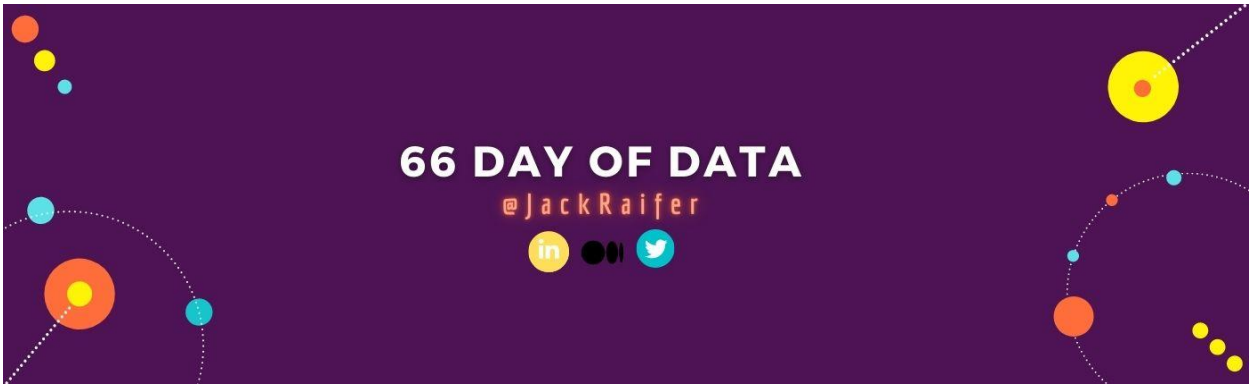

#66DaysOfData – Days 12 to 14: How to Create a Biostatistical Risk Formula – Risk of Lung Cancer for Smokers

On my last article from a couple of days back, I decided to start by building a risk formula using the information from the original studies we found, others I have spent the last days reading (and will share where I can in this article. The reason to do this, for me, is that this allows me to go deeper into the possible reality of being able to predict attrition from personality (using the OCEAN model in this case), using what scientist had to do before the existence of what we now refer to as Data Science.

How is a biostatistical risk formula built?

To explain, let me use the formula for risk of developing lung cancer in smokers within the next 6 years (meaning people who smoke or have smoked at any point in their lives). To build this formula, researchers had to jump into tons of data related to smokers and whether they had developed lung cancer. Then they had to figure out the correlations between the different factors that seem to affect the possibility of developing lung cancer. The most important ones are: sex, age, number of packs years, cigarettes per day, number of years since quitting, Body Mass Index, exposure (hours per day in a smoke filled environment) and cough (does the person cough daily during periods of the year, yes or no).

The formula looks something like this:

R6years = 1.18203062 + (0.3157 * sex) – (1.985 * [(age/100) ^ -1]) + (1.120 * [log(pack_years)]) – (0.040 * cigarettes per day) – (0.2402 * [log(years since quitting)]) – (1.7024 * [log(Body Mass Index)]) + (0.0807 * [log(exposure)]) + (0.4921 * cough)

Apologies for the vey long formula, as you can see, people before the data tools we had today had it rough. Let us take this equation piece by piece so we can understand it better.

R6years: simply means the risk of x happening in the next 6 years, so this would be the formula for the risk of developing lung cancer in the next 6 years. As you can see, there is a base risk defined as 1.18203062, this is the base risk that the researchers found, and hence, it is the base of our formula.

Sex: we add a risk factor variable for sex, 1 for male and 0 for female, this means that there is an increased risk for men to develop lung cancer of 0.3157.

Age: age tends to always be a risk factor, in this case, the researchers decided to build a bit of a complex relation from what they found, by subtracting a diminishing variable from the total risk. It is a diminishing variable since the higher the age, the smaller this variable becomes, and this formula simply represents the complexity of this relation with the lung cancer risk.

Pack Years: this adds the risk generated by the number of cigarettes you smoked during your time as a smoker, in other words, if you smoked 20 cigarettes a day for 20 years, that means you smoked 365 per year, times 20 years, or 7300 packs. This is the one variable that adds the largest amount of risk.

Cigarettes per Day: This one is complex to understand since the more cigarettes per day, the lower your risk. Turns out the pack years variable already contemplated the risk of the amount of smoking, and since there is a small diminishing increase in risk the more you smoke in a day, then this variable slightly reduces the risk the more you smoke per day. To try and explain better, it is worse to increase from 5 to 6 cigarettes per day that from 8 to 9, so this variable showcases by weighing that fact against the previous variable that already calculated the increased risk of smoking more.

Years since Quitting: Now we reduce the risk, again with a bit of complexity, depending on how many years it has been since the person quit smoking.

BMI: Body Mass Index, this one is counterintuitive. Turns out that the worse in shape you are, the lower your risk of developing lung cancer as a smoker. Now, this does not mean that it is good to be out of shape, it simply reflects the fact that out of shape people develop other diseases and die from them, before developing lung cancer, hence it lowers your risk of developing it.

Exposure: The amount of time you spend around other smokers, also increases your risk. Again, there is a complex relation here, but in general, more time equals more risk.

Cough: Finally, the cough, if you cough every day for a period of time with certain frequency, you are at higher risk, this variable can be either a 0 for no or a 1 for yes.

There are other parts of the formula, that adds coefficients, which are then used to transform the result into a percentage. These are all based on the research and models by Maria Maraki.

Here is a calculator you can easily use, so as not to spend a LOT of time on this, our biostatistics formula will be a LOT easier.

The idea was for me to explain how these kinds of formulas are built; these would be our hand-built algorithms. The reason I want to build one is simple, I have been reading tons of papers regarding attrition, or turnover and OCEAN personality, then problem is none of them give you access to the data used, just the results. Nonetheless, we can take all of the correlations found, work some statistical magic on them, and come up with the variables we could use to build a hand-built prediction model.

After all, it is always fun to start building "by hand", since it gives us deeper domain knowledge and allows us to discover interesting things.

Next Time – Building an OCEAN – Attrition Model Formula with Biostatistics.

Jack Raifer Baruch

Follow me on Twitter: @JackRaifer

Follow me on LinkedIN: jackraifer

About the Road to Data Science - #66DaysOfData Series

Road to Data Science series began after I experienced the first round of Ken Jee´s #66DaysOfData challenge back in 2020. Since we are starting the second round of the challenge, I thought it would be a good idea to add small articles every day where I can comment my progress.

I will be sharing all the notebooks, articles and data I can on GitHub:
https://github.com/jackraifer/66DaysOfData-Road-to-Data-Science

Please do understand I might have to withhold some information, including code, data, visualizations and/or models, because of confidentiality regards. But I will try to share as much as possible.

Want to follow the #66DaysOfDataChallenge?

Just follow Ken Jee on twitter **@KenJee_DS** and join the #66DaysOfData challenge.

You can also reach out to me at any time through **LinkedIN** or **Twitter.**