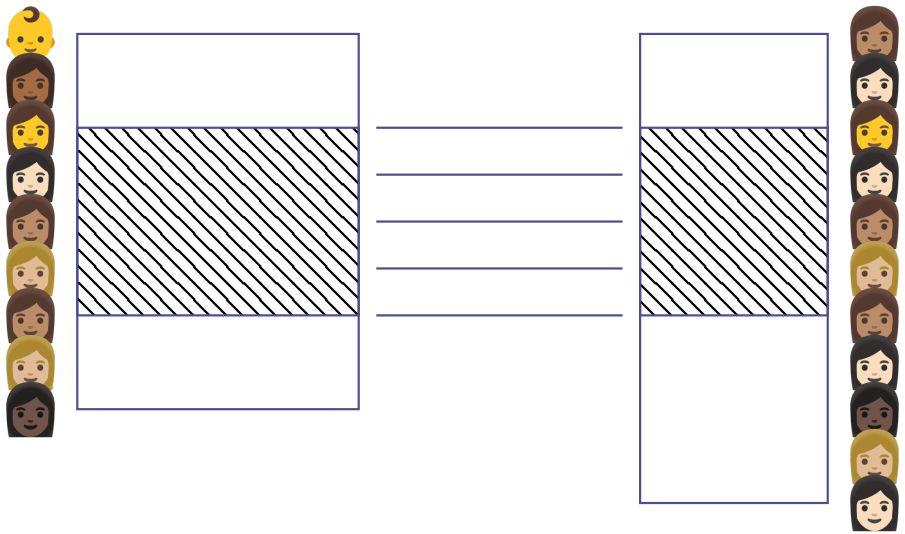
Causal Record Linkage

Kayané Robach

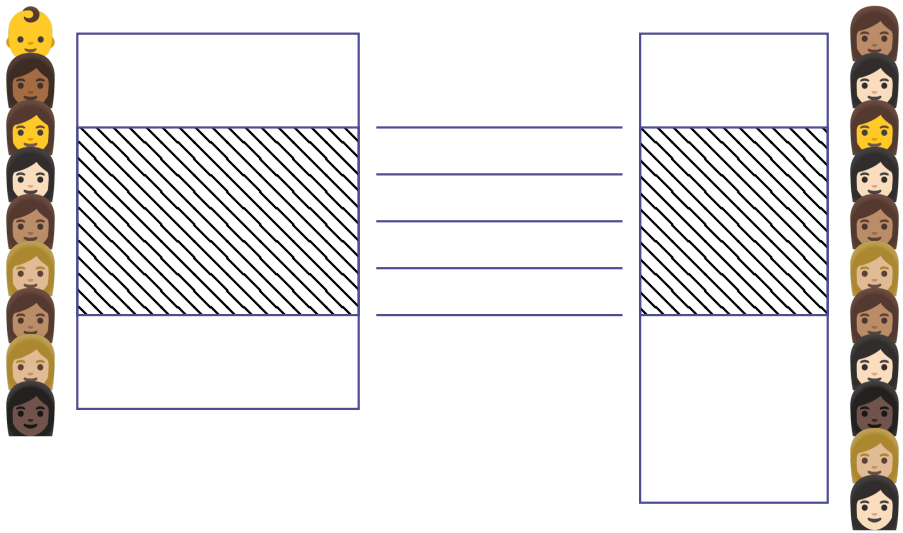
Joint work with Michel Hof, Mark van de Wiel, Stéphanie van der Pas

How to combine individuals records across heterogeneous data sets?

How to combine individuals records across heterogeneous data sets? *Record Linkage*

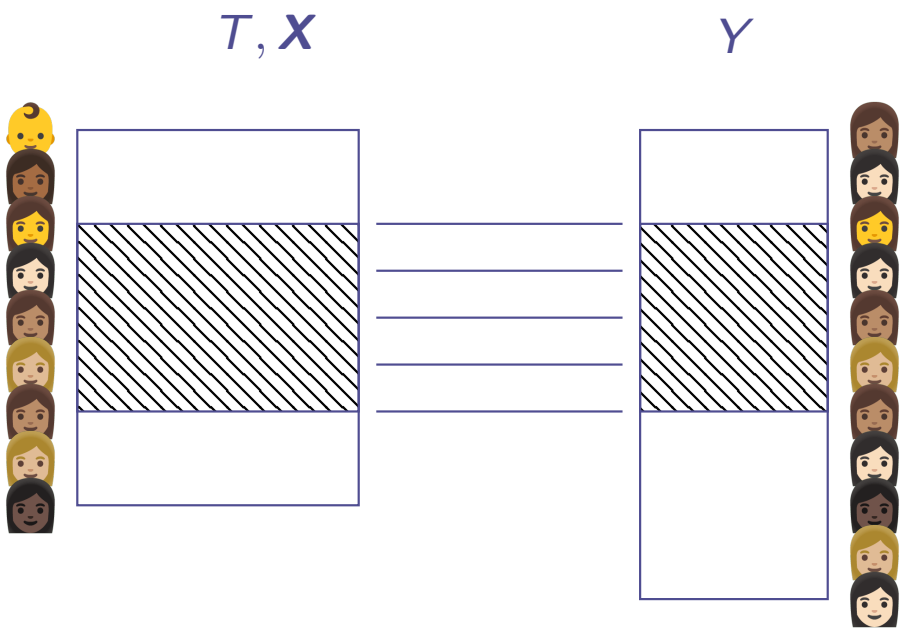


How to combine individuals records across heterogeneous data sets? *Record Linkage*



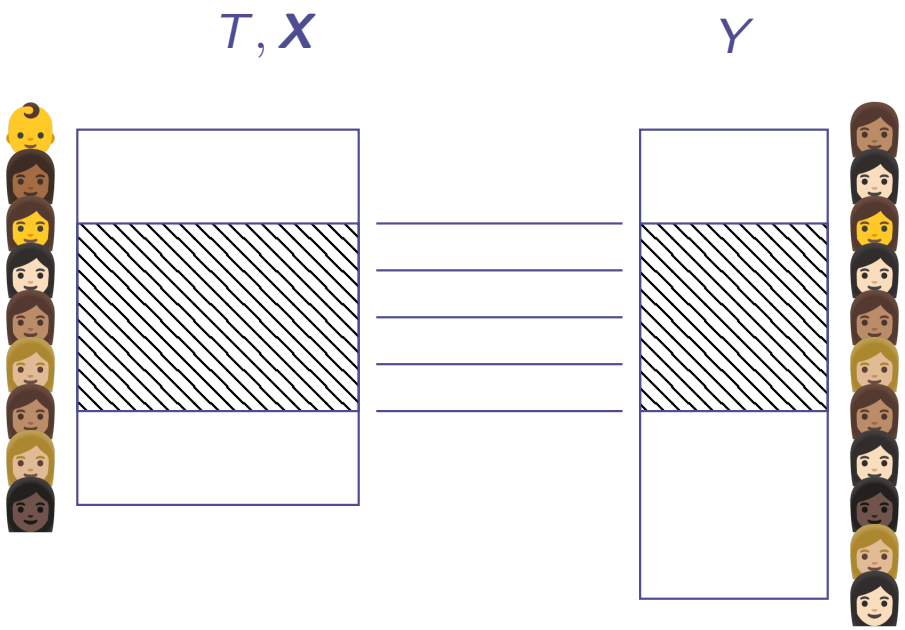
Connect information on the same individuals

How to combine individuals records across heterogeneous data sets? *Record Linkage*



baseline information and exposure $\longleftrightarrow$ outcomes

How to combine individuals records across heterogeneous data sets? *Record Linkage*



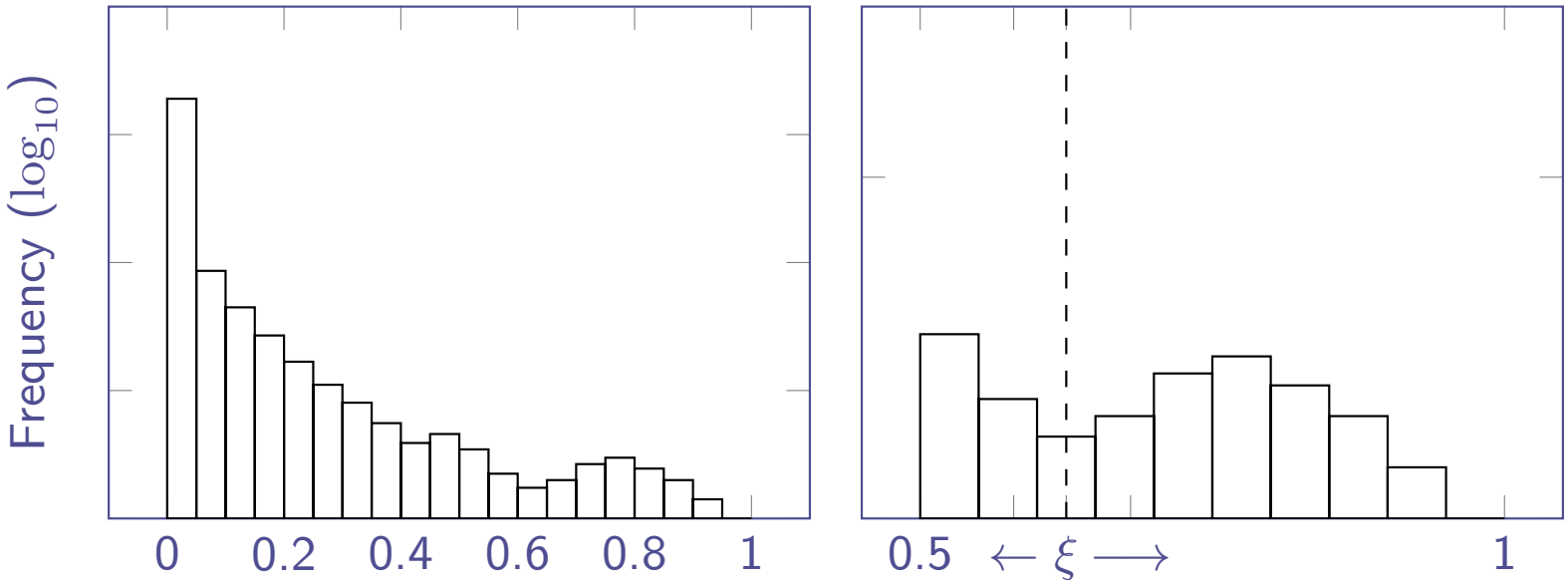
Perform causal inference

Without unique identifiers?

Without unique identifiers $\rightarrow$ use partially identifying information
(birth year, ethnicity, postal code) ⚠️ also causally relevant covariates $\boldsymbol{X}$

Probabilistically recover complete profiles $(T, \boldsymbol{X}, Y)$ with *Record Linkage*

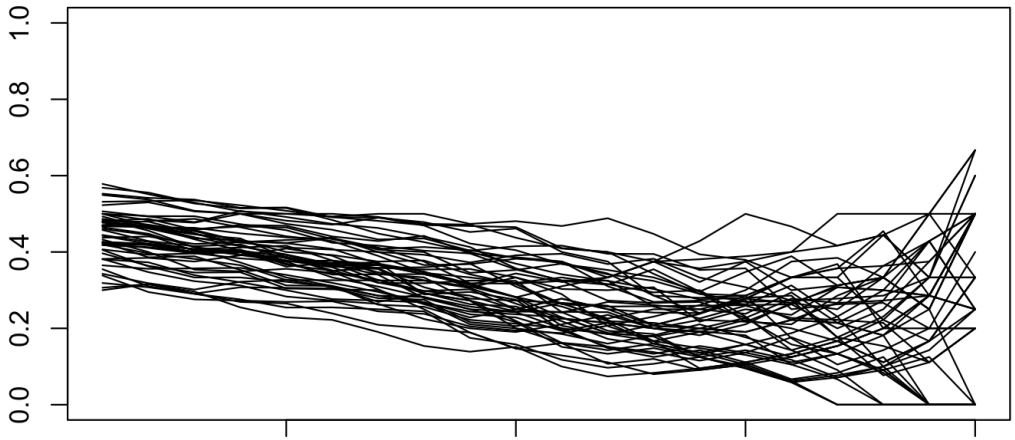
Rigorously select the profiles with linkage score $> \xi$



How to choose this threshold ξ?

We select the set of linked records, trying to minimise:

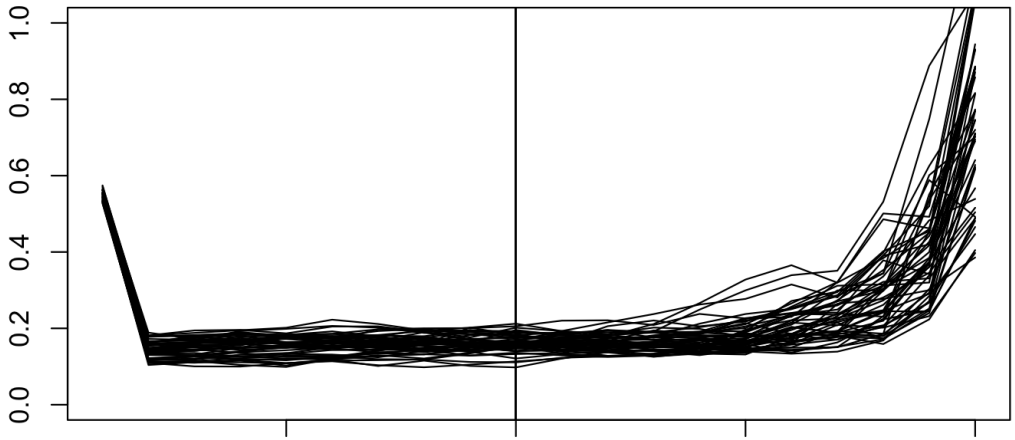
$\rightarrow$ the estimated False Discovery Proportion



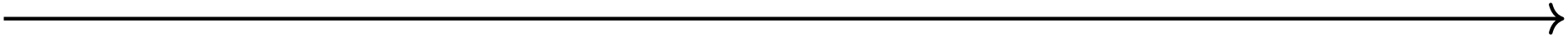
How to choose this threshold ξ?

We select the set of linked records, trying to minimise:

$\rightarrow$ the Maximum Mean Discrepancy



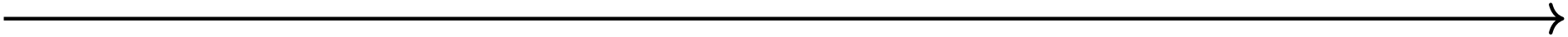
Stringency on the selection of linked records



Low (ξ closer to 0.5)

High (ξ closer to 1)

Stringency on the selection of linked records

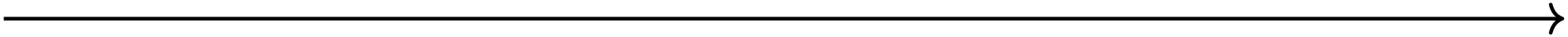


Low (ξ closer to 0.5)

Positivity 🙂

High (ξ closer to 1)

Stringency on the selection of linked records

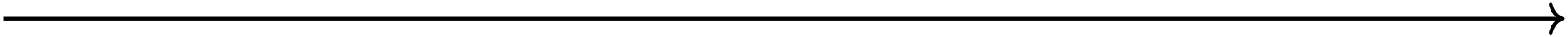


Low (ξ closer to 0.5)

Positivity 🙂
Conditional exchangeability 🥴

High (ξ closer to 1)

Stringency on the selection of linked records

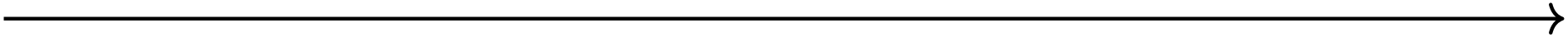


Low (ξ closer to 0.5)

Positivity 🙂
Conditional exchangeability 🥴
Consistency 😵‍💫

High (ξ closer to 1)

Stringency on the selection of linked records



Low (ξ closer to 0.5)	High (ξ closer to 1)
Positivity 🙂	
Conditional exchangeability 🥴	
Consistency 😵	Consistency 🙂

Stringency on the selection of linked records

Low (ξ closer to 0.5)	High (ξ closer to 1)
Positivity 🙂	
Conditional exchangeability 🥴	Conditional exchangeability 🙂
Consistency 😵‍💫	Consistency 🙂

Stringency on the selection of linked records

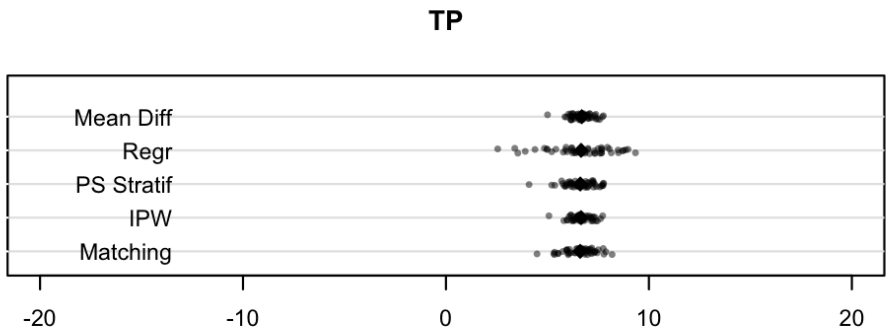
Low (ξ closer to 0.5)

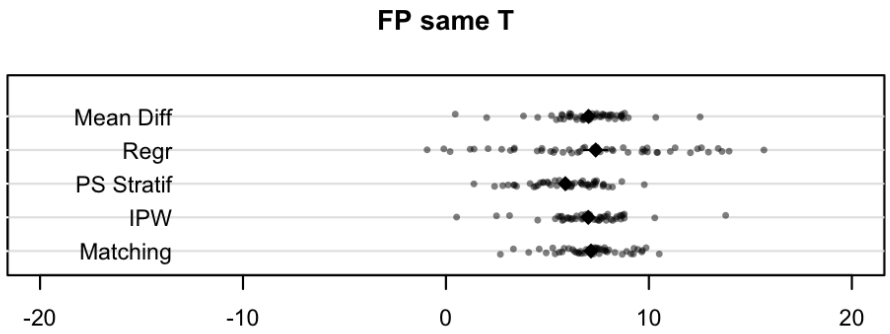
Positivity 🙂
Conditional exchangeability 🥴
Consistency 😵‍💫

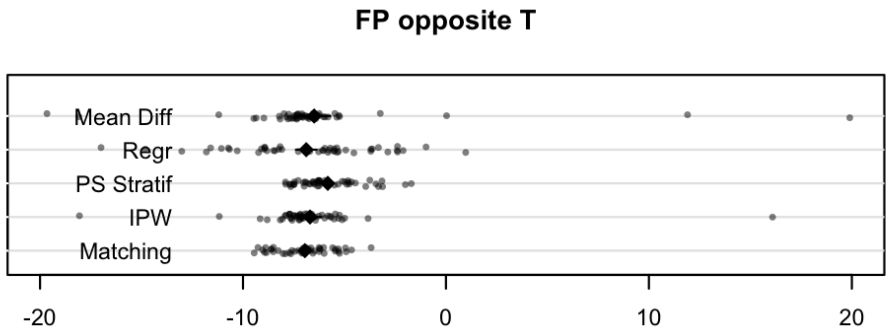
High (ξ closer to 1)

Positivity 🤯
Conditional exchangeability 🙂
Consistency 🙂

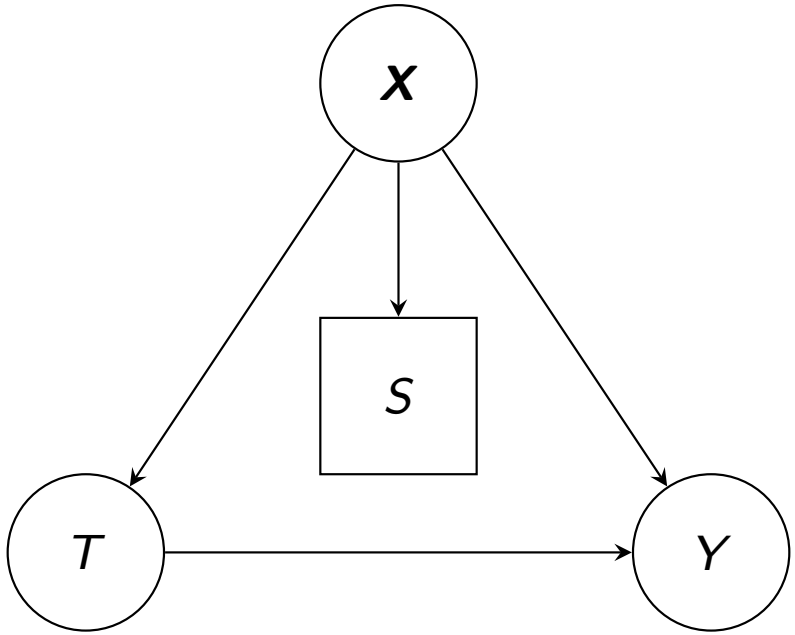
Atypical profiles over-represented!


TP
Mean Diff
Regr
PS Stratif
IPW
Matching
-20
-10
0
10
20

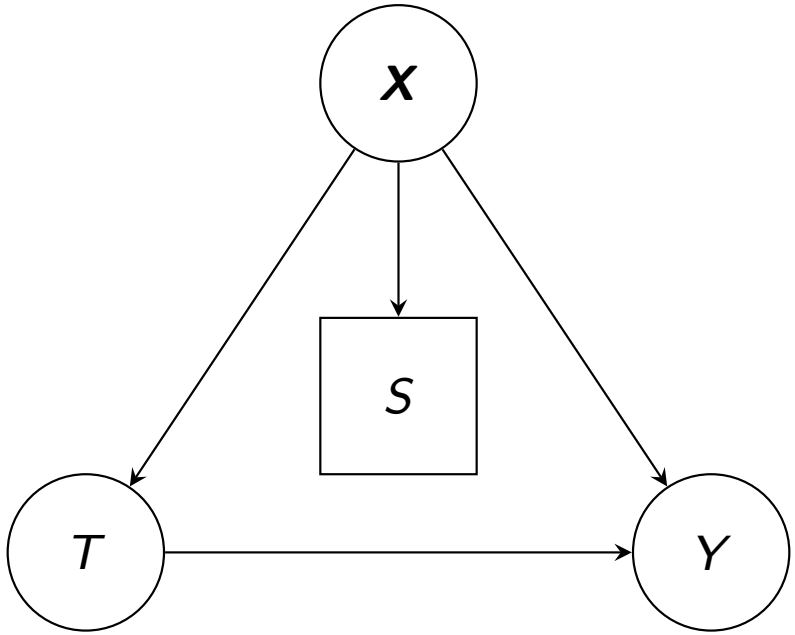

FP same T
Mean Diff
Regr
PS Stratif
IPW
Matching
-20
-10
0
10
20


FP opposite T
Mean Diff
Regr
PS Stratif
IPW
Matching
-20
-10
0
10
20

What is happening?

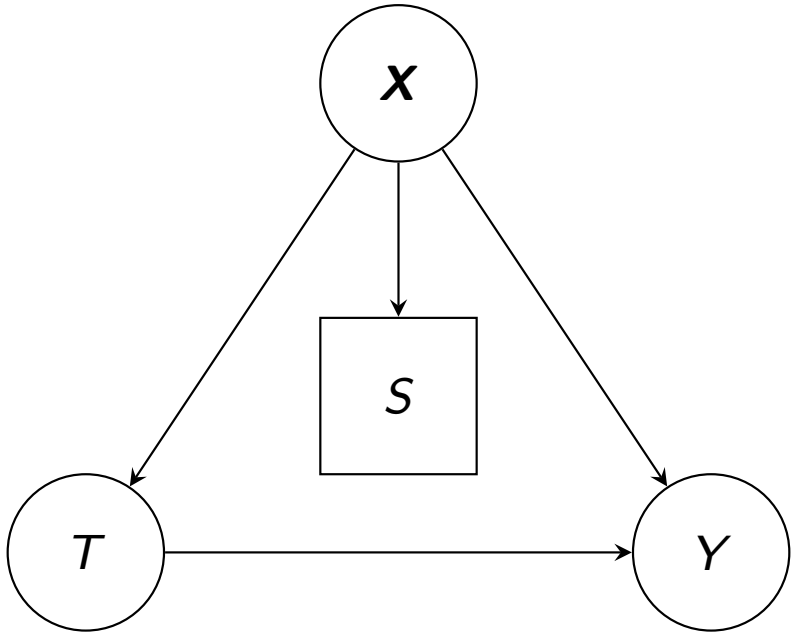


What is happening?



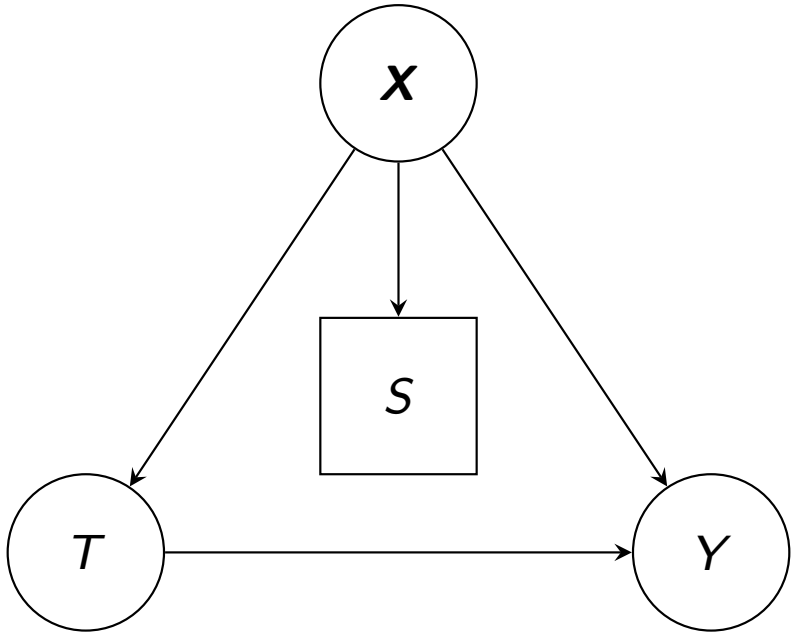
Selection S due to linkage

What is happening?



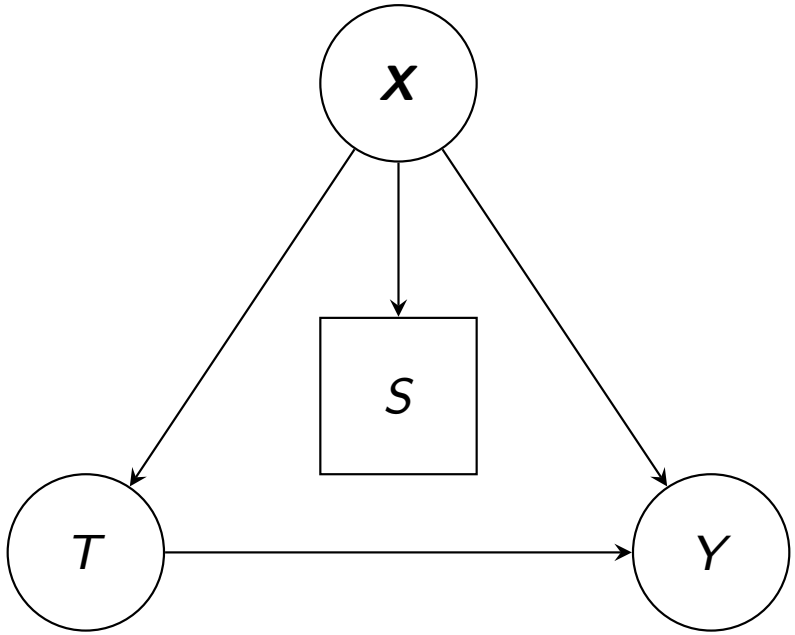
Selection S due to linkage
(*selection backdoor criterion, effect is transportable from selected to baseline population*)

What is happening?



Selection S due to linkage
(*selection backdoor criterion, effect is transportable from selected to baseline population*)
'Recoverability from sampling bias'

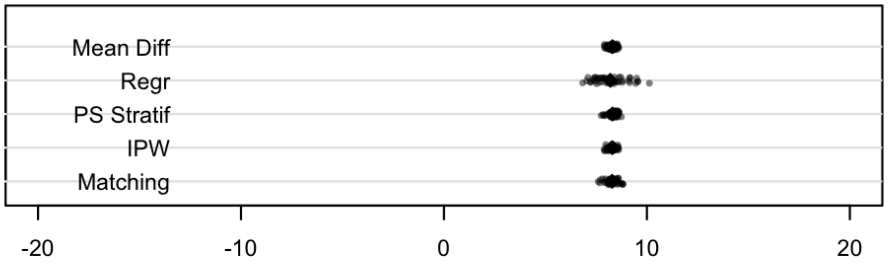
What is happening?



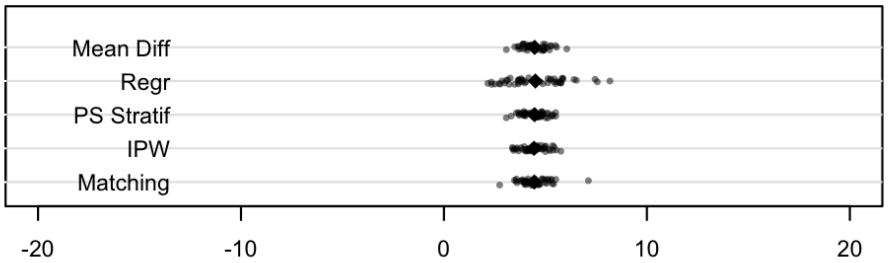
Selection S due to linkage
(*selection backdoor criterion, effect is transportable from selected to baseline population*)
'Recoverability from sampling bias'

$\rightarrow$ Use G-methods

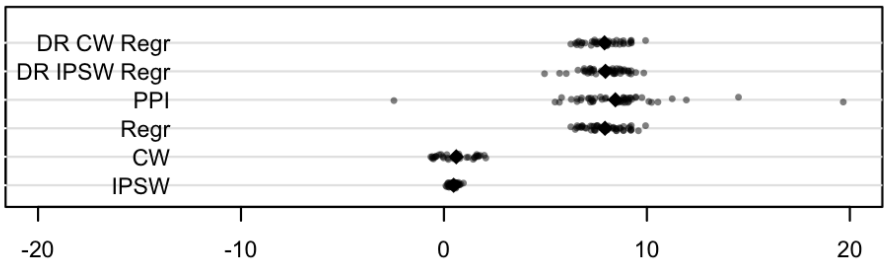
True Links

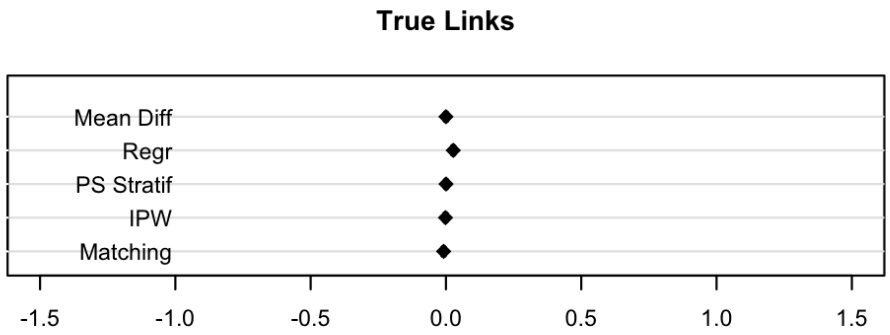


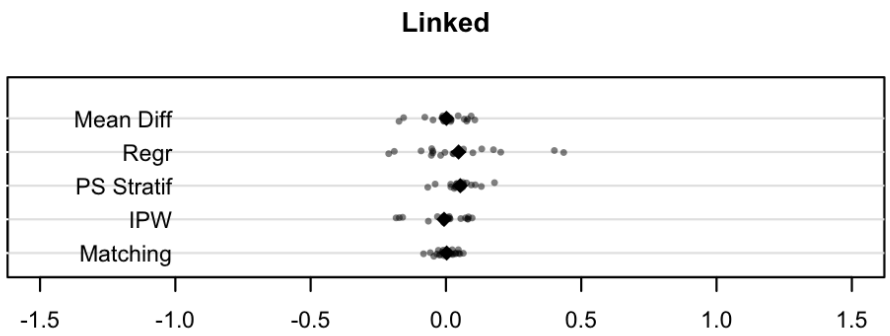
Linked

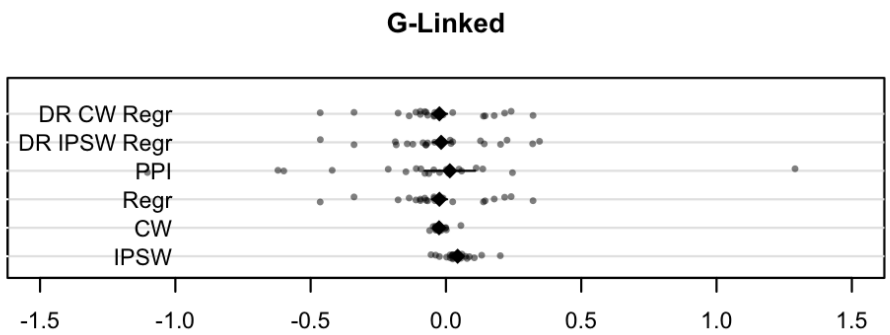


G-Linked




True Links
Mean Diff
Regr
PS Stratif
IPW
Matching
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


Linked
Mean Diff
Regr
PS Stratif
IPW
Matching
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


G-Linked
DR CW Regr
DR IPSW Regr
PPI
Regr
CW
IPSW
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Thank You!

k.c.robach@amsterdamumc.nl