DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 5

CAUSATION AND CONFOUNDING, PART I by Chris Wild

Welcome back. We're continuing with our series, why what I see is never quite the way it really is.

Today we'll talk about difficulties in determining cause and effect. We'll start by having a look at the effect of smoking on how well children's lungs work. We have data on 654 children from Boston in the '70s.

The lung function outcome is called FEV. FEV stands for forced expiratory volume, the amount of air the child blew out in a second. High values are good, strong, healthy lungs.

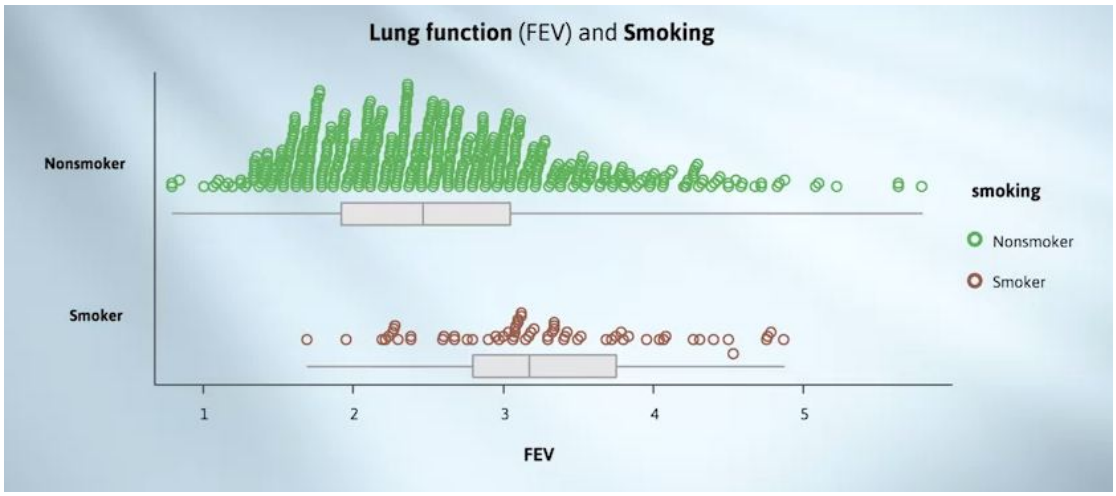


Fig 5.1

Here's a plot of lung function and smoking. Two things are immediately obvious. One is that there are a lot more non-smokers than smokers. The second, and more interesting thing, is the shift towards higher values for the smokers. Smokers tend to have higher lung function values than non-smokers. This is looking a whole lot like smoking is good for your lungs.

At this point, you are probably starting to smell a rat. What's going on here?

FEV	smoking	age	height	sex
...	...	...	...	...
3.25	Nonsmoker	10	66.0	Male
3.24	Nonsmoker	11	66.0	Female
3.44	Nonsmoker	14	62.5	Male
3.06	Nonsmoker	11	61.0	Male
3.01	Nonsmoker	10	62.0	Male
3.49	Nonsmoker	10	66.5	Male
2.86	Nonsmoker	10	60.0	Female
3.43	Smoker	14	64.0	Female
2.25	Nonsmoker	10	58.0	Female
4.68	Nonsmoker	14	68.5	Male
2.35	Nonsmoker	10	61.5	Male
4.39	Nonsmoker	12	68.5	Male
3.21	Smoker	13	61.0	Female
2.59	Nonsmoker	10	65.0	Male
3.19	Nonsmoker	13	70.0	Male
1.69	Smoker	11	60.0	Male
3.96	Smoker	14	72.0	Male
2.35	Nonsmoker	11	59.0	Female
4.79	Smoker	13	69.0	Male
3.52	Nonsmoker	11	67.5	Male
...	...	...	...	...

Fig 5.2

We know a bit more about these children than their FEV values and smoking status. The fact that there are so few smokers is also a bit of a clue. When did they start smoking?

I'm going to look at smoking and age.

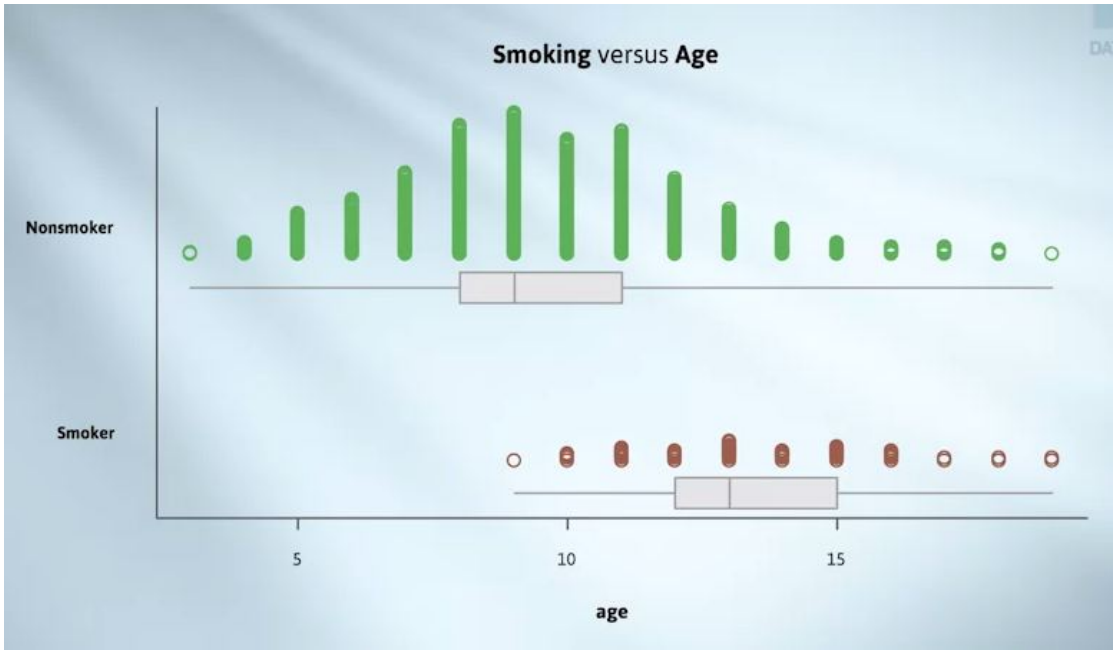


Fig 5.3

You may already have guessed it, the younger children didn't smoke.

What happens as children get older?

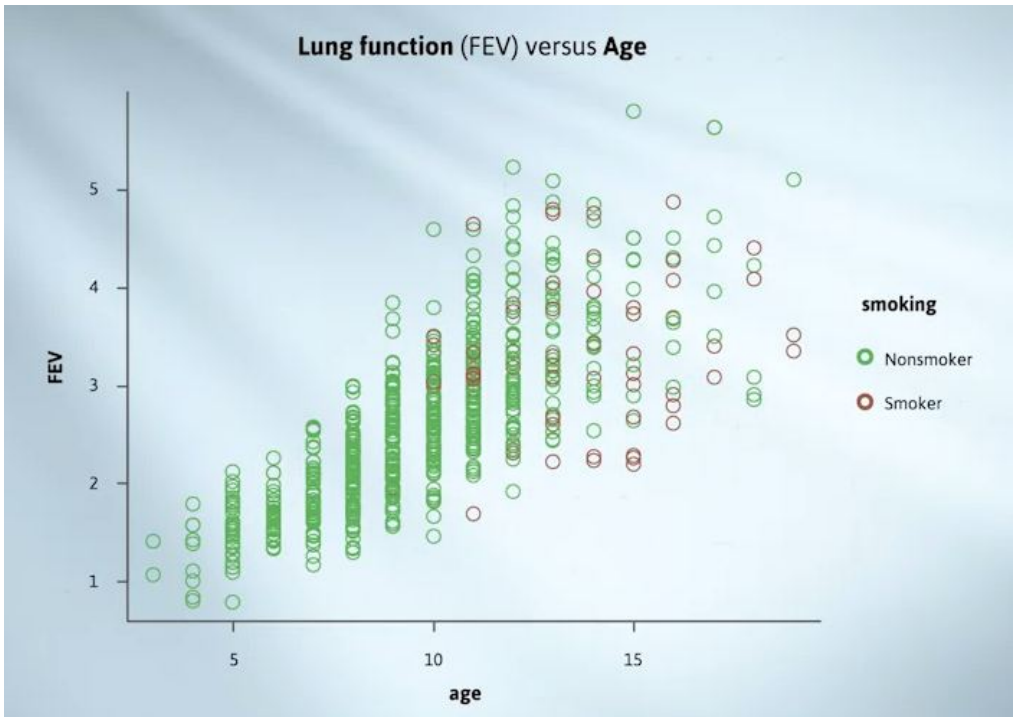


Fig 5.4

Well, their bodies, including their lungs, get bigger and stronger. Older children can blow out a lot more air.

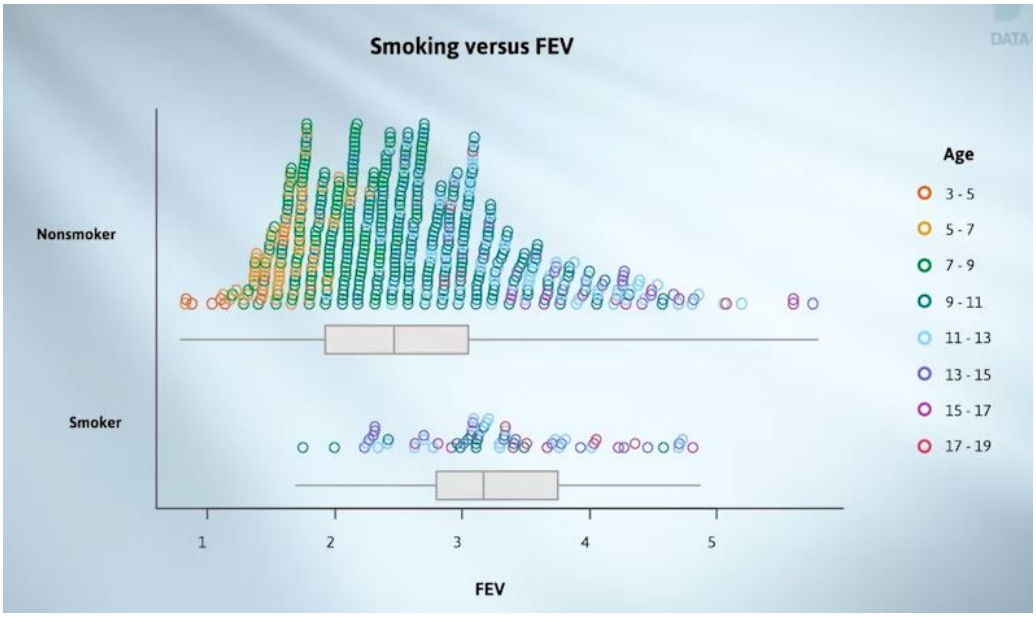


Fig 5.5

So in comparing smokers to non-smokers, we're comparing a group with lots of younger children with small lungs to a group of mainly older children with larger, stronger lungs. This is obviously not a fair comparison of the effect of smoking on lung function.

Before jumping to the lesson from all of this, I want you to try to second-guess your own thinking.

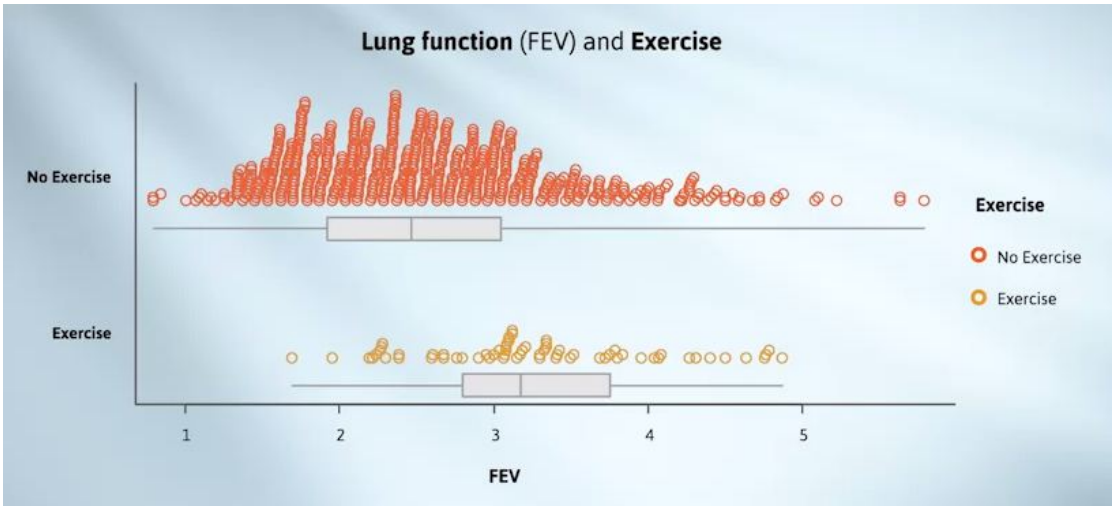


Fig 6

I'm willing to bet that if I'd shown you this instead, you'd have immediately been happy with concluding exercise improves lung function. We got suspicious in the first case because it clashed with our received wisdom that smoking is bad for you. But here we are much less likely to think critically and dig deeper because it aligns with another piece of received wisdom, exercise is good for you.

The big message is that we should never conclude, just from seeing a relationship in data like this, "That's what did it".

There is always the chance that the real culprit is that the groups we are looking at are unbalanced with respect to some other important factor, like age here.

Time for some name calling.

A variable is said to be a **cause** of changes in the outcome if actually changing its value leads to a change in the pattern of the outcomes. People tend to jump to causal conclusions too fast based on too little evidence. The language in common usage, like the word "effect" itself, suggests that a relationship is causal. We need to be more careful.

An **observational study** is one where your data comes from observing and recording conditions as they are in the world, rather than purposefully changing those conditions and then observing the consequences. For example, when we just record

who smoked and who didn't, that is observational. For observational data, we should restrict ourselves to the language of **association**. Positive association, when things tend to occur together, and negative association where the opposite occurs.

A **causal conclusion** (this is what did it) can never be justified on the basis of observational data alone. There has to be other supporting evidence.

In the context of our example, "purposefully changing the settings" would be making some people smoke and others not smoke.

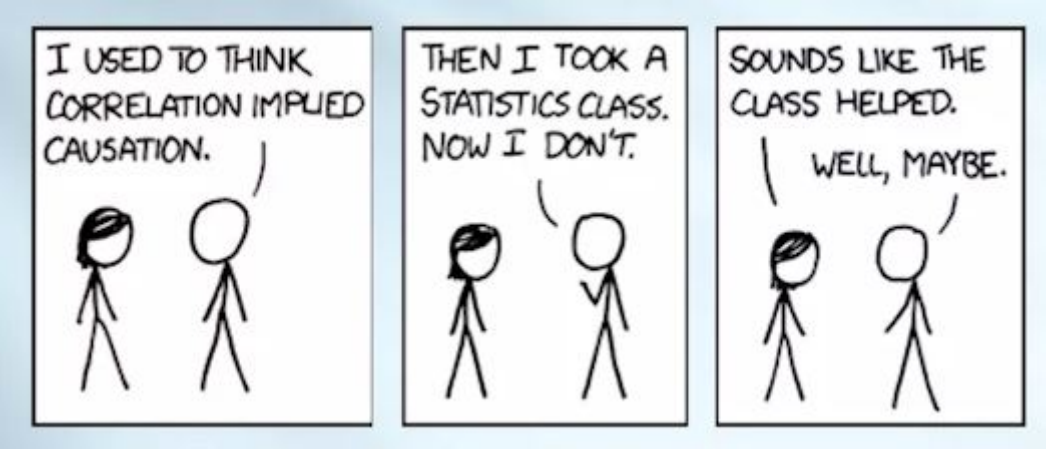
In a **randomised experiment**, we change conditions purposefully and use a random process to decide who will be subjected to what conditions. The **randomisation** is done to try to ensure balance on other causal factors.

A variable like age in our example is called a **confounder**, or **lurking variable**. Strictly, a confounder is something that causes changes in both the outcome and the predictor of interest.

When we look at smoking in a naive way, as in the graph (Fig 5.1), its effects are all mixed up (confounded) (Fig with the effects of the confounder variable Age.

So when we look at this graph, we are not looking at the true effect of the predictor smoking. The most reliable way people know of achieving comparable groups and reaching valid causal conclusions is to use randomised experiments. We'll talk more about them in Week Seven.

What we're seeing in this graph, an association between variables and observational data, just gives another instance of correlation does not imply causation.

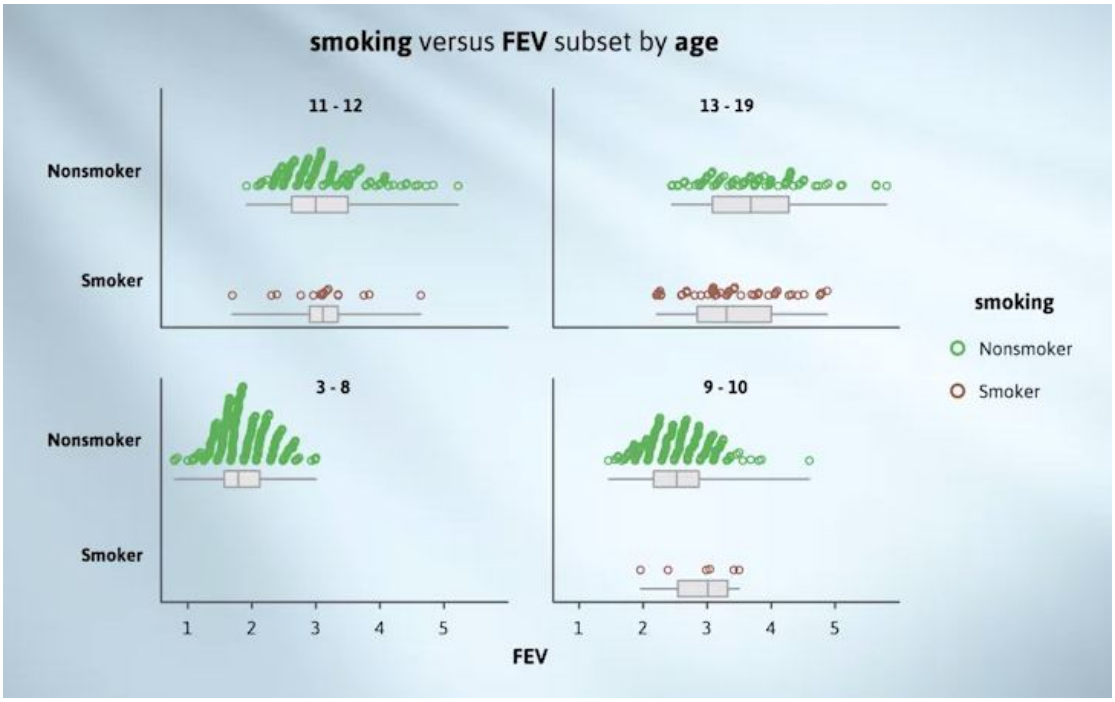


From: xkcd online comic site *http://xkcd.com/552/*

It does, however, as the cartoon suggests, point to some promising places to look. Correlation doesn't imply causation, but it does waggle its eyebrows suggestively, and gesture furtively while mouthing, "Look over there".

While an association between variables and observational data is not proof of anything, it is a clue that can be very helpful in on-going detective work. How can we adjust for a known lack of balance on a confounder?

Here's the basic idea. We make our comparisons within groups that have similar values of the confounder.



Here, the confounder is Age.

There were no smokers in the group aged eight and under, almost no smokers in the 9-10 age group. There were only 16 smokers in the 11-12 age group, and a suggestion of higher FEV values for smokers. But the bulk of the smokers are 13 and over. Among these teenagers, there is a clear shift towards larger FEV lung function values for non-smokers. The original simple plot comparing smokers to non-smokers got the effect backwards because of a lack of balance on the confounding variable Age.