Universitat de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

3. Cause
3.1.- DAGs

Medical Statistics

José Antonio González y Erik Cobo

Abril 2015

Outline

1. DAG: Directed acyclic graph
2. Common causes imply association
3. What do common effect imply?
4. Over-adjustment

1.- DAGs: Directed Acyclic Graphs

Diagrams to represent causal structures.

Directed: arrows have origin and end.
Acyclic: a variable cannot cause it self
Graphs: Diagrams

DAGs

$$\begin{tikzcd} \mathbf{Z} \arrow[r] \arrow[rr, bend left] & \mathbf{X} \arrow[r] & \mathbf{Y} \end{tikzcd}$$

Example (BP=Blood pressure):

Z: Baseline BP **X**: Treatment BP **Y**: Outcome BP

a.- Common causes imply association

$$\begin{tikzcd} \mathbf{Z} \arrow[r] \arrow[rr, bend left] & \mathbf{Z_2} & \mathbf{Y} \end{tikzcd}$$

Example:

Z: Smoking $\mathbf{Z_2}$: Yellow fingers **Y**: Lung cancer

Causal words: $\mathbf{Z_2}$ has no effect on **Y**
Associational words: $\mathbf{Z_2}$ and **Y** are associated

CONFOUNDING

Med. Stat. Cause: DAGs

Conditional independence

$$\begin{tikzcd} \boxed{Z} \arrow[r] \arrow[rr, bend left] & Z_2 & Y \end{tikzcd}$$

Example:

Z: Smoking Z_2 : Yellow fingers **Y**: Lung cancer

Causal words: Z_2 has no effect on **Y**

Associational words: Z_2 and **Y** are NOT associated conditional on **Z**

Med. Stat. Cause: DAGs

Example: High Lipids (Z) cause bad events (Y) such stroke, infarct,..

Gen (X) causes High Lipids (Z) Temporal order: X, Z, Y

Hypothesis: gen (X) causes bad events (Y) independently of High Lipids (Z)

$$\begin{tikzcd} X \arrow[r] \arrow[rr, bend left] & Z \arrow[r] & Y \end{tikzcd}$$

Without adjusting by lipids (Z) we see overall effects of X on Y

$$\begin{tikzcd} X \arrow[r] \arrow[rr, bend left] & \boxed{Z} \arrow[r] & Y \end{tikzcd}$$

Adjusting, we estimate direct effect, accounting by lipids

For example: can we eliminate all the gen effects by controlling lipids?

¿Que pasa con el gen, lipidemias e ictus?

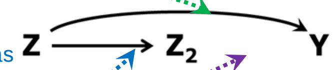
Lipid +	AVC+	AVC-	Lipid -	AVC+	AVC-	Todos	AVC+	AVC-
Gen +	503	185	Gen +	37	151	Gen +	540	336
Gen -	155	56	Gen -	83	337	Gen -	234	393
or = 1'01		$IC_{95\%}$ 0'7-1'4	or = 0'99		$IC_{95\%}$: 0'6-1'5	or = 2'7		$IC_{95\%}$:2,2-3,3

Full dichotomist example

Z_2+	Y+	Y-	Z_2-	Y+	Y-		Y+	Y-
Z+	80	40	Z+	40	20	Z+	120	60
Z-	20	40	Z-	40	80	Z-	60	120
OR = 4			OR = 4			OR = 4		
Y+	Z_2+	Z_2	Y-	Z_2+	Z_2-		Z_2+	Z_2-
Z+	80	40	Z+	40	20	Z+	120	60
Z-	20	40	Z-	40	80	Z-	60	120
OR = 4			OR = 4			OR = 4		
Z+	Y+	Y-	Z-	Y+	Y-		Y+	Y-
Z_2+	80	40	Z_2+	20	40	Z_2+	100	80
Z_2-	40	20	Z_2-	40	80	Z_2-	80	100
OR = 1			OR = 1			OR = 1'563		

Z has effect on **Y**

Z has effect on **Z_2**



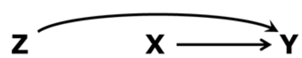
Z_2 has no effect on **Y**

Conditional Independence

Z_2 and **Y** are associated

CONFOUNDING

Full numeric example

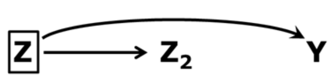


adjusted effect (-10) =

= raw effects (-10)

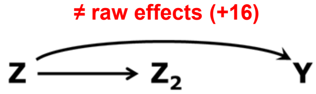
NO confounding

Mean (n)	Treated		Control		OR_{ZX} = 10·10 / 10·10 = 1
Z+	**110**	(10)	**120**	(10)	**115** (20)
Z-	**130**	(10)	**140**	(10)	**135** (20)
Total	**120**	(20)	**130**	(20)	125 (40), SD=10



adjusted effect (0) ≠

≠ raw effects (+16)



Collinearity &
Confounding!

Mean (n)	Treated		Control		OR_{ZZ_2} = 2·2 / 18·18 = 1/81
Z+	**110**	(2)	**110**	(18)	**110** (20)
Z-	**130**	(18)	**130**	(2)	**130** (20)
Total	**128**	(20)	**112**	(20)	120 (40), SD=10

Med. Stat. Cause: DAGs

Collinearity: the best friend of statisticians

If there is no collinearity, then no confusion in data, then less stat consultancy!

You have to be able to draw an imaginary data set with collinearity.

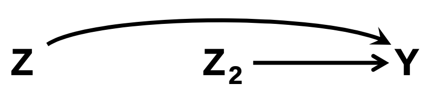
Both for a dichotomous and a continuous response

Z1	Y+	Y-	Z2	Y+	Y-		Y+	Y-
X+			X+	100		X+		
X-			X-			X-		

Mean (n)	X+	X-	All
Z1	20 (30)	30 (10)	
Z2	10 (10)	20 (30)	
All	17.5	22.5	

Med. Stat. Cause: DAGs

b. What do common effects imply?



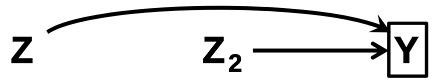
Example:

Z: Genetics **Z_2**: Environment **Y**: Lung cancer

Causal words: **Z** has no effect on **Z_2**

Associational words: **Z** and **Z_2** are independent

Conditioning on common effects



Examples:

Z: Genetics   **Z_2**: Environment   **Y**: Lung cancer

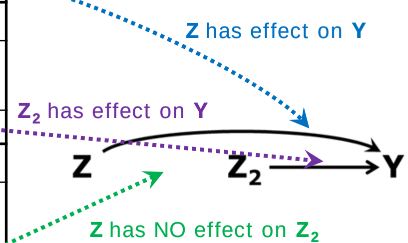
Causal words: **Z** has no effect on **Z_2**

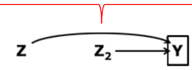
Associational words: **Z** and **Z_2** are associated conditional on **Y**

Selection BIAS

Conditioning on common effects

Z_2+	Y+	Y-	Z_2-	Y+	Y-		Y+	Y-
Z+	64	16	Z+	40	40	Z+	104	56
Z-	40	40	Z-	16	64	Z-	56	104
OR = 4			OR = 4			OR ≈ 3.45		
Z+	Y+	Y-	Z-	Y+	Y-		Y+	Y-
Z_2+	64	16	Z_2+	40	40	Z_2+	104	56
Z_2-	40	40	Z_2-	16	64	Z_2-	56	104
OR = 4			OR = 4			OR ≈ 3.45		
Y+	Z_2+	Z_2	Y-	Z_2+	Z_2-		Z_2+	Z_2-
Z+	64	40	Z+	16	40	Z+	80	80
Z-	40	16	Z-	40	64	Z-	80	80
OR = .64			OR = .64			OR = 1		





Z and Z_2 are associated conditional on Y: SELECTION BIAS

Med. Stat. Cause: DAGs

Example: Genetics and high lipids (new also invented data)

High Lipids (Z) cause bad events (Y) such stroke, infarct,..
Also gen (X) causes bad events (Y)

Temporal order: X, Z, Y

Imagine that we condition on Y. For example, if we only look at hospital data which has only patients with high bad events (Y+):

Y+	Z+	Z-	Y-	Z+	Z-
X+	80	45	X+	10	45
X-	45	10	X-	45	80
OR=0,4	$CI_{95\%}$=0,18 to 0,86		OR=0,4	$CI_{95\%}$=0,18 to 0,86	

	Z+	Z-
X+	90	90
X-	90	90
OR=1	$CI_{95\%}$=2/3 to 3/2	

Within (Y+) and outside (Y-) hospital,

gen (X) prevents lipids (Z)

Overall (previous in time to the occurrence of outcome Y), gen (X) is independent to lipids (Z)

Med. Stat. Cause: DAGs

INVENTING DATA for selection bias

	Z+	Z-
X+	100	100
X-	100	100
OR=1		

Y+	Z+	Z-	Y-	Z+	Z-
X+	50	50	X+	50	50
X-	50	50	X-	50	50
OR=1			OR=1		

Building relation between Z (columns) and Y (subtable)

Y+	Z+	Z-	Y-	Z+	Z-
X+	60	40	X+	40	60
X-	60	40	X-	40	60
OR=1			OR=1		

Building relation between X (rows) and Y (subtable)

Y+	Z+	Z-	Y-	Z+	Z-
X+	80	60	X+	20	40
X-	40	20	X-	60	80
OR= ½			OR= ½		

SELECTION BIAS

X has NO effect on **Z**

	Z+	Z-
X+	100	100
X-	100	100
OR=1		

Med. Stat. Cause: DAGs

Z_2+	Y+	Y-	Z_2-	Y+	Y-		Y+	Y-
Z+			Z+			Z+		
Z-			Z-			Z-		
OR =__			OR =__			OR ≈___		
Z+	Y+	Y-	Z-	Y+	Y-		Y+	Y-
Z_2+			Z_2+			Z_2+		
Z_2-			Z_2-			Z_2-		
OR =__			OR = __			OR ≈ ___		
Y+	Z_2+	Z_2	Y-	Z_2+	Z_2-		Z_2+	Z_2-
Z+			Z+			Z+		
Z-			Z-			Z-		
OR =___			OR = ___			OR = **1**		

Conditioning on common effects:
A) YOUR OWN DATA

$$\begin{tikzcd} Z \arrow[rr, bend left] & Z_2 \arrow[r] & Y \end{tikzcd}$$

B) YOUR OWN VARIABLES: Propose an example (variables, conditions) were selection bias may be devisable.

Med. Stat. Cause: DAGs

Another type of over-adjustment

Conditioning on effects originate selection bias.

But conditioning on causes/conditions may difficult interpretation:

effect of X on Y at a fixed level of the remaining Z variables

has to be possible and to make sense.

Counter-examples:

1) SBP and DBP: High relationship between both:

Higher is one of them, higher is expected the other.

If one BP is “fixed”, the other’ variability decreases.

Can we change one without modifying the other?

2) Idem for tobacco and alcohol

3) Risk of newborn with Down depending on the age of mother and father

Can we act on the mother’s age without modifying the father’s one?

Is the objective intervention or prediction?