Epidemiology and the causal enquire: the role of statistics

Milton Severo

November 25, 2019

1 Introduction

Data scientist is the new fashion and a statistician is out of fashion. A data scientist focus on data clean, data transformation, exploratory analysis, choosing prediction algorithms and the respective predictions quality. A statistician focus on the selection of the sample, p-values, confidences intervals and the assumptions of the models to do inference [Gutierrez, 2019].

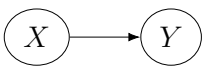
Independently of this differences the only important matter is to make good science. Data scientist wants to give good predictions, usually means that there is a strong correlation/association. In the case of a statistician good inference usually means study the association/correlation between variables. However, in both cases "correlation does not equal causation".

For example, there is a correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel prizes per 10 Million Population. This means that eating chocolate causes winning Nobel prizes or means that there is a common cause to both things, e.g., being a rich country causes at same time eating more chocolate and winning Nobel prizes [Maurage et al., 2013].

The focus on health and in particular on epidemiology is causation and not if there is good predictions or strong associations. We can have excellent predictions with none of the relations between the variables being causal. Thus, data scientist/statistician should not focus only in the prediction or in the model but if there is causation. A researcher should know how to to formalize and communicate causal questions and assumptions? The main tool to do that is causal diagrams, **direct acyclic graphs** (DAGs). DAGs are considered a useful tool for causal inference. Helpful for identifying which variables to control and to make assumptions explicit. This paper is divides in seven sections: introduction, causal diagrams language, probability link, association and causation, blocking paths, rules for d-separation and discussion.

2 Causal Diagrams Language

Graphs can be used to formalize causal inference [Spirtes et al., 2000]. This is a **direct graph**, which shows that X affects Y.



This is a **undirect graph**, which shows that X and Y are associated with each other.

A DAG can not have undirect paths and no cycles.

1. **Paths** are sequence of lines (edges) between two variables, regardless of the direction of arrows; A path is way to get from one vertex to another, travelling along edges (regardless of the direction of the edges); In figure 1 there are two paths from W to B, $W > Z > B$ and $W > Z > A > B$ and there is one path from Z to W ($Z < W$).

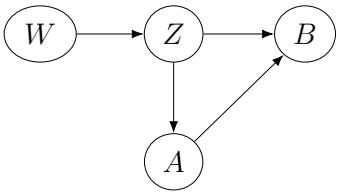


Figure 1: Example 1 of a causal graph

2. **Descendants** are the direct or indirect effects of a variable; In figure 1 Z is a descendant of W, W is Z parent, , W is ancestor of B or B has two parents (but we can have more than two). Consequently a family tree is a DAG!

3. **Colliders** Common effect of two variables in a path: where the arrows 'collide'. Any variable on a path that is not a collider is a "non-collider". In figure 1, B is a collider because is common effect of Z and A.

In a DAG, all common causes of two or more variables in the diagram have to be explicit, regardless of whether or not they are observed. The diagram should be parsimonious - causes of only one of the node (variables) should not be included. Unknown or unmeasured causes can and should be represented.

3 Probability

Where is the probability in DAGs? DAGs encodes assumptions about dependencies between variables. A DAG will tell us: which variables are independent from each other, which variables are conditionally independent from each other, ways that we can factor and simplify the joint distribution.
For a given graph and vertex W let **Parents(W)** be a set of parents of W, and **Descendants(W)** be a set of descendants of W.

Markov Condition: A direct acyclic graph G over **V** and a probability distribution $P(V)$ satisfy the Markov condition if and only if for every W in **V**, W is independent of $V \setminus \{\textbf{Descendants(W)} \cup \textbf{Parents(W)}\}$ given **Parents(W)**.

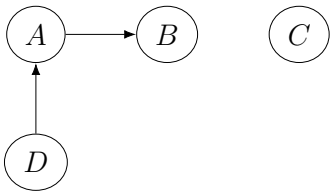


Figure 2: Example 2 of a causal graph

The DAG of figure 2 implies that:

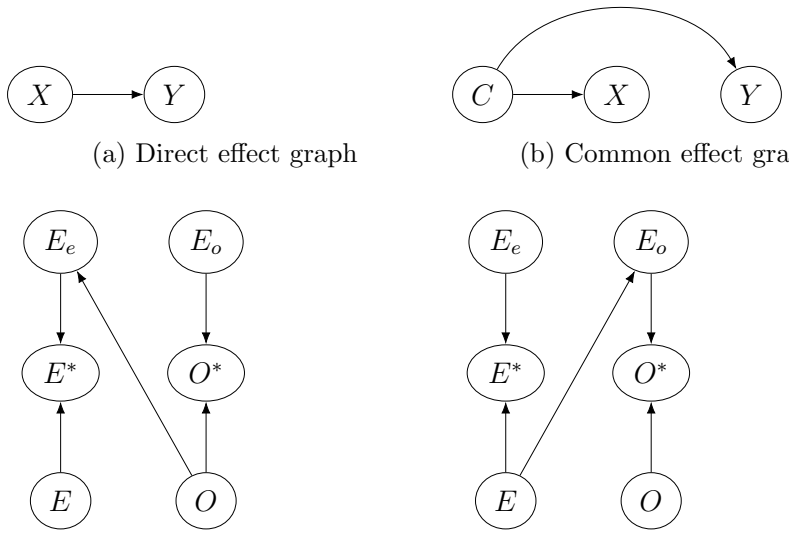
1. $P(C \mid A, B, D) = P(C)$;
2. $P(B \mid A, C, D) = P(B \mid A)$;
3. $P(B \mid D) \neq P(B)$;
4. $P(D \mid A, C, D) = P(D \mid A)$.

We can decompose the joint distribution by sequential conditioning only on sets of parents. Start with roots (nodes with no parents). Proceed down the descendant line, always conditioning on parents. Thus for the DAG of figure 2 implies that:

$$P(A, B, C, D) = P(C)P(D)P(A \mid D)P(B \mid A) \tag{1}$$

4 Association and causation

In which circumstances X and Y are statistically associated? All this DAGs implies association between X and Y.



(a) Direct effect graph

(b) Common effect graph

(c) Measurement error graph

(d) Measurement error graph

Figure 3: Four simple graphs

For graph 3a, X is parent so according to the Markov Condition they are not independent. Also graph 3b is not independent because according to the Markov Condition only $Y \perp\!\!\!\perp X \mid C$. We call **backdoor path** from treatment X to outcome Y if there are paths between X to Y trough arrows into X. Here, $X \leftarrow C \rightarrow Y$ is backdoor path from X to Y. Backdoor path confounded the relationship between X and Y, so this is not a causal effect. The path $X \leftarrow C \rightarrow Y$ is an unblocked path between X and Y - unless we condition on C (e.g. restrict, adjust, stratify). An observed statistical association between X and Y can be due to being a cause of Y or C being a common cause of X and Y.

Another not causal association is an association due to measurement error [Hernán and Cole, 2009] showed in graphs 3c and 3d. The true exposure (E)

affects both the outcome (O) and the measured exposure (E*). The causal diagram also includes the node E_e to represent all factors other than E that determine the value of E^* and similar case between E_o and O^*. We refer to E_e and E_o as the measurement error for E and O. The assumption implicit in many epidemiology studies is that the association between E^* and O^* approximates the association between E and O. However this is only true if E is not associated with measurement error of O (graph 3c) and the O is associated with measurement error of E (graph 3d). A previous study wanted to study the causal association between knee osteoarthritis (E) with knee Pain (O) [Pereira et al., 2017]. Both measures suffer from measurement error. Two medical doctors looking to the same x-ray can give different results in relation to the presence or not of osteoarthritis or two patients with the same pain level can given different scores in a pain questionnaire. The association found between E^* and O^* may not be causal because if a patient already know that has knee Osteoarthritis may recall and graded more pain (graph 3c). If a medical doctor already know that the patient has knee pain will more falsely detect the presence of knee osteoarthritis (graph 3d). Therefore is possible to have an association that is not really causal.

5 Blocking paths

Paths can be blocked by conditioning on variables (vertices) in the path. Consider the the following path:

Figure 4: Example of path with a mediator

If we condition on the mediator, M, (a node in the middle of chain), we **block** the path from E to O. Within each strata of M there is no association between E and O. E.g., a previous cohort study aimed to assess if maternal socio-economic position at 12 years of age (E) influences Child dietary patterns at 4 years of age (O) via-socio and demographic characteristics at child's delivery (M). The study found that E was positive associated with the O, higher socio-economic position at 12 years was associated with healthier dietary patten, however when maternal socio-economic and demographic characteristics at child's delivery (M) were added to the model, maternal

socio-economic position at 12 years ceased to be associated with children's dietary pattern [Durão et al., 2017].

Consider the following path:

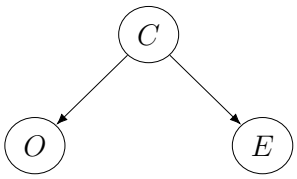


Figure 5: Example of path with a confounder

If we condition on confounder (C), the path from exposition (E) to the outcome (O) is blocked. E.g., a previous cohort study founded a crude association between occupational physical activity (E) and hypertension incidence (O), however after adjusting for age (C) the effect disappear. This showed that the crude association is not causal but due to age being a common cause of occupational physical activity and hypertension incidence [Camões et al., 2010].

Consider the following path:

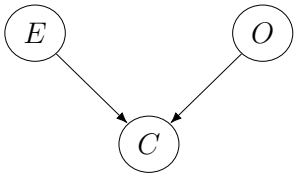
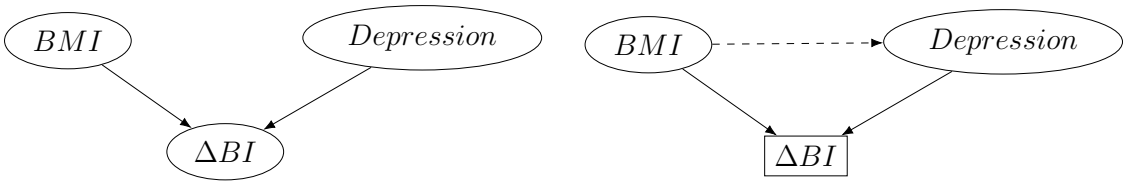


Figure 6: Example of path with a collider

If a **collider** is conditioned on, the opposite situation occurs comparing to the confounder example. In this path, E and O are not associated via this path. However, conditioning on C induces an association between E and O. Opens a door between E and O. This is call selection bias.
For example, a study had tried to explained the relationship between body mass index (BMI), depression and discrepancy between perceived and desired body image (ΔBI). [Almeida et al., 2012]. The study showed that there is no crude association between depression (E) and BMI (O), however BMI and depression has a positive association with discrepancy (C) between perceived and desired body image. The DAG is represented in figure 7a. If we adjust/conditional for discrepancy between perceived and desired body image

there will be a negative association between BMI and Depression. The DAG is represented in figure 7b. Why this? Among those individuals with high discrepancy, they have high probability of having BMI or high probability of depression (or both). Put it in another way, restricting our attention only to those individuals with high discrepancy or conditioning by (stratifying by) the variable discrepancy, we have conditioned on a 'common effect' of both BMI and Depression. Therefore, knowing that you have a high discrepancy, if we measure that you have low BMI, then we know that your high discrepancy is due to having depression; inversely, if we measure that you have no depression, then we know that your high discrepancy is due to high BMI. Conversely, all individuals with low discrepancy were exposed to neither high BMI nor high depression or at least one need to be very low to compensate a average value of the other.



(a) DAG for BMI, ΔBI and depression

(b) DAG for BMI and depression, after conditionally on ΔBI

Figure 7: DAG for BMI and depression and ΔBI

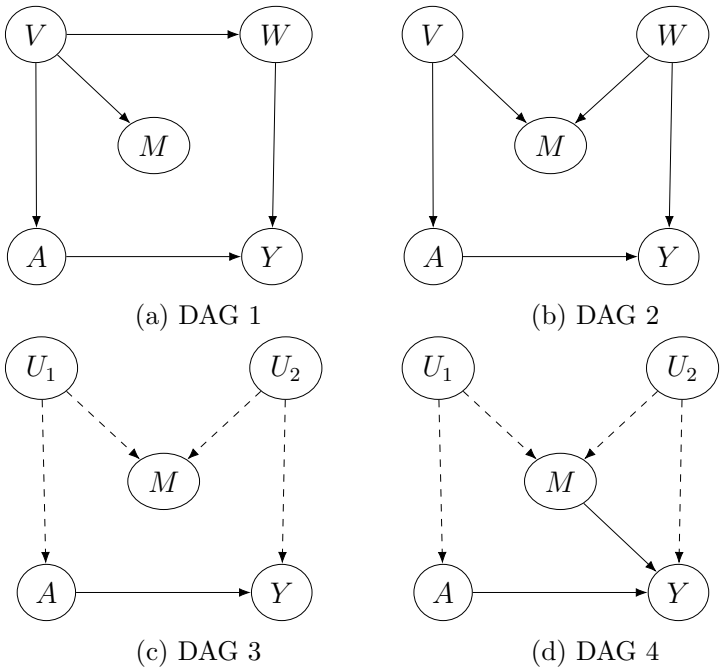
6 Rules for d-separation

The d-separation rules will allow the identification of the set of variables that is need to adjust to evaluate the direct effect of variable on another variable. A path is **d-separated** by the set of variables **C** if:

1. It contains a chain $D \rightarrow E \rightarrow F$ and the middle part is in C;
2. It contains a fork $D \leftarrow E \rightarrow F$ and the middle part is in C;
3. It contains an inverted fork ($D \rightarrow E \leftarrow F$) and the middle part is not in C, nor any descendants of it.

The d-separation rules can be applied to define the adjustment in statistical model to identify causality (Criterion 1). However there are other criteria to define the adjustment in a statistical model. The usual criterion for adjustment for statistical modelling is **adjusting for all variables** (Criterion 2). Another criterion is the **disjunctive cause criterion** [VanderWeele and Shpitser, 2011] (Criterion 3) that adjusts for all observed variables that causes the exposure, outcome or both.

Considering the following four DAGs.



(a) DAG 1 (b) DAG 2

(c) DAG 3 (d) DAG 4

Figure 8: Measurement error bias DAGs

Imagine that in the DAGs of figure 8 the researcher wants to assess the direct effect (causal) between A and Y, the variables V, W and M are observed variables and U_1 and U_2 are unobserved variable. Applying criterion 1 to DAG 1 of figure 8a we need to adjust for V or W or both, because the middle part of a fork should be in the adjustment; for DAG 2 and 3 of figures 8b and 8c we don't need to adjust for any variable because it contains an inverted fork and the middle part should not be in the adjustment. In DAG 4

of figure 8d we have two problems: from one side we have to adjust for $\{M\}$ because we have a fork $A < U1 > M > Y$; however, from the other side, the variable M is part of inverted fork so we should not adjust. In conclusion, in DAG 4 we can't obtain an unbiased association because by one-side the association is confounded and from the other side suffers from selection bias. To apply criterion 2 to DAG 1 and 2 of figures 8a and 8b we need to adjust for $\{V, W, M\}$ and this would be fine. In DAG 2 this adjustment would open a path between V and W by adjusting for M, however, by adjusting for the other variables this open path would be block. In DAG 3 and DAG 4 of figures 8c and 8d we would only adjust for M because the other variables $\{U_1, U_2\}$ are unobserved. This would open a path between U_1 and U_2, and consequently this would mean that the association between A and Y would be confounded because we couldn't block. Applying criterion 3 we would adjust for $\{V, W\}$ on DAG 1 and 2, on DAG 3 we would not adjust for any variable and for DAG 4 we would be adjusting for $\{M\}$. In the first three DAGs the adjustment set would be correct however in DAG 4 this would be incorrect because there would be an open path between U_1 and U_2.
The dagitty package from R can identify all this sets of adjustments as showed in following example [Textor et al., 2016].

```
> library(dagitty)
> #DAG 1
> g <- dagitty("dag{ v->{w m a} {a w}->y
+              a[exposure]
+              y[outcome]
+              }") # m-bias graph
> adjustmentSets(g)

 { w }
 { v }

> #DAG 2
> g <- dagitty("dag{ v->{m a} w->m {a w}->y
+              a[exposure]
+              y[outcome]
+              }") # m-bias graph
> adjustmentSets(g)

 {}

> #DAG 3
> g <- dagitty("dag{ u1->{m a} u2->m {a u2}->y
+              a[exposure]
+              y[outcome]
+              u1[unobserved]
+              u2[unobserved]
+              }") # m-bias graph
> adjustmentSets(g)
```

```
 {}

> #DAG 4
> g <- dagitty("dag{ u1->{m a} u2->m {a m u2}->y
+              a[exposure]
+              y[outcome]
+              u1[unobserved]
+              u2[unobserved]
+              }") # m-bias graph
> adjustmentSets(g)
> #none was founded
```

To explained more clear all this aspects we simulated DAG 2, 3 and 4 and adjust using linear regression model in R. As observed in this simulation when adjust for the correct set of variables a valid association is found if not adjust for the correct set of variables a bias association is found.

```
> set.seed(100)
> V<-U1<-rnorm(10000)
> W<-U2<-rnorm(10000)
> A<-2*U1+rnorm(10000)
> M<-3*U1+2*U2+rnorm(10000)
> Y<-A+U2+rnorm(10000)
> summary(lm(Y~A))       #DAG 2 criterion 1

Call:
lm(formula = Y ~ A)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4210 -0.9436  0.0051  0.9638  5.5306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001088   0.014199   0.077    0.939
A           0.995497   0.006372 156.233   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.42 on 9998 degrees of freedom
Multiple R-squared:  0.7094,        Adjusted R-squared:  0.7094
F-statistic: 2.441e+04 on 1 and 9998 DF,  p-value: < 2.2e-16

> summary(lm(Y~A+M+V+W)) #DAG 2 criterion 2

Call:
lm(formula = Y ~ A + M + V + W)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4344 -0.6766 -0.0142  0.6783  3.6164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 0.012971   0.010002   1.297    0.195
A           0.985226   0.009921  99.308   <2e-16 ***
M           0.001659   0.010077   0.165    0.869
V           0.024589   0.037668   0.653    0.514
W           1.005315   0.022565  44.553   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 9995 degrees of freedom
Multiple R-squared:  0.8559,        Adjusted R-squared:  0.8558
F-statistic: 1.484e+04 on 4 and 9995 DF,  p-value: < 2.2e-16

> summary(lm(Y~A+V+W)) #DAG 2 criteion 3

Call:
lm(formula = Y ~ A + V + W)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4350 -0.6767 -0.0142  0.6792  3.6170

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01297    0.01000   1.296    0.195
A            0.98522    0.00992  99.313   <2e-16 ***
V            0.02960    0.02223   1.332    0.183
W            1.00864    0.01001 100.793   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 9996 degrees of freedom
Multiple R-squared:  0.8559,        Adjusted R-squared:  0.8559
F-statistic: 1.979e+04 on 3 and 9996 DF,  p-value: < 2.2e-16

> summary(lm(Y~A))  #DAG 3 applying criterion 1 and 3

Call:
lm(formula = Y ~ A)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4210 -0.9436  0.0051  0.9638  5.5306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001088   0.014199   0.077    0.939
A           0.995497   0.006372 156.233   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.42 on 9998 degrees of freedom
Multiple R-squared:  0.7094,        Adjusted R-squared:  0.7094
F-statistic: 2.441e+04 on 1 and 9998 DF,  p-value: < 2.2e-16

> summary(lm(Y~A+M)) #DAG 3 applying criterion 2
```

```
Call:
lm(formula = Y ~ A + M)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4482 -0.7924  0.0017  0.8015  4.3060

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.007383   0.011869   0.622    0.534
A           0.639115   0.007603  84.061   <2e-16 ***
M           0.297997   0.004537  65.681   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.187 on 9997 degrees of freedom
Multiple R-squared:  0.797,        Adjusted R-squared:  0.797
F-statistic: 1.963e+04 on 2 and 9997 DF,  p-value: < 2.2e-16


> Y<-A+U2+M+rnorm(10000)
> summary(lm(Y~A))  #DAG 4 without adjusting


Call:
lm(formula = Y ~ A)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8845  -2.3743  -0.0744   2.4193  13.7696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03799    0.03589  -1.059     0.29
A            2.19113    0.01610 136.058   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.589 on 9998 degrees of freedom
Multiple R-squared:  0.6493,        Adjusted R-squared:  0.6493
F-statistic: 1.851e+04 on 1 and 9998 DF,  p-value: < 2.2e-16


> summary(lm(Y~A+M)) #DAG 4 applying criterion 2 and 3


Call:
lm(formula = Y ~ A + M)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3380 -0.7990  0.0110  0.7986  5.0323

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.010652   0.011908  -0.894    0.371
A            0.643507   0.007628  84.359   <2e-16 ***
M            1.294077   0.004552 284.285   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.191 on 9997 degrees of freedom
Multiple R-squared:  0.9614,        Adjusted R-squared:  0.9614
F-statistic: 1.245e+05 on 2 and 9997 DF,  p-value: < 2.2e-16
```

7 Discussion

This paper showed several ways that two variables can be associated without one being the cause of the other. The explanations can be: they share common causes, confounding, they share effects that have conditioned on, selection bias, or there is a differential measurement error that will induce a relationship between two variables that was not causal. One important point is that even that the association is causal this can be explain by **reverse causation**. It is not variable X that causes Y is Y that causes X.
Nevertheless, it is possible to identify causation. The more efficient way to identity causation is making a random trial, this will guarantee that there is no common causes between exposure and outcome because the exposure is randomly assign. By applying d-separation rules on observational studies we can guarantee causation. To do so we need to have enough measure variables that allows to block all backdoor paths for the measured and unmeasured confounding variables. For this last step to work well we need to represent all previous knowledge and assumptions using a DAG. The DAG will allow to understand the extent to which observed data are consistent with the causal model proposed by the researcher, to predict expected statistical associations and detect logical problems and contradictions in data analysis as we have shown.
A very usual library from R is 'dagitty' that helps to identify the adjustments we should do or if we can't find adjustment [Textor et al., 2016]. In conclusion, to claim causation a biostatistics needs to accompany all the knowledge on statistics and modelling with good theoretical/practical background on epidemiology and in more specific use tools like a DAG.

References

[Almeida et al., 2012] Almeida, S., Severo, M., Araújo, J., Lopes, C., and Ramos, E. (2012). Body image and depressive symptoms in 13-year-old adolescents. *Journal of paediatrics and child health*, 48(10):E165–E171.

[Camões et al., 2010] Camões, M., Oliveira, A., Pereira, M., Severo, M., and Lopes, C. (2010). Role of physical activity and diet in incidence of hypertension: a population-based study in portuguese adults. *European journal of clinical nutrition*, 64(12):1441.

[Durão et al., 2017] Durão, C., Severo, M., Oliveira, A., Moreira, P., Guerra, A., Barros, H., and Lopes, C. (2017). Association of maternal characteristics and behaviours with 4-year-old children's dietary patterns. *Maternal & child nutrition*, 13(2):e12278.

[Gutierrez, 2019] Gutierrez, D. (2019). Data scientists versus statisticians.

[Hernán and Cole, 2009] Hernán, M. A. and Cole, S. R. (2009). Invited commentary: causal diagrams and measurement bias. *American journal of epidemiology*, 170(8):959–962.

[Maurage et al., 2013] Maurage, P., Heeren, A., and Pesenti, M. (2013). Does chocolate consumption really boost nobel award chances? the peril of over-interpreting correlations in health studies. *The Journal of Nutrition*, 143(6):931–933.

[Pereira et al., 2017] Pereira, D., Severo, M., Ramos, E., Branco, J., Santos, R. A., Costa, L., Lucas, R., and Barros, H. (2017). Potential role of age, sex, body mass index and pain to identify patients with knee osteoarthritis. *International journal of rheumatic diseases*, 20(2):190–198.

[Spirtes et al., 2000] Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, prediction, and search*, chapter Formal Preliminaries. MIT press.

[Textor et al., 2016] Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894.

[VanderWeele and Shpitser, 2011] VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413.