UiT: The Artic University of Norway

I'm saving this line for later

Preparing Fit Futures data for analysis

Rafael Adolfo Nozal Cañadas

supervised by

Prof. Lasr-Ailo BONGO
Prof. Anne-Sophie FURBERG
Assoc. Prof. Anne Merethe HANSSEN

June 15, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Abstract

This document describe how we prepare the data for analysis. It covers up to... .

1.2 Definitions

This section is a brief layman description of the different concepts or entities involved in this project. Please refer to the reference section for more extensive info about each.

1.2.1 People

Lars-Ailo, [1] Anne-Sophie, [2] Anne-Merethe, Dina, Christopher, Whoever own the FF1+ data

1.2.2 Data

1.2.3 Fit Futures

Fit Futures is divided in a series of interviews during a long period of time. The dates are as follow:

FF1 - Fit futures 1 FF11 - Several samples of S.Aureus were taken during FF1. A week later FF11 is looking for the control samples.

FF12 - A few month later we take the sample two of the S.Aureus in FF12

FF2 - Fit futures 2

FF3 - Fit futures 3

1.2.4 Legal

1.2.5 Privacy levels

1.2.6 Statistics

Reproducibility

1.2.7 Programming languages

LaTeX

LaTeX [3] is a system for preparing written documents. The main characteristich is that, unlike in Microsoft Word, LibreOffice, and other similar software where you see the final result of what you are written live as you edit it, in LaTeX the user types in plain text keeping style and content separated, in a similar way as HTML+CSS works. Later on the code is compiled into a PDF document where you can see the final stylized result.

The main inconvinient of LaTeX is the learning curve, however LaTeX is widely used in all academia fields for the communication and publication of scientific documents. In my opinion, the inconvience of having to learn LaTeX outweights, by far, the amount of time that you are going to save editing documents in a "What

You See Is What You Get" [4] editor. To the point that mathematics communications alone, would be near impossible to perform without this software.

As opposite to Microsoft Word where you can syncronize automatically with OneDrive, Latex doesn't have by itself a colaborative interface and you are dependent of using a control software, such as GIT, or using online services, such as Overleaf. In any case, this problem can also be adverted by setting automatic version control scripts; which to be fair, scare people outside mathematics, informatics, and physics fields.

SPSS

The letters SPSS [5] stands for "Statistical Package for the Social Sciences". Is a propietary statistics software developed for social sciences researcher who has very limited, or none, knowledge of programing.

The advantage of SPSS is that is composed of easy to use drop-down menus. If a person has some basic knowedle of statistics, SPSS is very easy to use even for complex multivariate analysis.

The shortcomings are that you can't do extensive scripting in SPSS, so anything that we do here that generate results automatically, generate latex code automatically, generate websites automatically, won't be possible. During the course of the project we also tend to change definitions of things, such as, what does it mean to be a carrier of a bacteria. Running manually all the drop-down menus in SPSS, each time we decide to change the definition of a disease, or who is the target population, would be an insormuntable time consuming task.

Other negative issues are the propietary license cost, worsened by the fact that it has a software as service license, plus the issues with close software security. Adding to this, there are statistical method within several libraries available in R that aren't in SPSS, namely almost everything that has to do with network analysis. For fairness, mention that it has a trialware licese where you can use the software for free, but I would strongly advice agaitns wasting your time using it given all the limitations that presents.

R

R has however a lot of shortcommings and limitation as a programing language, which I discuss in great detail in section...

C/C++

While C++ would be my prefered program of choice, it is not the most popular language for data analysis, in part for the initial learning curve that has over R or Python.

Python

1.2.8 Technologies

SHA1

Let say two person have two files with thousand of rows and thousand of columns. They suspect that the files are different because they are not getting the same results after doing the same analysis.

Normally, we would have to check each of the million cells one by one until we find the the data that is wrong. But it might be that there is actually nothing wrong with the data and we are getting different result for another reason.

In order to avoid having to check all the data one by one every time we run a test, we can simply use what is call a hash function. SHA1 is one popular option of many different hash functions. This simply a function that takes all the bytes in the data, and generate a unique number based on those bytes. So if you input the same bytes you will get the same unique number. If the file is exactly the same for both parties, the generated number must also be the same. If it is not, it means that the error for getting different result is because we have different source files.

In order to get the SHA1 sum of a file, simply run this command on a terminal of your Linux machine:

sha1sum MyFile.csv > hash.txt

This will run the sha1sum algorithm, on the "MyFile.csv" file, and save the result in the ̈hash.txt ̈file. You will obtain a string of characters similar to this:

c4d047e998dd5d3701f4ce416b4fbebcd2da37a0

If you want to compare two files, you just need to compare this short string, instead of the million cells.

Inside the folder containing the data for the project, you will find the all the data, and for each datafile, you will also find the SHA1 text corresponding to each file.

Git

Git is a software for keeping track of changes in any set of files. It can be set up for offline use in only one computer, but typically is use for file sharing over multiple computers with multiple collaborators. Git is an open-source software under GPL2.0 license

Git is overwelmely the most poppular choice in the industry. It has a light leaning curve, but once you get use to it you will not want to use anything else. You can set up your own Git private server, but typically you will use one of the many free options already available.

A shortcoming is the lack of a easy user interface for "drag and drop" files outside Windows operative systems, which can put off new potential users who don't want to deal typing commands over the terminal. It can be adverted depending of what Git service you use.

Git only handle the store of data and doesn't offer any document collaboration functionality beside sharing the file itself, so things like adding comments into the documents need to be done within the document software that you are using for that particular file. I can't think of any major writing editor that doesn't offert that functionality already, so I would call this problem as adverted.

Storage

FileSender

FileSender is a web based application that allows authenticated users to securely and easily send large files to other users. This is use to send each other private data. The data here is store for a while until it self-destroy automatically. You need a Two Factor Authentication (2FA) in order to access the files.

We use FileSender in order to share the original data as this platform is complaince with the privacy levels required

TSD

Services for sensitive data (TSD) is a platform for collecting, storing, analyzing and sharing sensitive data in compliance with the Norwegian privacy regulation. TSD is used by researchers working at UiO and in other public research institutions (the UH-sector, universities, hospitals etc.). The TSD is primarily an IT-platform for research even if in some cases it is used for clinical research and commercial research.

The complete data of FF1 is stored here. However we lost access to this at the beginning of 2021 when the TSD project "484" expired.

TSD has the advantage that can be use a virtual remote machine, which is quite combinient if you don't want to be bothered with security issues related to data privacy. This include the possibility of several people working on the same documents (without version control software) and a common space to share results securely

The disadvantages is that TSD getting something out of TSD, is a burocratic nightmare. So for example, if you want to write an article, and you want to include a figure you generated in there, you might have to wait several weeks before you are able to do so. And this will repeat each time you generate any new file.

Other minor inconviennces is that, at the time of writing this, R software was limited to Windows arquitechture only. You also need to pass periodically another burocratic layer to keep access to your project, as project expire over time

TSD however has the potential to become the best cloud-sharing option, by far. It already has in place the physical infrastructure, with the hard drives with the actual data within Norwewian borders, and software that allow for the execution of a remote virtual machine which you can use anyware you want. It just need to include it own private GIT system where you can store, and importaly retrieve, the red and black data of your project.

Local folders

Within this context, a local folder is just whatever you store in your own computer. This might or might not be syncronized over "The Cloud" somewhere else. So typically your own laptop or desktop computer. Each person related to this project has his or her own private computers. Each person can have different permission to access the private data, and as such, has different copies of such data in their own computers. For example, as time pass by, I get more and more data that somebody has to transfer (typically Anne-Sophie) from her computer, to my computer, via FileSender.

As discused in the privacy levels, all private data is not allowed to leave the UiT computers. So this, the TSD, and the FileSender, are the only places where you can find the actual original data, or cleaned data.

The biggest shortcoming of this is the data inconsistency that can arise from working with multiple data sources that are not properly managed under any version control. This however is adverted since we work with a proper logging system everytime we run an experiment.

The next problem is that we had the situation in which I was working with a collaborator (Dina) in order to analyze some data. She have permission to read the data, but I didn't. And this run on for about 3 months where I had to work blind, and preparing the scripts using fake data, or preparing it in advance for when I received the actual data.

OneDrive

This is the infrastructure provided by the UiT to share files. You can use the OneDrive software to sync files automatically here; but because is limited to a Windows machine, if you are using Linux, you need to delete and update the project folder manually via website interface.

We use this platform mostly to write the manuscripts for the scientific papers in commond and share minor folders with results.

The greatest shortcoming of this is that you can't in any way automatize the writing. So once again, everytime you change a definition, or a variable, you need to write again, manually, the entire tables that might consist of several tables with hundreds of cell each with numerical information. This is very dangereous as it introduce too much risk of an error, not to mention time consuming. This is partially adverted as I generate all figures and tables in automaticall in Latex, and later on upload manually a PDF file with the most up to date results.

Also, OneDrive doesn't support anonymous link sharing. For any external collaborator, or simply curoius minded person who want to see the results of the analysis, you need to ask persmission that need to be granted manually (by me), and in order to do so you need a mail account, and on top of that a Microsoft mail account if you want to access the full range of functionality.

OneDrives can complay with UiT's privacy of data policies, however, further twiking is required via Azure in order to setup a proper secure server for your project. Dropping red and black data in a common OneDrive folder is not allowed.

Google Drive

The infrastructure provided by Google to share files. It has the same functionality as OneDrive, except that it can also work in Linux properly. However it present the same problems without solution of any other close software for cloud store platform; plus many ethical concern regarding how Google use your own private data. Furthermore, Google Drive, unlike OneDrive, is not UiT approved for the sharing of private data.

That said, Google products have become really popular within academic fields as they are generally clean and easy to use, specially since Android has the greatest piece of mobile market and it come integrated by default with all usual Google product (to the point that you can't even coop-out from it unless you root your

own phone). So, sooner or later you are bound to find a group or collaborator that demands working with Google Drive, either because of personal preference or imposibility of using something else.

You can get read only access with this link.

Overleaf

The infrastructure provi

GitHub

Git is the software for tracking changes in any set of files. Anyone can create his own git server. GitHub offert the convinience that it has the server already setup for you and is free for the vast majority of proejects.

I don't want to deal with the hassle of maintaining a server with my own GIT server, so I use this solution as the prefered one for file sharing. Our research group already use this too as a prefered option. If you want to access as a collaborator, you need a GitHub account though, otherwise, you can download almost everthing anonymously.

The repository can be found at: https://github.com/uit-hdl/mimisbrunnr/.

What you won't find here is any of the private individual data. That is of course an issue for reproducibility that we can't overcome anyway because chances are that you don't even have permission by the authorities to look at the data. You will find however some synthetic data which you can use to test the scripts.

1.3 Mimisbrunnr overview

The project structure is quite big. It is not complicated though, and if you are use to do any software project you will find the organization very intuitive and self-explanatory, described as follow:

/**data** Contain all the data use in the project.

/**dataframeReady** CSV files that are ready to be imported to R or any other software. All these files have data that has already being clean, normalized, or any other operation that was needed.

/**fakeData** CSV files that contain data with the same structure as the original data, but is completely made up randomly, so it is impossible that can be link to anybody in real life.

/**filtersReady** CSV files that has a filter apply to them, as in people whos BMI is greater than 30 and smoke frequenly. All of these files are also clean an ready to use.

/**metaData** CSV file with variables metainformation. Such as the name of each biomarker protein.

/**biomarkers** CSV and ODS with the biomarker metainformation.

/**blood** CSV and ODS with the blood metainformation.

/**originalData** Files with the original data. Each folder contain a SHA1 subfolder (not listed here) that contain the SHA1 checksum for each file.

/**csv** CSV files with the original data directly exported without any modification.

/**verbatim** XLS, SAV and DTA files with the orignal data exactly as it was received.

/**doc** General documentation that explain nuances of the project.

1. /Data metadata: The original files don't contain human readable values, instead they have cells with fields such as "Do you smoke?", with values "1", "2" and so for. Here it is explained what each of these values means, and the possible range that they can take. It also describe several variables that are available in FF1, which may or may not be available to us.
2. /Fit Future Brochures: The original FF1 information brochores intended to captivate teenagers into collaborating in the data collection and explaining what Fit Fitures will do.

- /reports : Everything that has to do with writting a document and presenting it to someone. Includes the manuscripts, HTML documents, meeting logs, and so for.

1. /Code tutorial: How to run the R or Latex code, and several naming conventions.

2. /Latex: Everything that is written in Latex is inside this folder.
3. /Meetings: The logs of collaborators meeting, teeling what it need to be done and who is going to do it.
4. /Notebook: The generated notebooks that can run the code, notice that this is not the same as the plain HTML webs that you can find later in /Web folder.
5. /ODT: The LibreOffice text documents, generally there is nothing of intetest here as is just use for quick writing before importing it to Latex.
6. /Web: Generated simple plain website with the HTML+CSS code.

- /res : Resources folder. This contain documents and images that were not generated by the code. For examples, icons use in the website, ODG files with diagrams of the definitions we have for carrier, UiT logos for using in Latex documents, empty HMTL templates, and so.
- /results : All the results generated by the scripts. Just a lot of images, tables, and logs. It is divided by topic.
- /src : All the source code.
- LICENSE.html : License description (GPL3.0) for the GitHub web.
- README.md : Initial project description in the GitHub web.
- .gitignore : A description of what folders are not kept in the git repository. Mainly all the private data and all the result folder. You can still get the results in the report document with proper explanations.

All the public available files can be founnd in the GitHub repository, Google Drive, and OneDrive. Private data such as the patient data, or irrelevant files such as each individual table and image generated automatically, you can only find it in the local folder of each collaborator.

If you want to gain access to any of these sites, please write an email to "rca015@uit.no".

Chapter 2

Metadata

2.1 Intro

This chapter describes all the data we have in the original files. It does not cover any transformation or cleaning of the data, this will be explained in the next chapter.

2.2 Original files

The data use in the analysis comes from several different original files that we merge together. Here we describe the originins of each file and a brief description of what it contains. A detailed description for each variable is found in the next section of this document.

All of these files were converted into a more user friendly CSV format without any modification to the data itself. We build upon these CSV files later on to apply all the transformations. Notice that some of the variables are repeated and contained across these files. In here we only list the first instance in which we encounter them.

2.2.1 Basic information

Filename	Description	Date received
PERSKEY_Rafael_s aureus_19022020.xls	For the FF1 period only, all the S.Aureus information regarding direct cultures and enrichment broths, SPA types and dates of cultures. All the social network information including network representativeness grading. Some phenotypes variables: ID, sex, age, highschool information, smoking habit, snuff habits, sports habits, BMI, use of actibiotics including frequency and brand	2020/02/21
data_ut_11Juni2019.dta	For the FF1 period: Antropometry data, diseases, some medication usage, menstruation cycle, hormonal contraceptive information, some of the blood serum analysis variables, and puberty development. For the FF11 and FF12 periods, the follow up status on colonization.	2021/02/26
eutro_rafael_paakoblet.sav	For the FF1 period only. Full medication data, full blood serum, full biomarkers, household information, etchnicity, hygiene and sunbathing	2021/08/04
eutro.sav	For FF1 period, diet information. For FF12 period, follow up in the social network information.	2021/10/05
eutro.sav	For FF2 period, basic anthropometry variables (no DEXA scans available so far for any period).	2021/10/05

Table 2.1: Summary of the available Fit future data and received date.

The original file with all the metadata for the phenotype is called "20180601-Komplett Metadata FF1.xls", in here we can find 1514 different variables about many different topics. Furthermore, this file doesn't contain all the possible variables as we will see later; for example the high school ID is not described here.

The file for the S.Aureus metadata is named metadata_nasal_throat_swabs_FF1_20022020.xlsx. We only have access to a subset of all those variables.

The first file where the actual data is stored is called "PERSKEY_Rafael_s aureus_19022020.xls" , regarding the phenotypes, network, and S.Aureus. This however is a proprietary file of ".xls" format that can't be read directly by many programs without a proper transformation. To solve that, I converted that file into a ".csv" file which is a better standard for the computers. The file that I use to read all the data that we use in the scripts is located in:

"../data/aureus/csv/saureus_19022020.csv"

with SHA1:

c4d047e998dd5d3701f4ce416b4fbebcd2da37a0

The second file that contain actual info is called "data_ut_11Juni2019.dta", and this contain redundant data which we partially have in the previous file, plus the variables regarding anthropometry, medicine use, contraceptive use, and some limited biomarkers. Again, this .dta file is proprietary, and popular with Stata users. So I converted this into another csv file:

"../data/hc/csv/data_ut_11Juni2019.csv"

with SHA1:

be0593bf4d5e62bcf92ca523ced0a12225a8a02a

So far, all these data files contain data which is still "dirty". Meaning that we have numbers instead of categories for each variable (what does 1 means? man or woman?), everything is mixed together, there are a lot of missing numbers that indicates that something is unknown, and so on.

In order to clean the data, I use the script "dataCleaning.R". Later on I will explain what this script does specifically in another section of this document. Once the data is "clean" and we filter out the values that we don't want, then we can proceed with our analysis.

For now, let focus on describing the actual variables we have in these files. The following tables are divided by topic, so variables related to the same topic are in the same table.

2.3 Data description

2.3.1 Personal key

Each individual have a personal key described in the "pers_key_ff1" variable. The original key looks like this: "12345678" and is simply a 8 digit unique key for each of the 1038 individuals in that file.

2.3.2 Basic information

Name	Description
SEX_FF1	Sex
AGE_FF1	Age at the time of screening

Table 2.2: Table with the original data for the basic information variables. Note that we don't have a date of birth, so the age variable is a very limited numerical variable.

2.3.3 School and education

Name	Description
HIGH_SCHOOL_NAME_FF1	School ID (Not described in the metadata.)
HIGH_SCHOOL_CLASS_FF1	Class ID (Not described in the metadata.)
HIGH_SCHOOL_PROGRAMME_FF1	Study subprogram (Not described in the metadata.)
HIGH_SCHOOL_MAIN_PROGRAM_FF1	Main high school program (vocational program)

Table 2.3: Table with the original data for the education variables.

2.3.4 Recreational Drugs

Name	Description
SMOKE_FF1	Do you smoke?
SNUFF_FF1	Do you use snuff?
ALCOHOL_FREQUENCY_FF1	How often do you drink alcohol?

Table 2.4: Table with the original data for the recreational drugs variables.

2.3.5 Physical Activity

Name	Description
PHYS_ACT_LEISURE_FF1	Exercise and physical exertion in leisure time. If your activity varies much, for example between summer and winter, then give an average. The question refer only to the last twelve months.
PHYS_ACT_OUTSIDE_SCHOOL_FF1	Are you actively doing sports or physical activity (e.g. skateboarding, football, dancing, running) outside school hours?

Table 2.5: Table with the original data for the physical information variables.

2.3.6 Anthropometry

Name	Description
HEIGHT_FF1	Body height in cm measured at screening
WEIGHT_FF1	Body weight in kg measured at screening
WAIST1_FF1	Waist circumference, first measurement (cm)
HIP1_FF1	Hip circumference, first measurement (cm)
WAIST2_FF1	Waist circumference, second measurement (cm)
HIP2_FF1	Hip circumference, second measurement (cm)
BMI_FF1	BMI at the time of screening (Not described in the metadata.)

Table 2.6: Table with the original data for anthropometric variables. Note that in the TSD we have extended information about the DXA scans.

2.3.7 Medicines

Name	Description
ANTIBIOTICS_FF1	Have you taken any antibiotics (tablets or oral suspensions, nasal ointments, eye drops or eye ointment applicated in the nose/eye) the last 24 hours?
ANTIBIOTICS_BRAND1_FF1	If you have taken any antibiotics the last 24 hours, what brand (inc.strength) did you take?
ANTIBIOTICS_ATC1_FF1	If you have taken any antibiotics the last 24 hours, ATC-code?
ANTIBIOTICS_BRAND2_FF1	If you have taken any antibiotics the last 24 hours, what brand (inc.strength) did you take?
ANTIBIOTICS_ATC2_FF1	If you have taken any antibiotics the last 24 hours, ATC-code?
MEDICATION_DAILY_FF1	Do you take any medicine daily or regularly?
MEDICATION_BRAND1_FF1	If you take any medication, what brand (inc.strength) do you take - 1?
MEDICATION_ATC1_FF1	If you take any medication, ATC-code - 1?
MEDICATION_REGULAR1_FF1	How frequent do you take the medication - 1?
– Rest of regular medicines –	The previous rows repeat for Regular medicines 1, 2, 3, 4, and 5.
MEDICATION_OTHER_FF1	If you take any medication, unknown or not listed brand?
MEDICATION_OTHER_DESC_FF1	If you take any medication, unknown or not listed medicine description

Table 2.7: Table with the original data for medicine intake related variables. Contraceptives are also medicine, but we explore then in a different section specifically through out the whole document.

2.3.8 Diseases

Name	Description
CHRONIC_DISEASE_FF1	Do you have any chronic or persistent disease?
DIAGNOSIS_CHRONIC_DISEASE1_FF1	If you have any chronic or persistent disease, what diagnosis - 1?
ICD10_CHRONIC_DISEASE1_FF1	If you have any chronic or persistent disease, ICD10 -code 1?
– Rest of Chronic Diseases –	The previous rows repeat for Chronic Disease 1, 2, 3, 4, and 5.
CHRONIC_DISEASE_OTHER_FF1	If you have any chronic or persistent disease, other symptoms?
CHRONIC_DISEASE_OTHER_DESC_FF1	If you have any chronic or persistent disease, other chronic symptom description?
DIABETES_FF1	Do you have diabetes?
ICHY_SKIN_FF1	Have you had ichy skin rash during the last 12 months?
ICHY_SKIN_LOCATION_FF1	If you have had ichy skin rash during the last 12 months, did the skin rash affect the following locations: round your neck, around your ears or eyes, in the crook of your elbows, on your bottocks, behind your knees, or at the front of your ankles?
PSORIASIS_LIFETIME_FF1	Do you have or have you ever had psoriasis?
PSORIASIS_SEVERITY_FF1	If you have or have ever had psoriasis, how severe is your psoriasis today? Please, indicate on a scale from 0 (no disease symptoms) to 10 (most severe disease symptoms).
ALLERGIC_RHINITIS_FF1	Have a doctor ever said that you have hay-fever or allergic rhinitis?
ASTHMA_FF1	Have a doctor ever said that you have asthma?
ATOPIC_ECZEMA_FF1	Have a doctor ever said that you have children's eczema or atopic eczema?

Table 2.8: Table with the original data for diseases related variables.

2.3.9 Menstrual information

Name	Description
MENSES_FF1	Have you started menstruating?
MENSES_REGULARITY_FF1	If you have started menstruating; how regular are your periods?
MENSES_CYCLE_LENGTH_FF1	If you have started menstruating and your cycles always or usually are regular; what is the usual number of days between start of each period?
MENSES_START_DATE_CERTAIN_FF1	If you have started menstruating; do you know the date of the start of your last menstruation?
MENSES_START_DATE_FF1	If you have started menstruating and if you know the date of your last menstruation; what was the date of the first day of your last menstruation??
CHANCE_PREGNANT_FF1	If you have started menstruating; Is there any chance that you may be pregnant?
PREGNANCY_TEST_RESULT_FF1	If pregnancy consent; - pregnancy test result

Table 2.9: Table with the original data for the menstrual variables.

2.3.10 Puberty

Name	Description
MENARCHE_FF1	Girls: have you started menstruating?
MENARCHE_AGE_YEAR_FF1	Girls: if you have started menstruating, how old were you when you had your first menstrual period? Years
MENARCHE_AGE_MONTH_FF1	Girls: if you have started menstruating, how old were you when you had your first menstrual period? Months
PUBIC_HAIR_FEMALE_FF1	Girls: if you have not started menstruating, have you got or started to get pubic hair?
BREASTS_FEMALE_FF1	Girls: if you have not started menstruating, have your breasts enlarged or started to enlarge?
PUBIC_HAIR_MALE_FF1	Boys: have you got or started to get pubic hair?
PUBIC_HAIR_AGE_MALE_FF1	Boys: if you have got or started to get pubic hair, how old were you when you started to get pubic hair?
PUBERTY_BOYS_HEIGHT_FF1	Boys: Would you say that your growth in height,
PUBERTY_BOYS_HAIR_BODY_FF1	Boys: Would you say that your body hair growth,
PUBERTY_BOYS_VOICE_FF1	Boys: Have you noticed a deepening of your voice?
PUBERTY_BOYS_HAIR_FACE_FF1	Boys: Have you begun to grow hair on your face?

Table 2.10: Table with the original data for the puberty variables.

2.3.11 Nutrition

Name	Description
FAT_FISH_FF1	How often do you usually eat fat fish (e.g. salmon, trout, mackerel, herring)
LEAN_FISH_FF1	How often do you usually eat lean fish (e.g. cod, saithe, haddock)
SEAGULL_EGGS_FF1	How often do you usually eat seagull eggs?
REINDEER_FF1	How often do you usually eat reindeer meat?

Table 2.11: Table with the original data for diet and nutrition. The TSD variable contain about 75 variables in total regarding eating habits.

2.3.12 Biomarkers

Name	Description
S_ESTRADIOL_FF1	Serum estradiol, E2 (nmol/L)
S_PROGESTERONE_FF1	Serum progesterone (nmol/L)
S_TESTOSTERONE_FF1	Serum testosterone (nmol/L)
S_SHBG_FF1	Serum sex hormone binding globuline (SHBG) (nmol/L)
S_LH_FF1	Serum luteinizing hormone (LH) (IU/L)
S_FSH_FF1	Serum follicle-stimulating hormone (FSH) (IU/L)
S_HBA1C_FF1	Glycated haemoglobin (%). EDTA whole blood
ALBUMIN_FF1	Albumin (g/L). Serum
S_25_VITD_FF1	25(OH)D (nmol/L). Serum
S_TESTOSTERON_LCMSMS_FF1	Serum testosterone (nmol/L), analyzed by LC-MSMS
S_ANDROSTENDION_LCMSMS_FF1	Serum androstendione (nmol/L), analyzed by LC-MSMS
S_17OHPROG_LCMSMS_FF1	Serum 17-hydroxyprogesterone (nmol/L), analyzed by LC-MSMS
S_PROGESTERON_LCMSMS_FF1	Serum progesterone (nmol/L), analyzed by LC-MSMS
S_ESTRADIOL_LCMSMS_FF1	Serum estradiol (pmol/L), analyzed by LC-MSMS
S_E2_BELOW_LIMIT_FF1	Serum estradiol below 0,10 nmol/L
S_PROG_BELOW_LIMIT_FF1	Serum progesterone below 1 nmol/L
S_SHBG_ABOVE_LIMIT_FF1	Serum sex hormone binding globuline (SHBG) above 200 nmol/L
S_LH_BELOW_LIMIT_FF1	Serum luteinizing hormone below 0,5 IU/L
S_FSH_BELOW_LIMIT_FF1	Serum follicle-stimulating hormone below 0,5 IU/L
S_SHBG_ABOVE_LIMIT_FF1	Serum sex hormone binding globuline (SHBG) above 200 nmol/L
S_PROG_BELOW_LMT_LCMSMS_FF1	Serum progesteronoe below 0,5 (nmol/L), analyzed by LC-MSMS
S_ESTR_BELOW_LMT_LCMSMS_FF1	Serum follicle-stimulating hormone below 0,5 IU/L

Table 2.12: Table with the original data for the biomarkers information variables. Note that in the TSD we have extended information regarding fatty acids, iron related variables, calcium, and much more.

2.3.13 Contraceptives

Name	Description
CONTRACEPTIVES_FF1	If you have started menstruating; do you use any kind of contraceptives?
CONTRACEPTIVES_TYPE_FF1	If you use any kind of contraceptives; what types?
ORAL_CONTRACEPT_NAME_FF1	If you use any oral contraceptive pill, what is the name of the medicine?
INJECTED_CONTRACEPT_NAME_FF1	If you use any injected contraceptive, what is the name of the medicine?
SUBDERMAL_CONTRACEPT_NAME_FF1	If you use any hormonal contraceptive subdermal implant, what is the name of the medicine?
CONTRACEP_SKIN_PATCH_NAME_FF1	If you use any hormonal contraceptive skin patch, what is the name of the medicine?
VAGINAL_CONTRACEPT_NAME_FF1	If you use any vaginal contraceptive ring, what is the name of the medicine?
ORAL_CONTRACEPT_ATC_FF1	If you use any oral contraceptive pill, what is the ATC-code of the medicine?
INJECTED_CONTRACEPT_ATC_FF1	If you use any injected contraceptive, what is the ATC-code of the medicine?
SUBDERMAL_CONTRACEPT_ATC_FF1	If you use any hormonal contraceptive subdermal implant, what is the ATC-code of the medicine?
CONTRACEP_SKIN_PATCH_ATC_FF1	If you use any hormonal contraceptive skin patch, what is the ATC-code of the medicine?
VAGINAL_CONTRACEPT_ATC_FF1	If you use any vaginal contraceptive ring, what is the ATC-code of the medicine?

Table 2.13: Table with the original data for the use of contraceptives variables. The contraceptives are linked only to girls who started menstruating.

2.3.14 Sociology

Name	Description
HOUSHOLD_SIBS1TO2_FF1	Who do you live with now: 1-2 siblings?
HOUSHOLD_SIBS3PLUS_FF1	Who do you live with now: 3 or more siblings?
HOUSHOLD_FRIENDS_FF1	Who do you live with now: Friends?

Table 2.14: Table with the original data for the basic information variables. The TSD contain hundreds more variables.

2.3.15 Network

Name	Description
FRIEND_1_FF1	Which students have you had most contact with the last week? Name up to 5 students at your own school or other schools in Tromsø and Balsfjord.
FRIEND1_PHYSICAL_CONTACT_FF1	Do you have physical contact?
FRIEND1_CONTACT_SCHOOL_FF1	Are you together at school?
FRIEND1_CONTACT_SPORT_FF1	Are you together at sports?
FRIEND1_CONTACT_HOME_FF1	Are you together at home?
FRIEND1_CONTACT_OTHER_FF1	Are you together at other places?
– Rest of friends –	The previous rows repeat for FRIEND2, FRIEND3, FRIEND4 and FRIEND5
NETWORK_DATE_FF1	Network date (When the interview for the network was recorded).
NETWORK_SIGNATURE_FF1	Network signature (Who recorded the interview).
NETWORK_OVERVIEW_FF1	To what degree does this table of friends give an overview of your social network? Please, indicate on a scale from 0 (small degree) to 10 (high degree).
NETWORK_COMMENT_FF1	Comments network - friends. For example "Ingen kontakt pga vinterferie."

Table 2.15: Table with the original data for the network variables.

2.3.16 S.Aureus

Name	Description
DATE_CULTURE_DAY0_FF1	Nasal and Throat swabs: Date of culturing in the laboratory.
CONTROL_NASAL_DAY2_FF1	Nasal swab: Any growth of bacterial colonies on control agar plate.
CONTROL_THROAT_DAY2_FF1	Throat swab: Any growth of bacterial colonies on control agar plate.
STAPH_NASAL_DAY2_FF1	Nasal swab: Any growth of bacterial colonies on Staphylococcus aureus selective agar plate.
STAPH_GROWTH_NASAL_DAY2_FF1	Nasal swab: Classification of growth of bacterial colonies on Staphylococcus aureus selective agar plate.
STAPH_THROAT_DAY2_FF1	Throat swab: Any growth of bacterial colonies on Staphylococcus aureus selective agar plate.
STAPH_GROWTH_THROAT_DAY2_FF1	Throat swab: Classification of growth of bacterial colonies on Staphylococcus aureus selective agar plate
STAPH_NASAL_ENRICH_FF1	Nasal swab in enrichment broth: Any growth on Staphylococcus aureus selective agar plate after enrichment
STAPH_GROWTH_NASAL_ENRICH_FF1	Nasal swab in enrichment broth: Classification of growth of bacterial colonies on Staphylococcus aureus selective agar plate after enrichment.
STAPH_THROAT_ENRICH_FF1	Throat swab in enrichment broth: Any growth on Staphylococcus aureus selective agar plate after enrichment
STAPH_GROWTH_THROAT_ENRICH_FF1	Throat swab in enrichment broth: Classification of growth of bacterial colonies on Staphylococcus aureus selective agar plate after enrichment
STAPH_COAGULASE_NASAL_FF1	Nasal swab: Coagulase test.
STAPH_COAG_NASAL_ENRICH_FF1	Nasal swab in enrichment broth: Coagulase test.
STAPH_COAGULASE_THROAT_FF1	Throat swab: Coagulase test
STAPH_COAG_THROAT_ENRICH_FF1	Throat swab in enrichment broth: Coagulase test

Table 2.16: Table with the original data for the S.Aureus variables. 1/2

Name	Description
SPA_THROAT1_FF1	Throat swab: Spa-type of S. aureus isolate.
CCN_THROAT1_FF1	(Not described in the metadata).
CC_THROAT1_FF1	Throat swab: S. aureus clonal complex based on spa-type.
SPA_NASAL1_FF1	Nasal swab: Spa-type of S. aureus isolate.
SPA_NASAL2_FF1	Second nasal swab: Spa-type of S. aureus isolate.
SPA_THROAT2_FF1	(Not described in the metadata).

Table 2.17: Table with the original data for the S.Aureus 2/2

Bibliography

[1] "Bongo, lars-ailo." https://uit.no/ansatte/person?p_document_id=66818, Jun. 2022.

[2] "Furberg, anne-sofie." https://uit.no/ansatte/person?p_document_id=168042, Jun. 2022.

[3] "An introduction to latex." https://www.latex-project.org/about/, Jun. 2022.

[4] "Wikipedia: What you see is what you get." https://en.wikipedia.org/wiki/WYSIWYG, Jun. 2022.

[5] "Ibm spss statistics." https://www.ibm.com/products/spss-statistics, Jun. 2022.