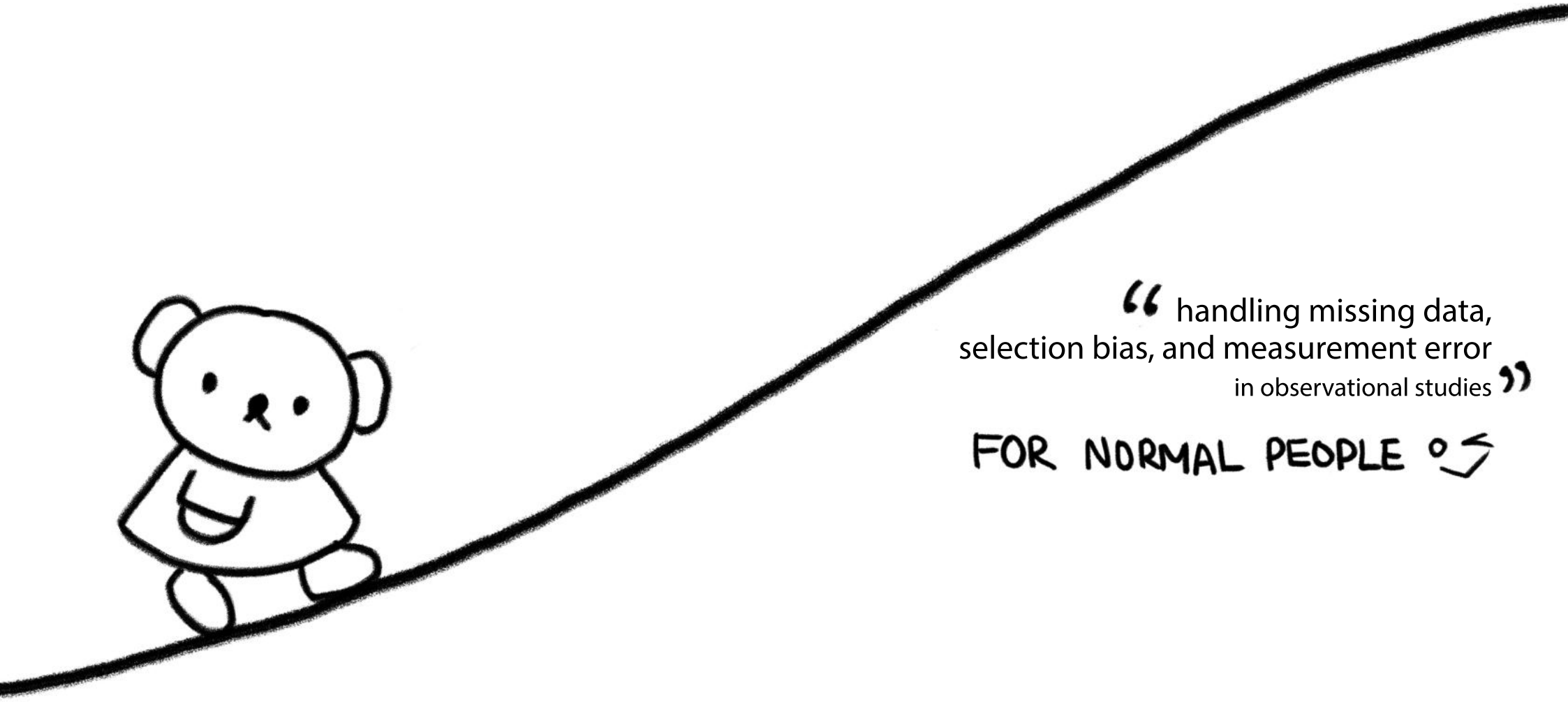
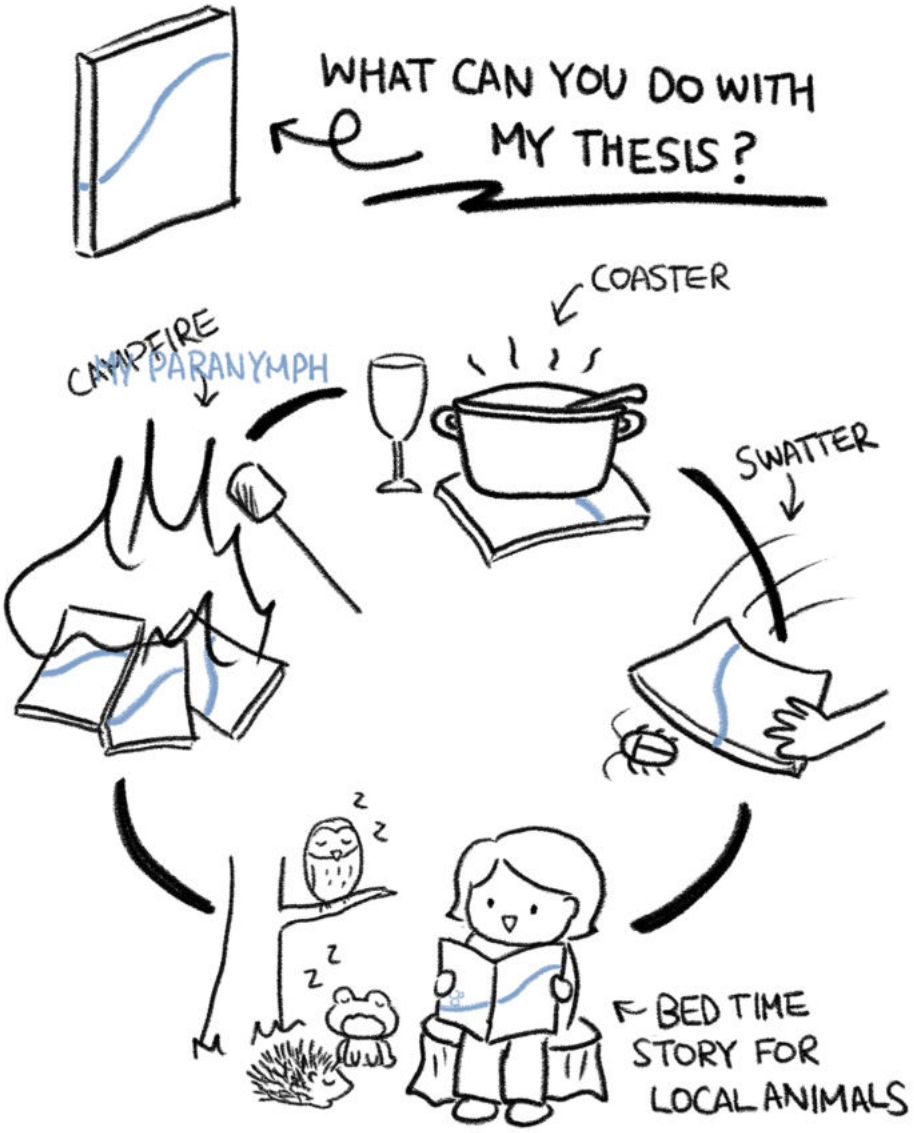


“ handling missing data,
selection bias, and measurement error
in observational studies ”

FOR NORMAL PEOPLE

jungyeon choi

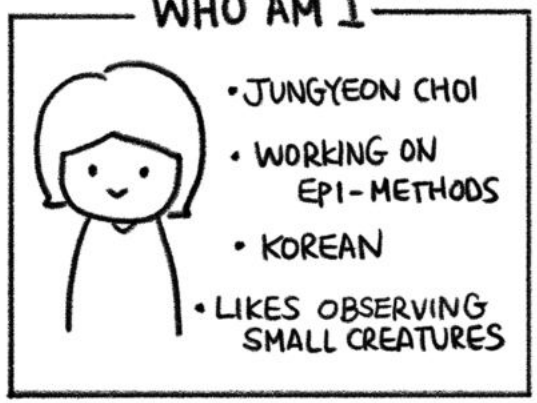


FOR THOSE HAPPENED TO BE INTERESTED IN THE CONTENT OF MY THESIS, READ THIS ZINE!

TABLE OF CONTENTS

Disclaimer

This Zine is to communicate my thesis to friends and family less familiar with epidemiology. For this purpose, I simplified the examples. In reality, what my epidemiologist and statistician colleagues are working on is much more contextual than the given examples.



1. WHAT IS EPIDEMIOLOGY?

Epidemiology is the study of how health-related events and health-related factors are distributed in specified populations and why [1. 2].

WHAT DO WE MEAN WITH THESE TERMINOLOGIES?

- **Health-related events** could be a disease such as diabetes or cancer. Or, it could be a state like being overweight or having high blood pressure.

- **Health-related factors** could be many things - biological factors (such as physiological or genetic conditions), socio-economic factors (like income level or education), behavioral factors (like smoking or exercising), or clinical care (such as medication or surgeries).

- **A specified population** means a group of people you want your study to say something about. It can vary depending on how big your study is and your interest. Do we want to say something about all Dutch people? About white women under 65? Or maybe about adults who are diagnosed with breast cancer?

ALL THE SUDDEN YOU STARTED TO HEAR 'EPIDEMIOLOGY' IN THE MEDIA :)

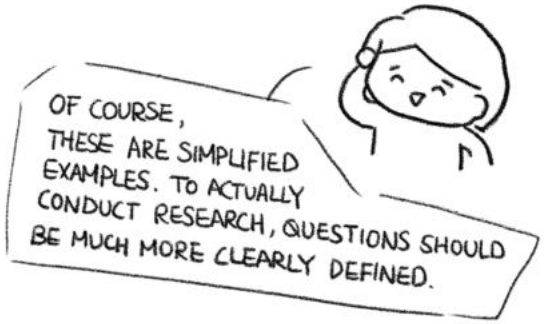
It may all still sound vague. So, let's say we are interested in Covid-19. As an epidemiologist, you could be asking these types of questions.

Descriptive: What are the characteristics of people who got infected with Covid-19? Where and when did the outbreaks happen?

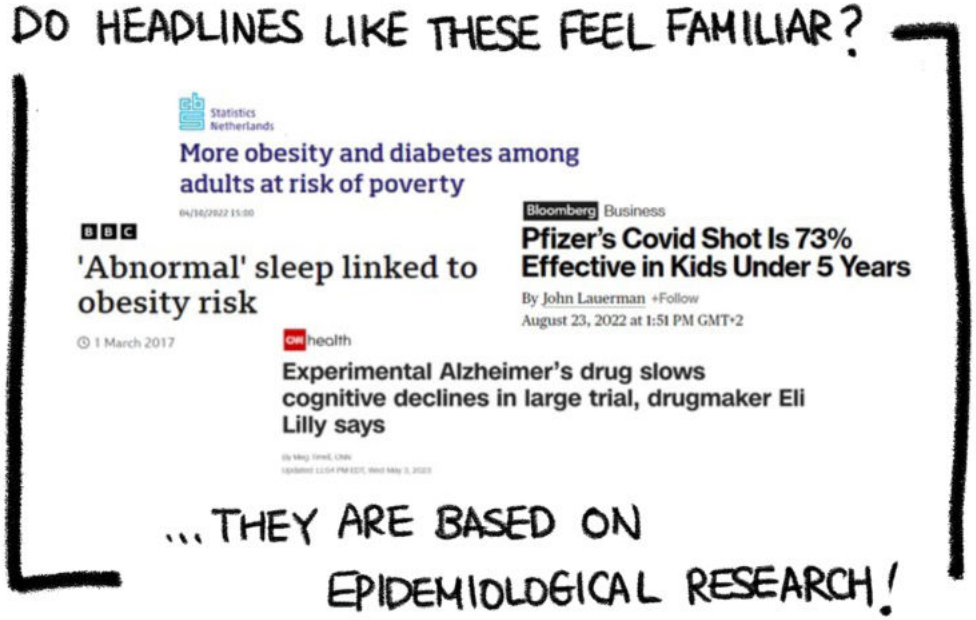
Diagnostic: How do we detect whether a person is infected with Covid-19? Does rapid antigen testing perform as well as PCR testing?

Prognostic: Who will be infected with Covid-19? How does the health-status of a person infected by Covid-19 develop over time?

Interventional: Does the lock-down policy reduce the risk of Covid-19 infection? Is Covid-19 vaccination affective?



DO HEADLINES LIKE THESE FEEL FAMILIAR?

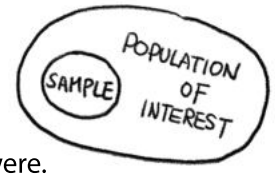


...THEY ARE BASED ON EPIDEMIOLOGICAL RESEARCH!

2. How is my thesis related to epidemiology?

The title of my thesis does not seem to say anything about health-related issues. So, why am I even talking about epidemiology?

Let's say we want to know among the Dutch population whether people who received Covid-19 vaccination experience milder symptoms of Covid-19 compared to people who have not recieved the vaccination.



It is extremely resource-heavy (and probably impossible) to follow every single person in the Netherlands to check whether they ever got Covid-19 and how severe their symptoms were. Instead, what we can do is to select a group (a sample) who can represent the overall population of the Netherlands well. Then, we collect data from the selected group and study the relationship between Covid-19 vaccination and severity of Covid-19 symptoms. From the results, we can say something about the overall Dutch population, which is our population of interest.

Now, we can randomly select a half of the sample and give Covid-19 vaccination and not give the vaccination to the other half. Then, we follow them until they get infected by Covid-19 and see if the symptom severity differs between the two groups - in other words, we conduct a randomized controlled trial (RCT) for Covid-19 vaccination vs. no vaccination.

THE ESSENCE OF RCT IS THAT COMPARING TREATMENT OPTIONS ARE GIVEN RANDOMLY. THEREFORE, THERE IS NO SYSTEMATIC DIFFERENCE BETWEEN THE GROUPS. (IT'S LIKE THROWING A COIN TO DECIDE WHICH TREATMENT TO GIVE FOR EACH INDIVIDUAL.)

However, conducting a randomized controlled trial like this is often not feasible. First, randomizing treatment can be unethical. For instance, you cannot justify randomly withholding the vaccination from some people if you already know that it reduces the risk experiencing severe covid-19 symptoms. Also, if you are looking into very rare disease, lilke myocarditis, your trial may have to follow a huge number of people for years - it gets expensive and inefficient.

RCTS OFTEN HAVE STRICT PATIENT INCLUSION CRITERIA. THEREFORE, PEOPLE INCLUDED IN A RCT MAY NOT BE A REPRESENTATIVE OF A GENERAL POPULATION.

In such cases, observational studies can be an alternative choice. In observational studies, researchers do not have control over a treatment allocation but observe what happens in the real-world as it is. Therefore, they may provide better insight to the real-world than RCTs. But the perk comes with a price. Unlike RCTs, comparing treatment options are not randomly assigned. Thus, in observational studies, you cannot guarantee that the comparing groups are similar to each other. In fact, they could be quite different.

THERE ARE MANY AMAZING DEVELOPMENTS IN THE WORLD OF CLINICAL TRIALS, WHICH I KNOW VERY LITTLE ABOUT

GOING BACK TO OUR COVID-19 EXAMPLE ...

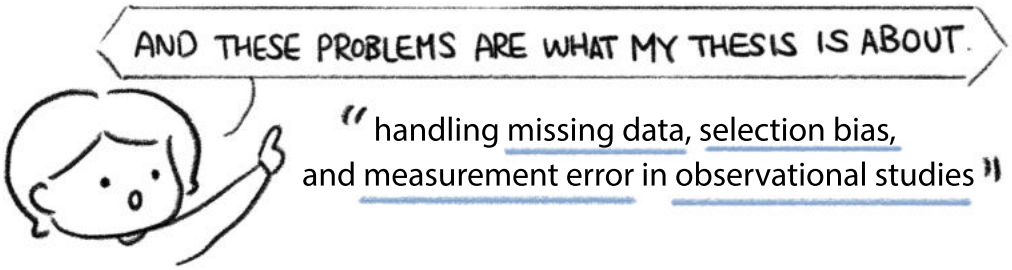
During the pandemic, people who did not get any covid-19 vaccination could be different from people wo did in terms of their life styles, health status, health-related behaviors etc. These factors also could influence how severely one experiences Covid-19 symptoms. This makes it difficult to isolate the effect of the vaccination from other factors.

Adjusting for differences between the comparing groups is one of the most important issues in observational studies. If not correctly done, the comparison could be false - which in epidemiology is called 'confounding bias'.

Apart from confounding bias, there are other issues that bring challenges in observational studies. Typically, these issues are missing data, selection bias, and measurement error.

- **Missing data** can occur if not all information is recorded in the data. Younger people, for example, may be less worried about their health and do not report their symptoms when infected with Covid-19.
- When people included in your data are systematically different from your population of interest, **selection bias** may occur. For example, people living in the in the randstad may have behaved differently during the pandemic compared to people living in other provinces. Your data cannot well-represent the overall Dutch population if you only includes people living in the randstad.
- Your measuring tools are not always accurate. You might have heard that rapid antigent testing has a high false-negative rate (the test result shows you are not infected when you are actually infected with Covid-19). This can introduce **measurrement error** bias into your study.

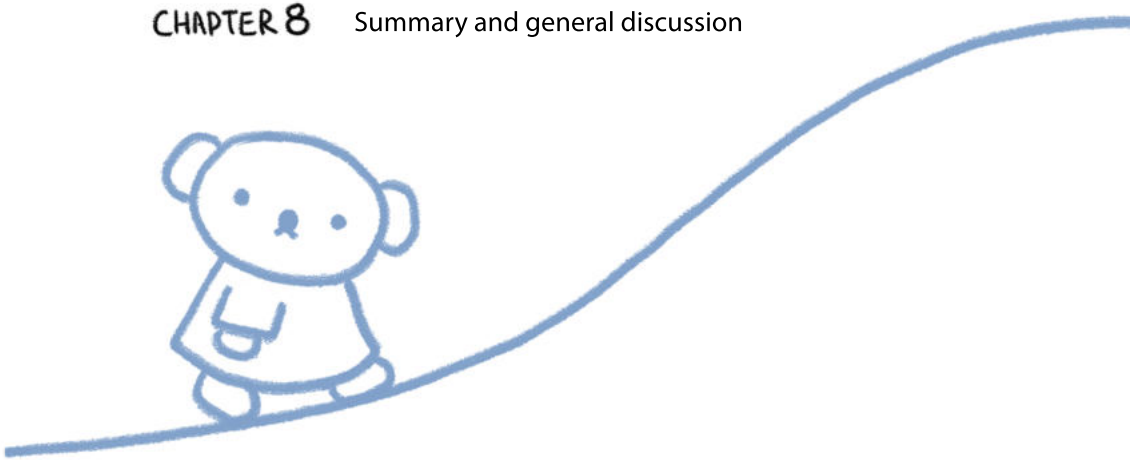
Without correctly addressing these potential biases, your study results would be less trustworthy, and your interpretation could be invalid.

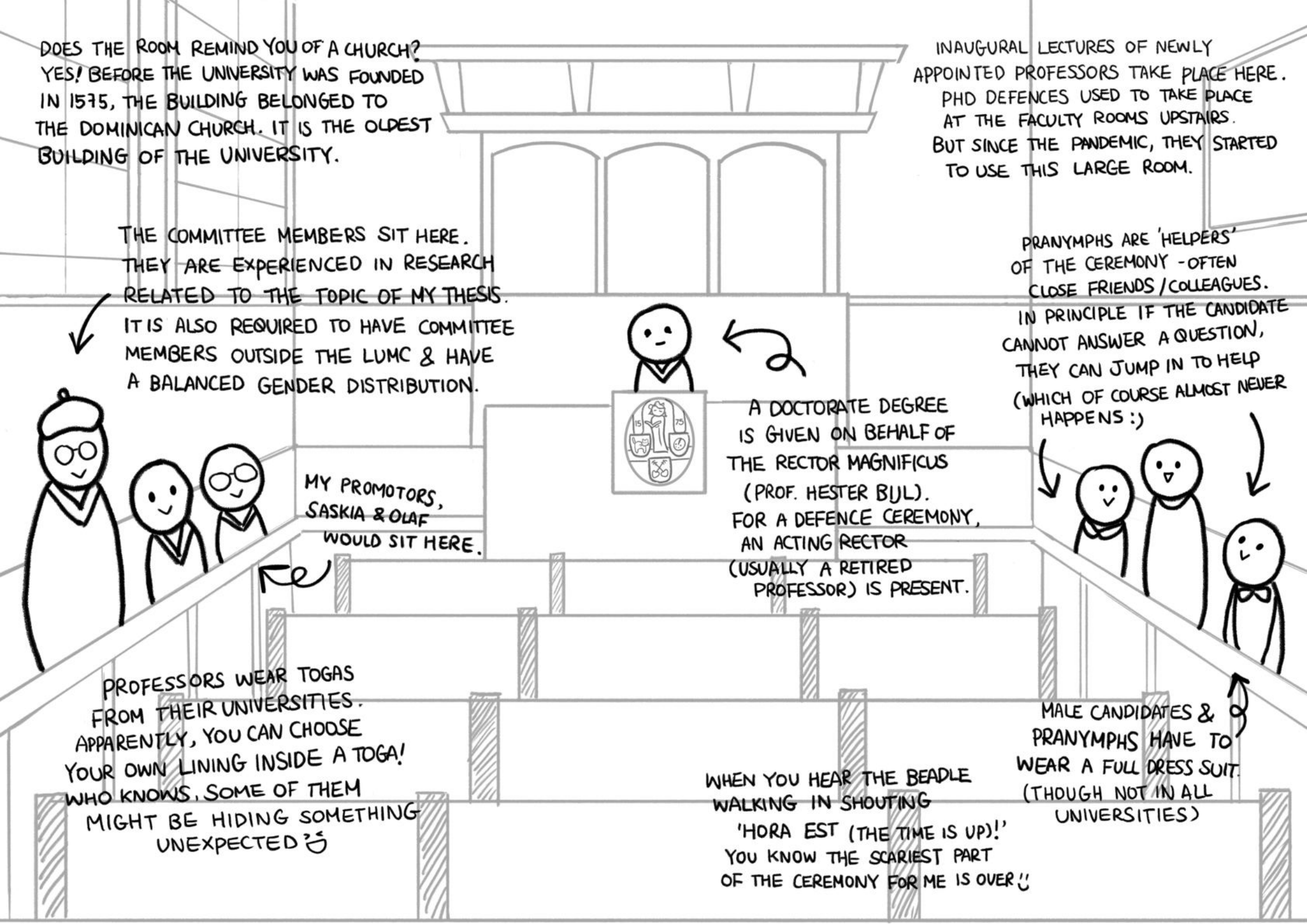


Of course these are still huge topics! My thesis look at these issues in specific narrowed-down research settings.

3. WHAT IS EACH CHAPTER OF THE THESIS ABOUT?

CHAPTER 1	Introduction
CHAPTER 2	A comparison of different methods to handle missing data in the context of propensity score analysis
CHAPTER 3	Comparing methods for measurement error detection in serial 24-hour hormonal data
CHAPTER 4	A comparison of different methods for handling measurement affected by medication use
CHAPTER 5	Estimating medication effects using routinely collected electronic health records
CHAPTER 6	How measurements affected by medication use are reported and handled in observational research: a literature review
CHAPTER 7	Tying research question and analytical strategy when variables are affected by medication use
CHAPTER 8	Summary and general discussion




DOES THE ROOM REMIND YOU OF A CHURCH? YES! BEFORE THE UNIVERSITY WAS FOUNDED IN 1575, THE BUILDING BELONGED TO THE DOMINICAN CHURCH. IT IS THE OLDEST BUILDING OF THE UNIVERSITY.
INAUGURAL LECTURES OF NEWLY APPOINTED PROFESSORS TAKE PLACE HERE. PHD DEFENCES USED TO TAKE PLACE AT THE FACULTY ROOMS UPSTAIRS. BUT SINCE THE PANDEMIC, THEY STARTED TO USE THIS LARGE ROOM.
THE COMMITTEE MEMBERS SIT HERE. THEY ARE EXPERIENCED IN RESEARCH RELATED TO THE TOPIC OF MY THESIS. IT IS ALSO REQUIRED TO HAVE COMMITTEE MEMBERS OUTSIDE THE LUMC & HAVE A BALANCED GENDER DISTRIBUTION.
PRANYMPHS ARE 'HELPERS' OF THE CEREMONY - OFTEN CLOSE FRIENDS / COLLEAGUES. IN PRINCIPLE IF THE CANDIDATE CANNOT ANSWER A QUESTION, THEY CAN JUMP IN TO HELP (WHICH OF COURSE ALMOST NEVER HAPPENS :)
A DOCTORATE DEGREE IS GIVEN ON BEHALF OF THE RECTOR MAGNIFICUS (PROF. HESTER BIJL). FOR A DEFENCE CEREMONY, AN ACTING RECTOR (USUALLY A RETIRED PROFESSOR) IS PRESENT.
MY PROMOTORS, SASKIA & OLAF WOULD SIT HERE.
PROFESSORS WEAR TOGAS FROM THEIR UNIVERSITIES. APPARENTLY, YOU CAN CHOOSE YOUR OWN LINING INSIDE A TOGA! WHO KNOWS, SOME OF THEM MIGHT BE HIDING SOMETHING UNEXPECTED 😏
WHEN YOU HEAR THE BEADLE WALKING IN SHOUTING 'HORA EST (THE TIME IS UP)!' YOU KNOW THE SCARIEST PART OF THE CEREMONY FOR ME IS OVER!!
MALE CANDIDATES & PRANYMPHS HAVE TO WEAR A FULL DRESS SUIT. (THOUGH NOT IN ALL UNIVERSITIES)

<CHAPTER 2>

'Propensity score analysis' is a statistical method to mimic a randomized control trial with observational data. In chapter 2, we compared several approaches for dealing with missing data in the context of propensity score analysis.

A BIT ABOUT 'MISSINGNESS'...

Missing data can happen through different mechanisms. If something is missing completely at random - for example, a lab technician accidently dropped a tube with blood sample - for this, it may be ok to ignore the missing data and only use the data you have. Sometimes, however, there could be systematic reasons why certain information is missing. For example, general practitioners only record BMI if patients are noticeably overweight or underweight. Therefore, BMI information is missing for people with normal weight. In this case, only using the recorded BMI in your analysis may lead to bias.

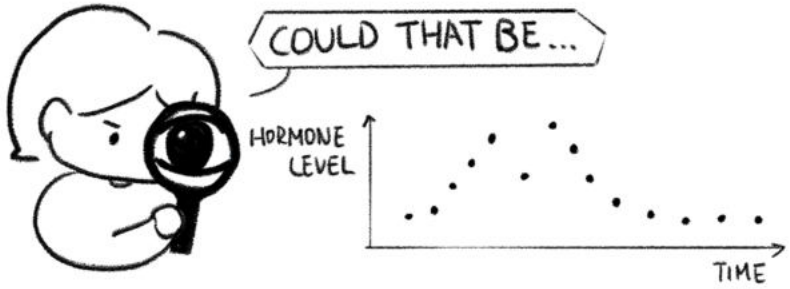
Therefore, using an appropriate method for handling missing data according to the assumed missing data mechanism is important for making a valid conclusion with your study.

<CHAPTER 3>

The Leiden Longevity Study (LLS) collected 24-hour hormonal data of some of the study participants. Blood was withdrawn every 10 minute from the participants. Later, the blood sample was analyzed to measure levels of hormones.

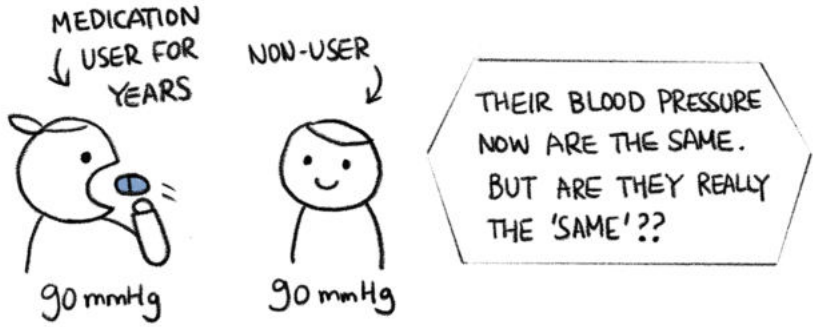
From the data, Evie, a reasearcher in the LLS, noticed that sometimes severalhormones from the same blood sample were measured lower than expected. She suspected that some of the blood samples might have been diluted and resulted in under-reporting of the hormonal levels - in other words, some hormones were measured with errors.

Together, we compared methods for detecting measurement errors in hormonal data and the impact of removing detected measurement errors when analyzing the data.



<CHAPTER 4&5>

Biological measurements such as blood pressure or blood glucose levels are often of interest in medical research. One of the challenges when using these variables is that many people use medication to control their blood pressure or blood sugar level. Ignoring the fact that some people's blood pressure has been lowered due to the medication use would lead to bias.



In Chapter 4, we compared available methods and provided guidance on which methods to use when. One of these methods is to use external information on the medication effect to correct for potential bias. The medication effect may be obtained from previous clinical trials. However, trial settings are often quite different from what happens in the real world. Therefore, in Chapter 5, we tried to estimate the medication effect using observational data and routinely collected electronic patient records.

<CHAPTER 6>

In Chapter 4, we showed that which methods you use for handling medication-effect can lead to different results. To see whether appropriate methods are used in clinical research, we performed a literature review.

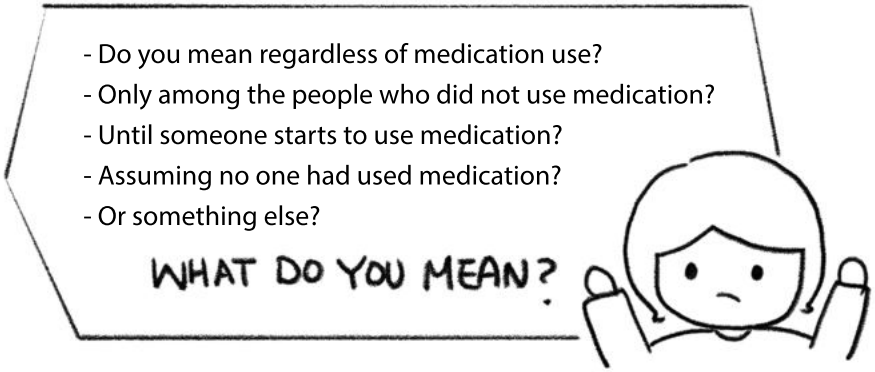
WE GOT HELP FROM A LIBRARIAN JAN SCHOONES. LITERATURE SEARCH IS A REAL SKILL!

We evaluated whether the research aim of each study matches the method they used for handling medication use. What have we found? Many studies overlooked potential bias due to medication use, and often there was a mismatch between their research aim and the method used. Also, statistical methods recommended in previous methodological research seemed not being well used in applied research.

<CHAPTER 7>

Is a research question 'what is the effect of X on Y?' enough? Let's think about the following question: what is the effect of having a certain genetic factor on blood pressure at age 40?

The question is seemingly straight forward - but what do you mean by 'blood pressure at age 40'?



Without clearly specifiying your research question, you will likely make arbitrary decisions while anlyzing data. A danger here is that you may end up having a mismatch between the interpretation of your study results and what you have done in your statistical analyses. Therefore, in Chapter 7, we discussed different types of questions that could be of interest when variables are affected by medication use and what are the appropriate analytical strategies under each type of question.

4. PROPOSITIONS

↳ OPPOSABLE AND DEFENDABLE STATEMENTS RELATED TO THE THESIS AND THE FIELD.

1. There is no one optimal statistical method that can handle biases across every study setting. Each source of bias should be handled on the basis of content specific knowledge.

2. Multiple imputation is not a panacea to handle missing values and should be used more consciously.

↳ A STATISTICAL METHOD FOR REPLACING MISSING DATA WITH SUBSTITUTED VALUES, WHILE ALLOWING THE UNCERTAINTY ABOUT THE MISSING DATA

3. Incorporating experts' content knowledge is recommended to detect measurement errors in time-serial data rather than solely relying on automated approaches.

4. A research question such as 'what is the effect of X on Y?' requires further elaboration. One should consider whether and how medication use or other factors have affected the measurements of interest.

5. Problems of confounding, selection, and measurement bias can be addressed with a question; what is the missing information? This calls for unified perspectives for addressing these biases.

↳ BASICALLY, THE INFORMATION NEEDED TO MAKE A VALID CONCLUSION IS MISSING.

6. Conducting comparison studies to evaluate existing methods should be incentivised. For many analysis problems, the issue is not a lack of available methods; rather, it is a lack of accessibility to available methods - After STRATOS initiative

↳ STRENGTHENING ANALYTICAL THINKING FOR OBSERVATIONAL STUDIES

7. Simulations allow empirical comparisons between available methods under various data structures and violation of assumptions. Utilizing simulation studies will benefit clinical researchers.

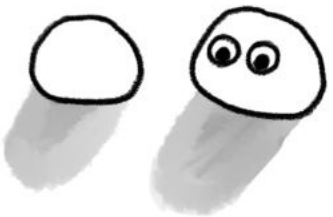
↳ COMPUTER EXPERIMENTS THAT INVOLVE GENERATING DATASETS WHILE ACCOUNTING FOR THE STRUCTURE & RANDOMNESS OF REAL DATASETS

8. Even in the emergence of big data and machine learning, careful considerations of the research setting, clinical knowledge, and study designs remain highly important - possibly more than ever..

'BIG DATA' IS NOT A CORRECTION FOR BIASES

9. Prisoners in a cave we (epidemiologists) are, looking at shadows (data) cast upon the cave wall. The shadows reflect a fragment of the real world (medical reality). - After The Allegory of the Cave

10. "Every new discovery is just a reminder - we are all small and stupid. [...] all of that exists inside of one universe out of who knows how many." (Everything Everywhere All At Once, 2022) Because nothing matters, everything we give meaning matters.



Immeasurable support, help, and fun memories made it possible to make this zine.

MY PROMOTORS,
SASKIA & OLAF

FIRST PHD ROOMIES
MONA & AMIR

MY PARANYMPH!
RUIFANG

CLINICAL EPIDEMIOLOGY COLLEAGUES @ LUMC

WITH THEM
COFFEE & COOKIE HOURS

AJDA & VID
& SUSHI

SEBAS, CHAVA, TARIQ
OFFICE MATES

WEON, ISCB, CIM
WITH LINDA, KIM, BAS

JULIUS CLINICAL DS TEAM
COOL PEOPLE COOL TEAM

VAN WIJK FAMILY
TOO MANY TO LIST ALL

CHOI & WOO FAMILY
♡아빠 엄마 재호♡

SHIMMY
WILLEM

진아 엄친 도윤 선영
FANTASTIC ~~SIX~~ FIVE

SMALL FRIENDS
설기, 연지, 천천, 부기...

POOMIE

HALMO

AND SO MANY MORE

REFERENCES

Coggon D, Barker D, Rose G. 2009. Epidemiology for the Uninitiated. John Wiley & Sons.

Last JM, editor. Dictionary of epidemiology. 4th ed. New York: Oxford University Press.

Sterne JAC, White IR, Carlin JB, Spratt M, RoystonP, Kenward MG et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls BMJ 2009; 338: b2393 doi:10.1136/bmj.b2393

Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods, Statistics in Medicine. 2019; 38: 2074– 2102. https://doi.org/10.1002/sim.8086

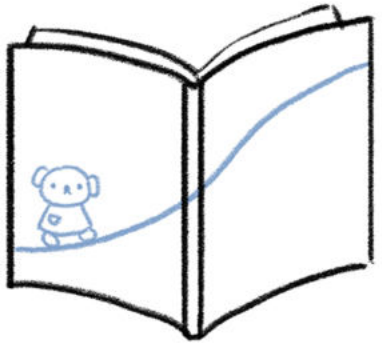
<NEWS ARTICLES>

https://www.cbs.nl/en-gb/news/2022/40/more-obesity-and-diabetes-among-adults-at-risk-of-poverty

https://www.bloomberg.com/news/articles/2022-08-23/pfizer-s-shot-is-73-effective-against-covid-in-children-under-5#xj4y7vzkg

https://www.bbc.com/news/uk-scotland-glasgow-west-39127601

https://edition.cnn.com/2023/05/03/health/alzheimers-drug-donanemab-eli-lilly/index.html



THE COVER IMAGE IS A WORK OF DICK BRUNA, 'BORIS OP DE BERG'.

THANK YOU FOR READING THIS ZINE & DO NOT HESITATE TO ASK ANY QUESTIONS OR TO SEND ME AN EMAIL!

Choi_Jungyeon@outlook.com

DOWNLOAD THE FULL THESIS HERE, OR AT

github.com/Yeon-Choi-git/PhDthesis