Example 1:

We define the treatment and the outcome as follows:

Treatment : getting a dog W=1, not getting a dog W=0

Outcome : severe depression symptoms Y=1, mild depression symptoms Y=0 measured **after** the treatment assignment W

For the depression/dog example, a potential confounder is the severity of depression symptoms (denoted by X) **before** treatment assignment.

It is reasonable to believe that individuals with severe symptoms of depression pretreatment (X=1) are more likely to adopt a dog (W=1) than people with mild symptoms of depression (X=0).

Furthermore, individuals with severe symptoms of depression before the treatment assignment (X=1) are more likely to have severe symptoms of depression after the treatment assignment Y, than individuals with mild symptoms of depression (X=0).

RCTs are the gold standard study design used to estimate causal effects. To assess the causal effect on survival of getting a new drug compared to a placebo, we could randomize half of the patients enrolled in our study.

Half would receive the dog (W=1), and the other half would not receive the dog (W=0).

Randomization is particularly important to establish the efficacy and safety of drugs (new and existing). This is because randomizing patients eliminates systematic differences between treated and untreated observations. In other words, randomization ensures that these two sets of observations are as similar as possible with respect to all potential confounders, regardless of whether we measure these potential confounders, and are identical on average. If the distribution over the measured and unmeasured confounders are the same in the two groups, then we can use the treated observations to infer what would have happened to the untreated observations.

Unfortunately, randomization is often not possible, either because there are ethical conflicts or because it is challenging to implement.

In the latter case, the most constraining factors are the time and monetary expense of data collection. Additional limitations of randomization include inclusion criteria that are too strict and cannot study large and representative populations. Moreover, inclusion criteria usually focus on simplified interventions (e.g., randomization to a drug versus placebo) that do not mirror the complexity of real-world decision-making. While the credibility (*internal validity*) and ability to advance scientific discovery of RCTs is well accepted, there are large classes of interventions and causal questions for which results that have a causal interpretation can only be gathered from observational data.

Let's now assume that in an observational study we compare two samples: one that adopts a dog (W=1) (exposed) and one that does not (W=0) (unexposed).

Then within each of these two populations, we calculate the rate of experiencing severe symptoms of depression Y=1:

	W=0	**W=1**	**Tot**
Y=1	780	830	1610
Y=0	220	170	390
Tot	1000	1000	2000

We found that adopting a dog appears to make the symptoms of depression worse: if you have a dog you are 5% more likely (83% versus 78%) to experience severe symptoms of depression:

$$P(Y = 1|W = 0) = 78\% < P(Y = 1|W = 1) = 83\%$$

Should we advise people not to own dogs? The problem with this analysis is that we ignore the fact that the subjects ***might be different in ways that would bias the conclusions***.

As mentioned before, a key potential confounder is the degree of severity of their depression symptoms ***before*** they were "assigned" the treatment (X).

For example, let's **stratify** the two populations (treated and untreated) based on whether they experience severe or milder depression symptoms of (X=1 versus X=0) before treatment assignment:

X	**Y**	**W=0**	**W=1**	**Tot**
X=1	**Y=1**	231	670	901
	Y=0	18	102	120
	Tot	*249*	*772*	*1021*
X=0	**Y=1**	549	160	709
	Y=0	202	68	270
	Tot	*751*	*228*	*979*
Total		1000	1000	2000

$$P(Y = 1|X = 1\,, W = 0) = 93\% > P(Y = 1|X = 1, W = 1) = 87\%$$

$$P(Y = 1|X = 0\,, W = 0) = 73\% > P(Y = 1|X = 0, W = 1) = 70\%$$

We find that **within these two population strata**, adopting a dog reduces the rate of experiencing severe symptoms of depression. This is an example of what is known as Simpson's paradox. Here the paradox occurs because people with severe depression symptoms before treatment assignment **are more likely** to adopt a dog. The solution that we applied corresponds to calculate a "CATE", i.e. a **conditional average treatment effect**, conditioning on the baseline depression level. [But how to estimate an ATE in this kind of situations? See the next example]. If we define:

$$e_i = P(W_i|X_i)$$

as the ***propensity*** of adopting a dog conditional to the level of depression symptoms pretreatment assignment, then in this example:

$$P(W_i = 1\,|X_i = 1) = \frac{772}{772 + 249} = 0.76$$

is higher than:

$$P(W_i = 1 \,|X_i = 0) = \frac{228}{228 + 751} = 0.23$$

In other words, the assignment to treatment, who gets a dog and who does not, is not completely random, as in an RCT. It is influenced by the pre-existing level of depression of the study subjects. Situations like these are very common in observational studies!!! We will see in Block 3 how to generalize the estimate of the propensity according to ***multiple*** confounders.

Example 2:

Imagine that first we observe mortality rates among subjects who received treatment A versus treatment B in an observational study without measuring their baseline condition:

	Death	**Alive**	**Tot**
Treat A	240	1260	1500
Treat B	105	445	550
Tot	345	1705	2050

From these data:

$$P(Death|A) = 16\% < P(Death|B) = 19\%$$

Treatment A seems better than Treatment B.

Let's now stratify subjects based on their baseline condition **before** treatment assignment:

Condition	**Death**	**A**	**B**	**Tot**
Mild	**Yes**	210	5	215
	No	1190	45	1235
	Tot	*1400*	*50*	*1450*
Severe	**Yes**	30	100	130
	No	70	400	470
	Tot	*100*	*500*	*600*
Total		1500	550	2050

Now, if we compute the probability of death under Treatment A and B as follows:

$$\sum_C P(D|A,c) * \boldsymbol{P(C = c|A)} = \frac{210}{1400} * \frac{1400}{1500} + \frac{30}{100} * \frac{100}{1500} = 0.16$$

$$\sum_C P(D|B,c) * \boldsymbol{P(C = c|B)} = \frac{5}{50} * \frac{50}{550} + \frac{100}{500} * \frac{500}{550} = 0.19$$

We obtain exactly the initial estimate: this is a naïve approach since we are not considering here that the probability of mild and severe conditions is distributed *differently* between the two

treatments. This is equivalent to the first computation, when we *ignored* the fact that there was a very different distribution of the baseline condition. But, the vast majority of subjects in Treatment A are in a Mild condition, and viceversa for treatment B. The baseline condition is therefore associated both to the treatment received and to the final outcome.

Note that the CATE here is:

$$P(D|mild\ ,A) = 15\% \ > \ P(D|mild, B) = 10\%$$

$$P(D|severe\ ,A) = 30\% \ > \ P(D|severe, B) = 20\%$$

So, the direction of the effect is always indicating that B is better than A **when conditioning on the baseline condition**. How can we recover ATE that is coherent with this finding?

To taking into account this problem, we could adopt the following approach:

$$\sum_C P(D|A,c) * \boldsymbol{P(C=c)} = \frac{210}{1400} * \frac{1450}{2050} + \frac{30}{100} * \frac{600}{2050} = 0.194$$

This represents the effect of Treatment A if we had treated all the Mild subjects in the population and all the Severe subjects (all in the population treated with A).

$$\sum_C P(D|B,c) * \boldsymbol{P(C=c)} = \frac{5}{50} * \frac{1450}{2050} + \frac{100}{500} * \frac{600}{2050} = 0.129$$

This represents the effect of Treatment B if we had treated all the Mild subjects in the population and all the Severe subjects (all in the population treated with B).

In other words, in this way we *weight* each treatment effect as if *all subjects in the population were treated* versus *all were untreated* [and we **standardize** the effect to the *overall observed distribution* of the confounder in the population].

This approach is known in causal inference literature as the **G-formula***.

This example is a non-parametric approach to estimate the ATE effect; then it is also possible to use regression models (i.e. a parametric approach) to obtain this kind of estimate. Note that ATE is a **marginal** causal effect.

*It is among a broad class of so-called "**g methods**" (where the "g" stands for "generalized") developed by James Robins, with deep roots in causal inference research and is widely used in biostatistics, epidemiology, and medical sciences to assess time-varying treatment effects in longitudinal data; see, e.g., Hernán and Robins (2020, Part III) and Naimi et al. (2016) for introductions to g-methods.

Hernán, M. A., & Robins, J. M. (2023). Causal inference: What if. Chapman & Hall - CRC.

Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2016). An introduction to g methods. International Journal of Epidemiology, 46 (2), 756–762. https://doi.org/10.1093/ije/dyw323