Põhjuslikkus
(Causality)

motivated by

Vanessa Didelez & Naula Sheehan (2007)

Mendelian randomization as an instrumental variable approach to causal inference

Stat Methods Med Res 16: 309-330

Märt Möls
Mart.Mols@ut.ee

Prognoosimine vs toimuva mõjutamine

Tulekahjut kustutama sõitnud tuletõrjeautode arv ja põlengu poolt tekitatud kahju suurus on tugevalt korreleeritud – mida enam autosid, seda suurem kahju

Kas nimetatud seose teadmisest on kasu, kui olete:

- Päästeameti direktor (saata välja vähem tuletõrjeautosid, et vähendada kahju?)
- Ajakirjanik (kumba kahest tulekahjust kajastamiseks valida?)

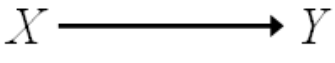
Kureteooria

Summary

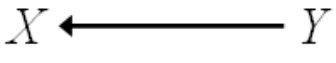
Data from Berlin (Germany) show a significant correlation between the increase in the stork population around the city and the increase in deliveries outside city hospitals (out-of-hospital deliveries). However, there is no correlation between deliveries in hospital buildings (clinical deliveries) and the stork population.

New evidence for the Theory of the Stork. Thomas Höfer, Hildegard Przyrembel, Silvia Verleger. *Paediatric & Perinatal Epidemiology*. Volume 18 Page 88 - January 2004

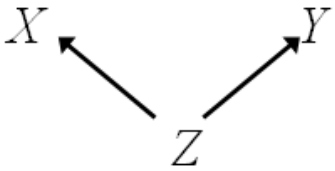
Miks on tunnuste *X* ja *Y* vahel statistiline seos?



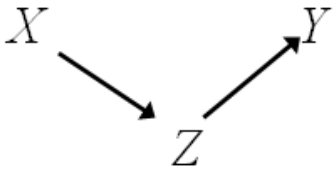
a) X-i mõju Y-le



b) Y'i mõju X-le

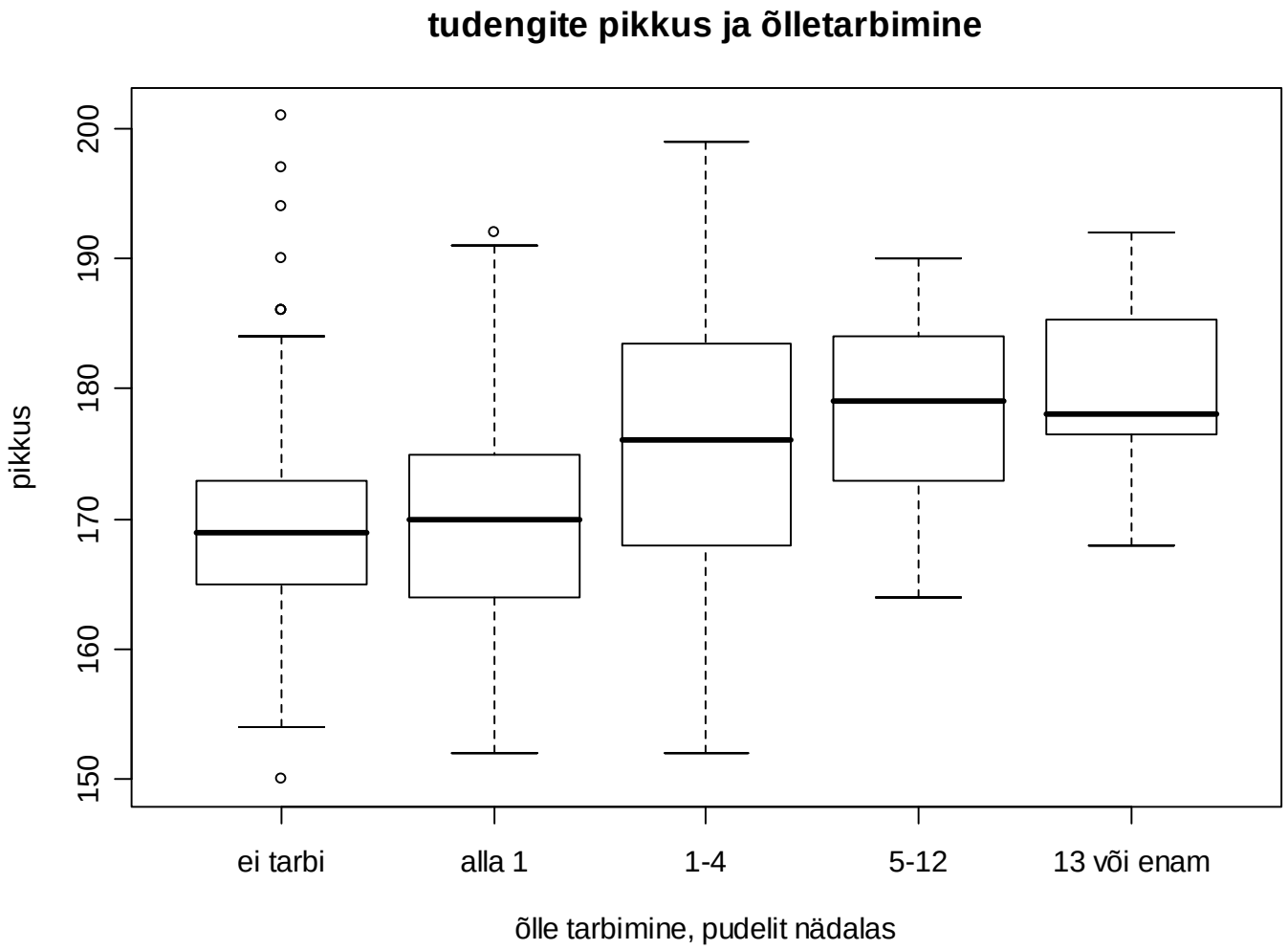


c) kolmanda teguri mõju nii X-le kui Y-le.



d) X mõju Z-le, Z mõju Y-le

Või kombinatsioon ülaltoodud seostest


tudengite pikkus ja õlletarbimine
pikkus
150
160
170
180
190
200
ei tarbi
alla 1
1-4
5-12
13 või enam
õlle tarbimine, pudelit nädalas

Causal relationship

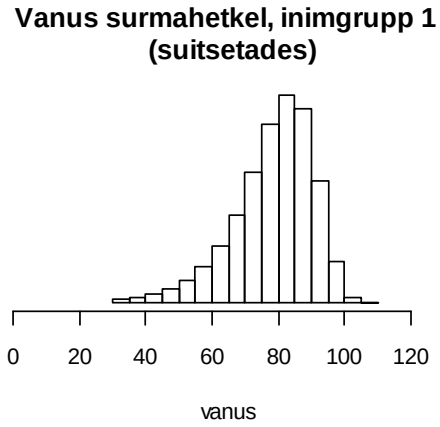
- Counterfactuals

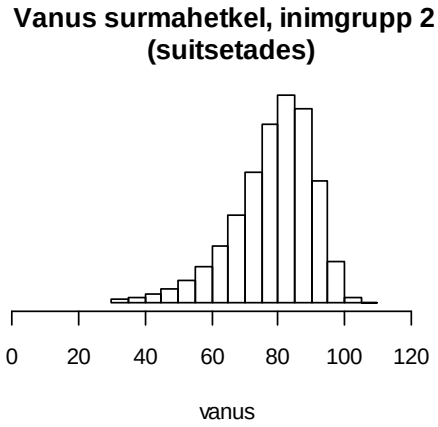
  Jaan suitsetas ja suri noorena.

  Kui Jaan poleks suitsetanud, poleks ta noorena surnud.

  Järelikult põhjustas suitsetamine Jaani surma.

Üksikisiku puhul saavutamatu, inimeste grupi puhul saavutatav



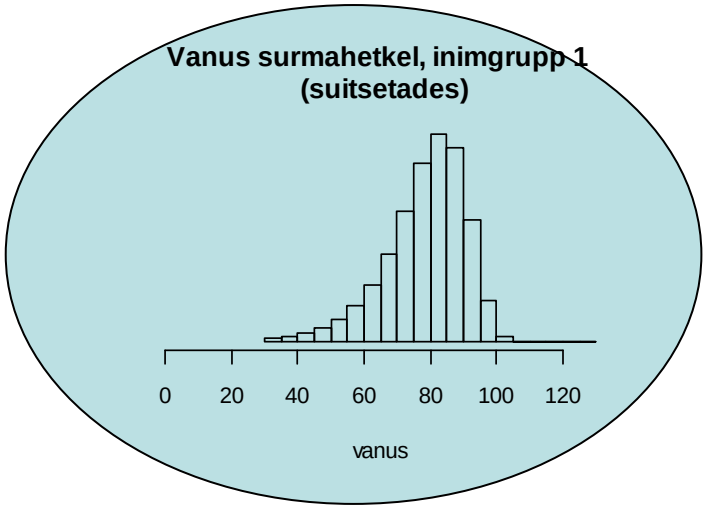


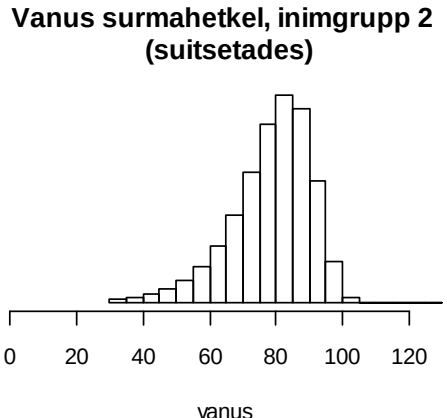
Ideaalvariant: Randomization

Vaese mehe alternatiivid:

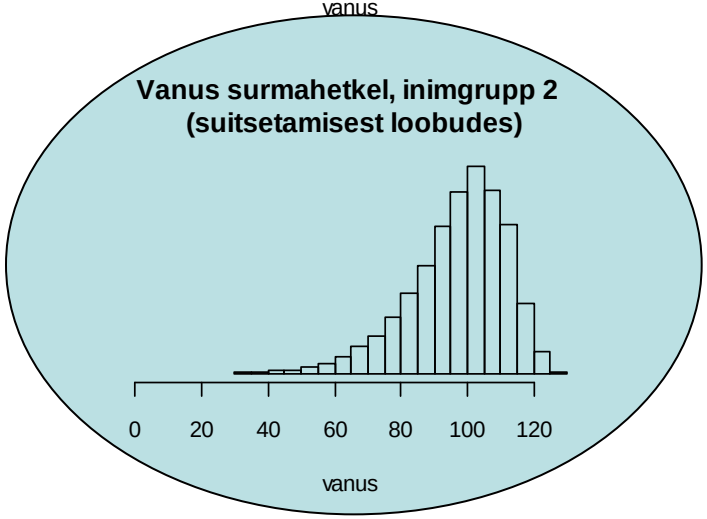
- Matched Cohort Study
- Matched Case-Control Study

Üksikisiku puhul saavutamatu, inimeste grupi puhul saavutatav

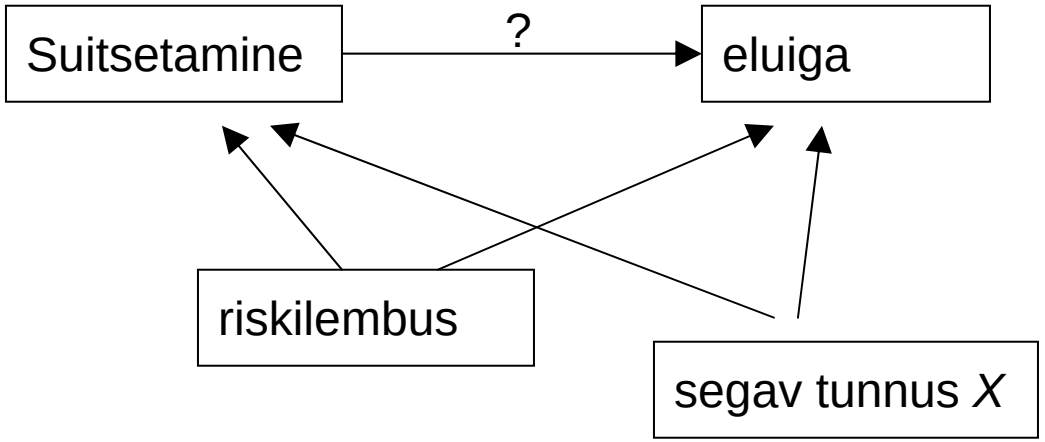




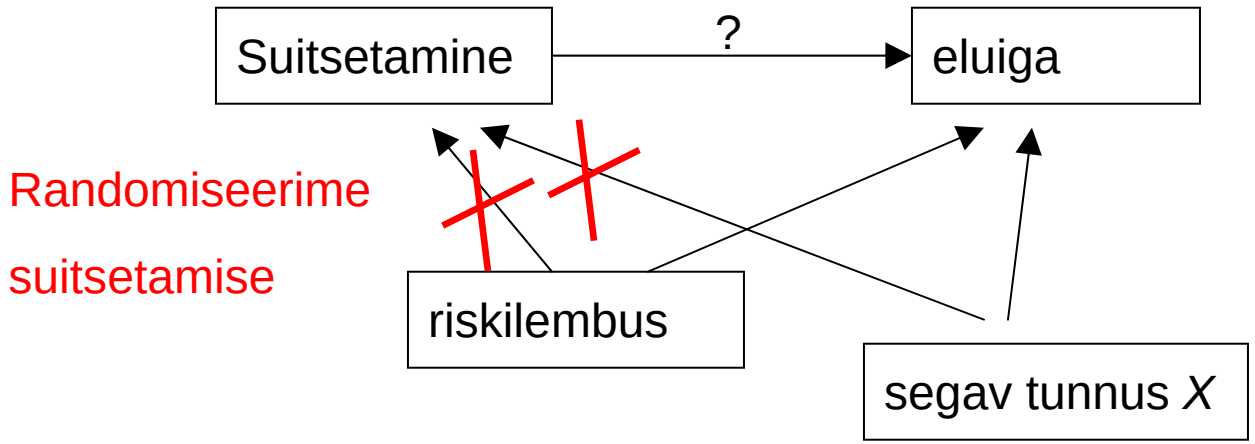
Vaadeldud



Suitsetamise ja kopsuvähi vaheline seos



Suitsetamise ja kopsuvähi vaheline seos, randomiseerime



Alternatiivid randomiseerimisele

- Katsetingimuste (segavate faktorite) fikseerimine ehk kontrollimine

  Võtame vaatluse alla ainult riskilembesed inimesed ja vaatame, kas riskilembesed suitsetajad elavad lühemat aega kui riskilembesed mittesuitsetajad

Segavate faktorite kontrollimine statistiliselt

$$Eluiga = c_0 + c_1\ suitsetab + c_2\ riskilembus + e$$

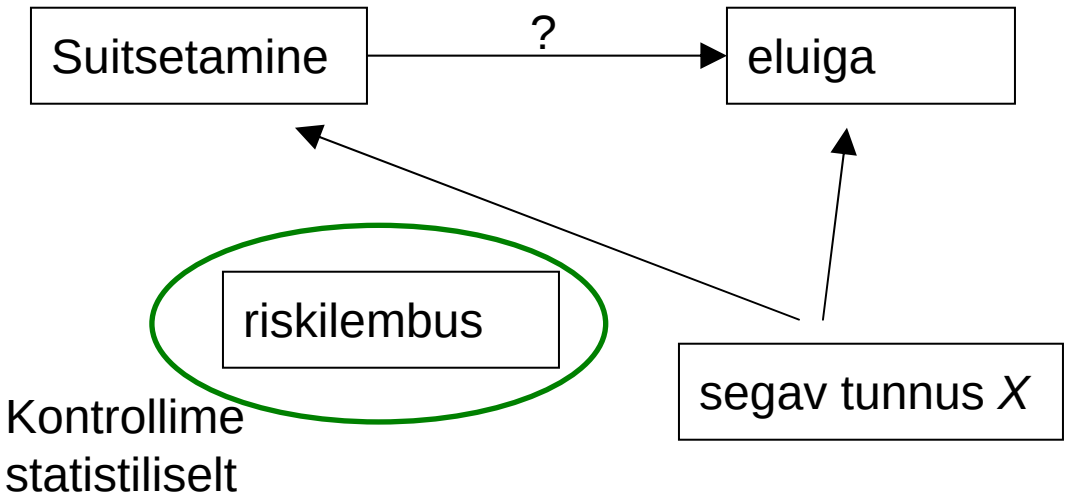
või

$$Eluiga = c_0 + c_1\ suitsetab + c_2\ log(riskilembus) + e$$

või

$$Eluiga = c_0 + c_1\ suitsetab + c_2\ riskilembus^2 + e$$

Suitsetamise ja kopsuvähi vaheline seos



Hoiatus!

Kui tahame uurida põhjusliku mõju olemasolu tunnuste *X* ja *Y* vahel, siis ei tohi mitte iga tunnust, mis on tunnustega *X* ja *Y* korreleeritud, segavaks faktoriks kuulutada!

Tahame uurida, kas vanemate nahavärvi vahel on seos. Oletame hüpoteetiliselt, et inimesed lähevad paari täiesti sõltumata teineteise nahavärvist. Meie aga üritame segava faktorina arvesse võtta lapse nahavärvi. Mis juhtub?

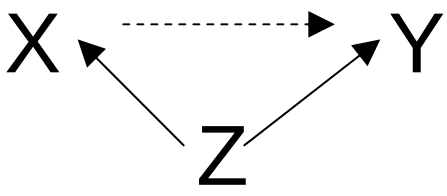
Teades lapse nahavärvi (mulatt), muutub ema nahavärvi teadmine vägagi oluliseks papa nahavärvi ennustamisel

laps: mulatt, ema: valge -> papa on mustanahaline;

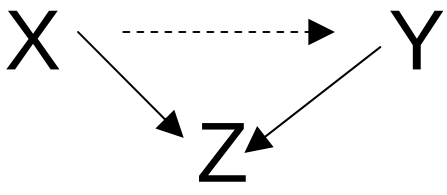
laps: mulatt, ema: mustanahaline -> papa on valge.

Seega võttes mudelis arvesse lapse nahavärvi saame me kunstliku tugeva sõltuvuse vanemate nahavärvide vahele.

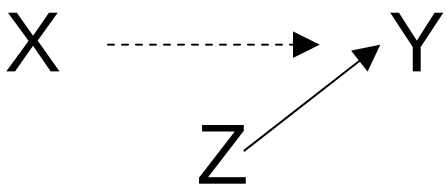
Milliste tunnuste järgi kontrollida?



Lisa mudelisse Z!



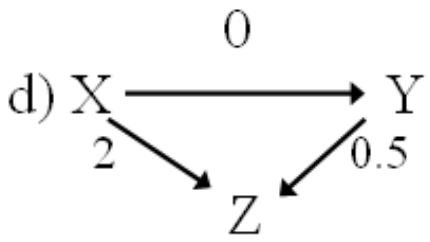
Ära lisa mudelisse Zi!



Sinu valik (pigem lisa)!

Näiteid vigadest

Skeem ja tegelikud kordajate väärtused



Andmete tekitamine:

```
> x=rnorm(100)
> y=10+rnorm(100)+0*x
> z=100+2*x-0.5*y
```

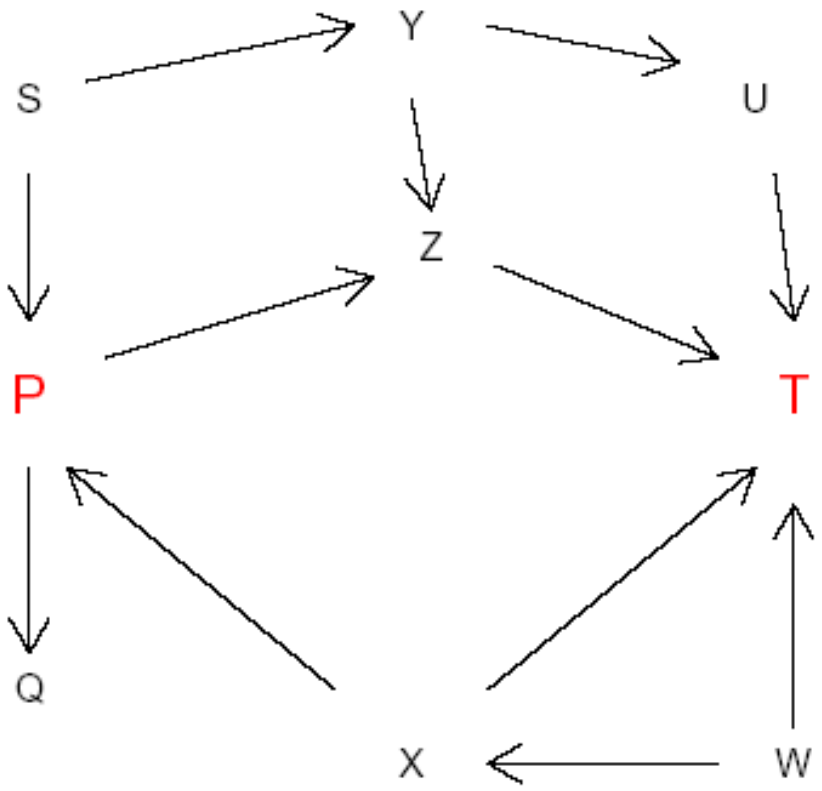
Vigane analüüs:

```
> summary(lm( y ~ x + z))

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  2.000e+02  3.350e-13  5.971e+14   <2e-16 ***
x            4.000e+00  7.288e-15  5.488e+14   <2e-16 ***
z           -2.000e+00  3.528e-15 -5.669e+14   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

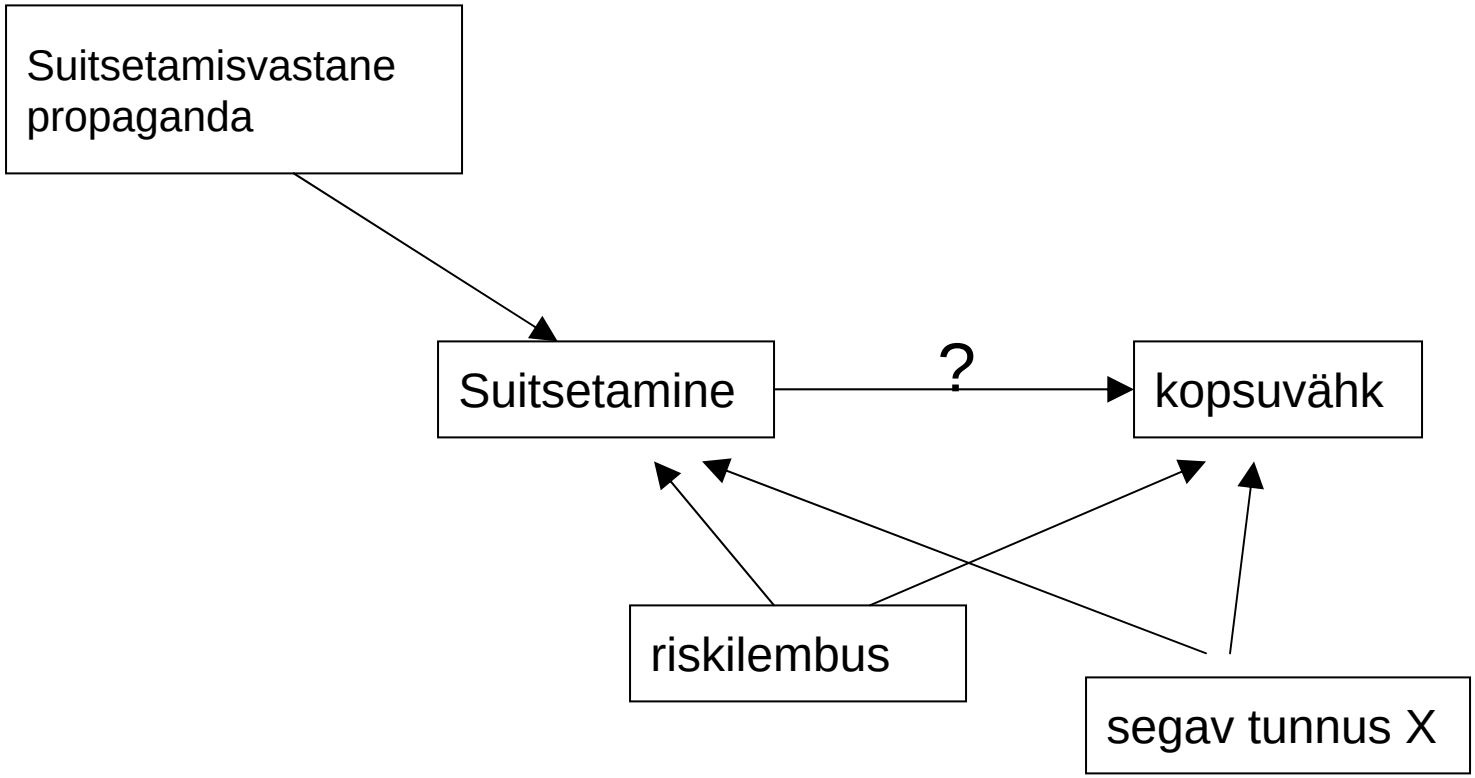
Korrektne analüüs:

```
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.10487    0.09263   109.1   <2e-16 ***
x           -0.02787    0.09301    -0.3    0.765
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Kõrvalepõige: Milliseid tunnuseid tuleks mudelisse lisada ehk graafi moraliseerimine



Suitsetamise ja kopsuvähi vaheline seos – “instrumental variable”

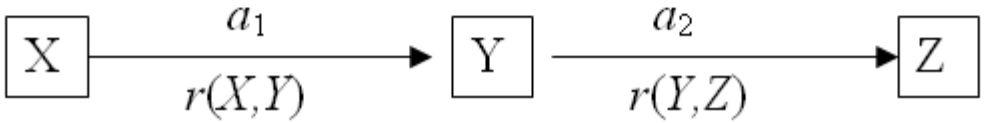


Lihtsaid reegleid juhu tarvis, kui mõjud on lineaarsed I

Andmete tekkimine:

$Y = a_1 X + e_y$; $\quad Z = a_2\, Y + e_z$

$r(X,Y) = a_1$ SQRT(DX / DY) $\quad r(Y,Z) = a_2$ SQRT(DY / DZ)



$r(X,Z) = r(X,Y)\, r(Y,Z)$

$Z = a_1 a_2 X + e$ $\quad$ *võõi* $\quad Z = a_2\, Y + 0{*}X + e$

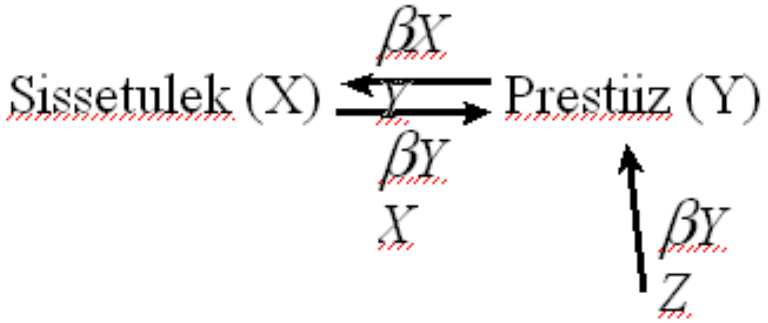
Cov(X,Z) = Cov(X,Y) a_2

Parim põhjuslik mudel Y-le: $Y = c + a_1 X + e$

parim prognoosiv mudel Y-le: $Y = c + a_1\,(1\text{-}v)X + 1/a_2\; v\; Z + e$,

kus v = D(e_y)/(D(e_y)+ D(e_z)/ a_2^2)

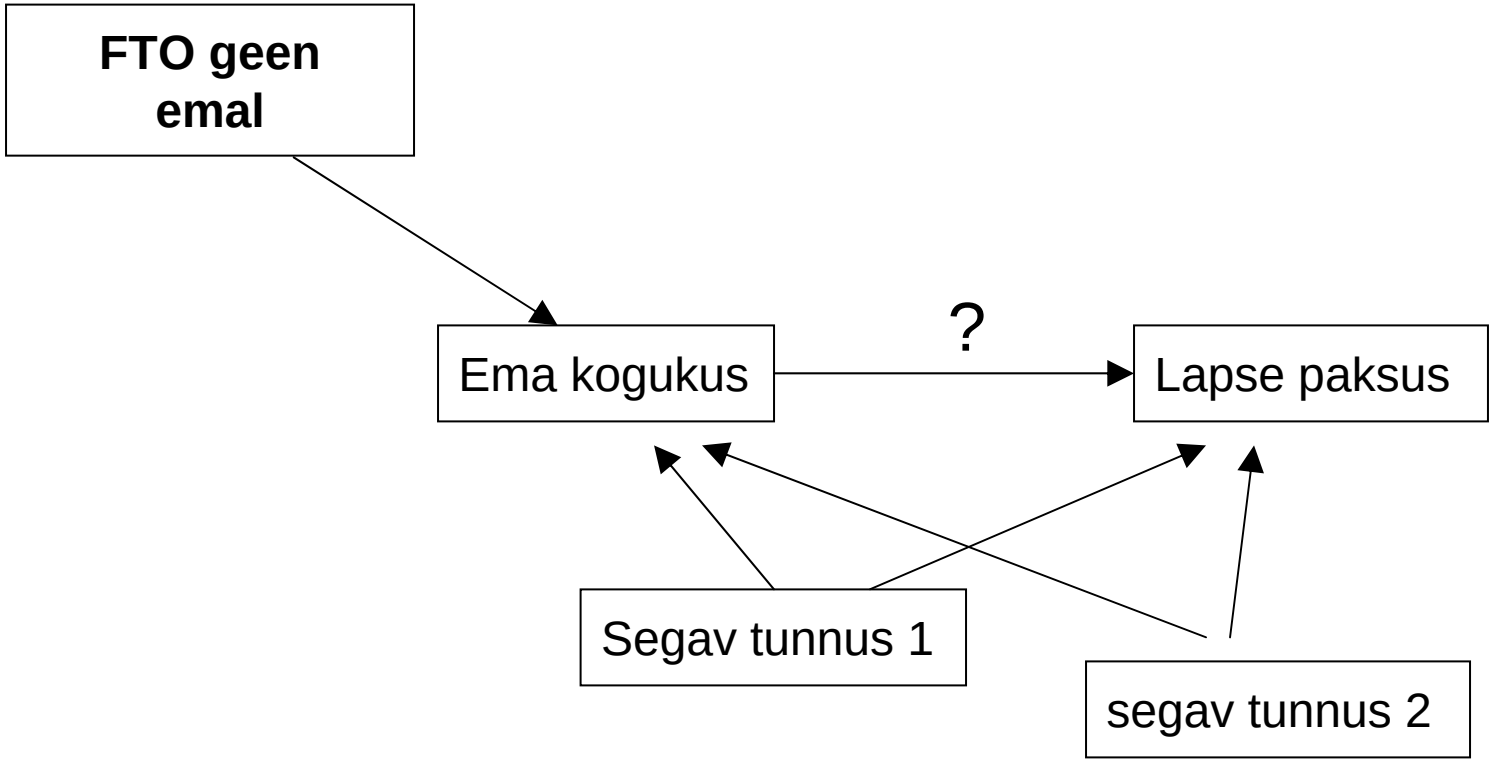
Regressioonikordajate leidmine, kui kovariatsioonimaatriks on teada



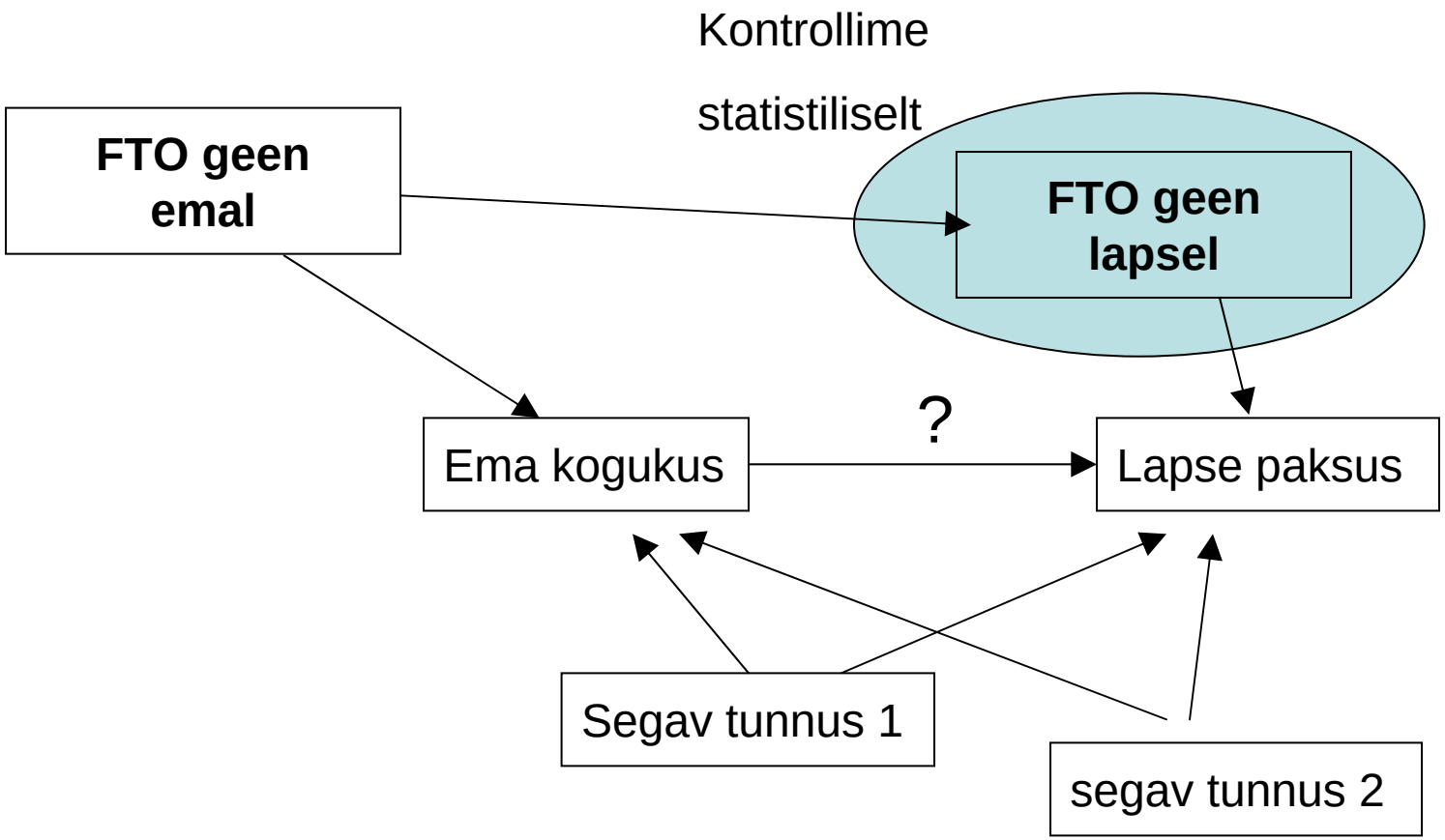
$$cov\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} DX & & \\ \beta_{XY}DY+\beta_{YX}DX & DY & \\ \beta_{XY}\beta_{YZ}DZ & \beta_{YZ}DZ & DZ \end{bmatrix}$$

Võrdsustame vaadeldud ja teediagrammi poolt eeldatad kovariatsioonimaatriksid ja lahendame saadud võrrandisüsteemi.

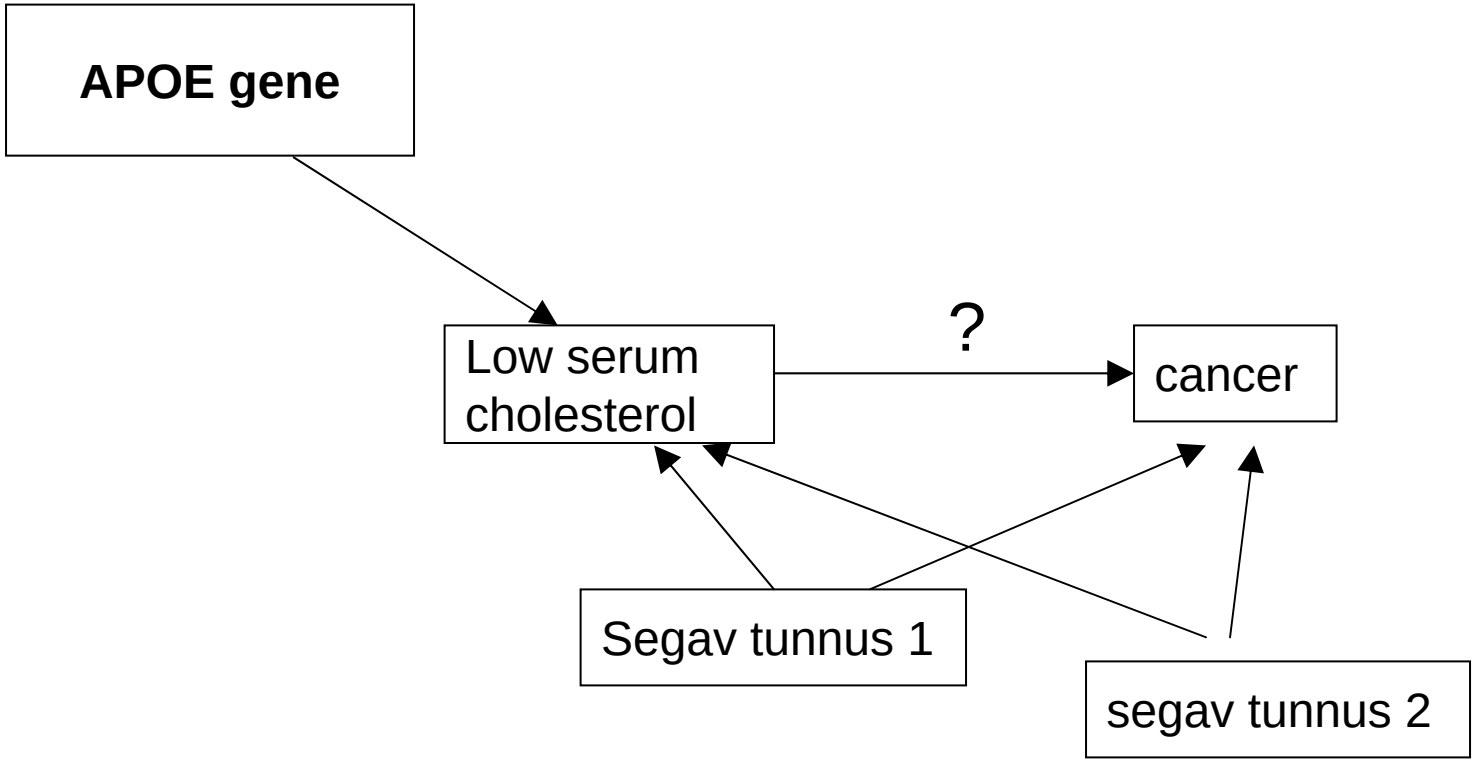
Mendelian randomization



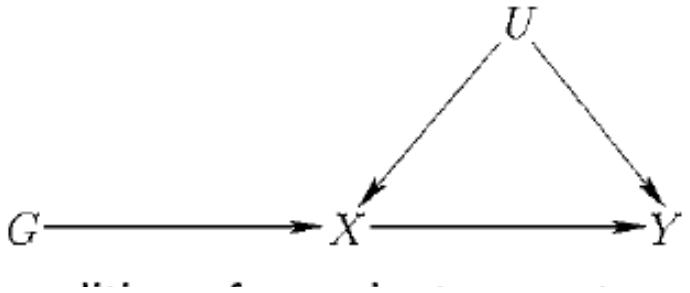
Mendelian randomization I

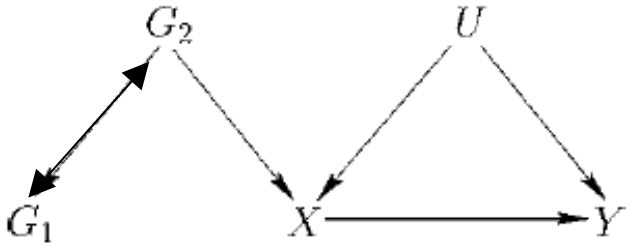


Mendelian randomization II

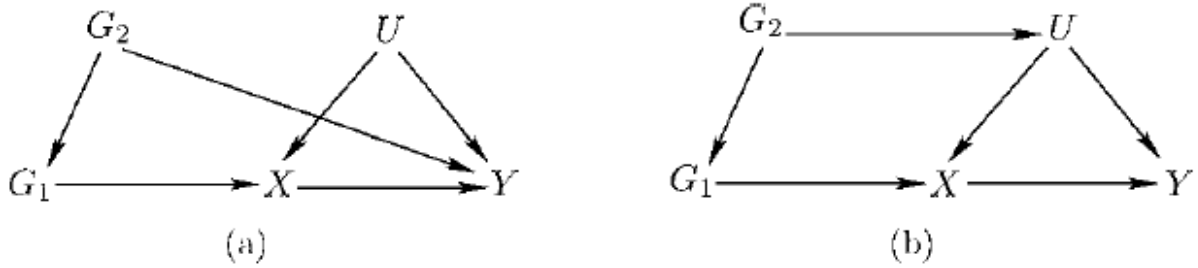


Genotype is IV variable, if

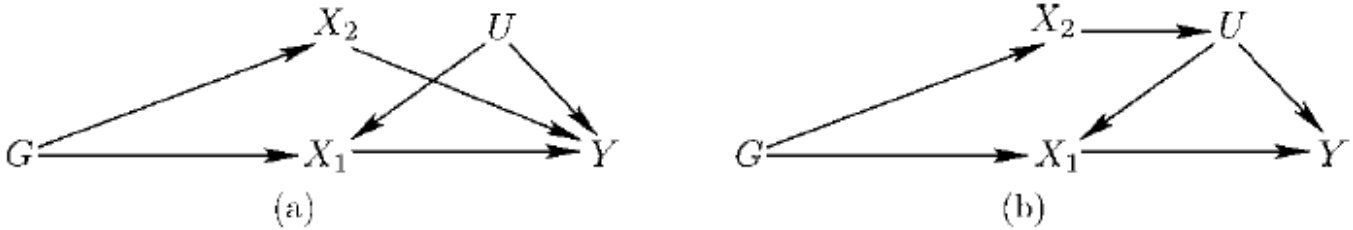




Limitations of Mendelian randomization



Linkage disequilibrium in a Mendelian randomization application.



Pleiotropy in a Mendelian randomization application.

Additional problematic situations

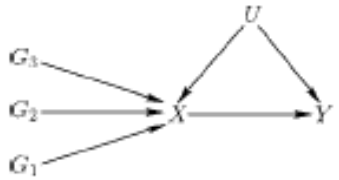


Figure 7 Genetic heterogeneity in a Mendelian randomization application.

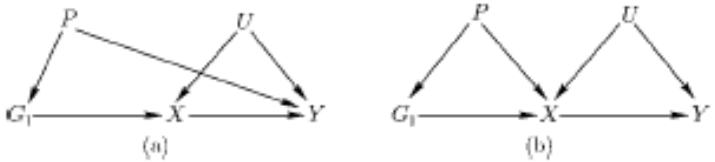


Figure 8 Two examples of population stratification for Mendelian randomization.

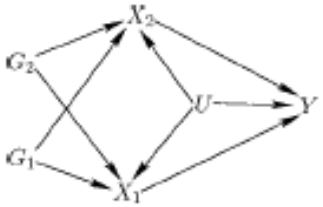


Figure 9 A more complicated example for gene–phenotype–disease relations.