UPDATE 8.2.2017

DATA CLEANING
I had cleaned the data, mostly with Alaitz code that she implemented based on the instructions.
For discrete variables, the values 5555, 6666, 7777, 8888, 9999 were converted to NA.
For continuous variables, total cholesterol(skol), Triglycerides(stg), blood pressure(sbt and dbt), ldl and hdl cholesterol were corrected due to different measurement techniques and medication use.
Values for variables langd, bmi, vikt, midja, skol, hdl, stg, blods0, blods2, sbt and dbt were checked and set to NA for those out of the limits specified, numbers of subjects out of limits for each variable:

bmi	49
langd	22
vikt	5
midja	97
skol	7
hdl	400
stg	9769
blods0	56
blods2	1028
sbt	0
dbt	1

Additionally I excluded those that had insufficient or implausible diet data.
I had excluded 3343 subjects that had the variable exclude==1, which indicates insufficient diet data.
For the biologically implausible energy intake, I had excluded subjects based on the variable FIL, where I had excluded the bottom 5% of the distribution, resulting in 7441 subjects and the top 2.5% of the distribution, resulting in 3534 subjects.
After the merge and correction of the besok with the missing besok, the dataset had 168327 rows.
After the exclusion due to insufficient or implausible diet data, the dataset had 154009 rows.
The cleaned dataset is called VIP_170206_cleaned.csv in the VIP_data folder on the server.

EXTRACTING SUBJECTS WITH TWO VISITS APPROXIMATELY 10 YEARS APPART

I had created a subset of subjects, which have two visits apart by either 9,10 or 11 years.
I had used the cleaned dataset (VIP_170206_cleaned) with the merged and corrected besok, using the besok1:

-I started by looking at those that have the besok1==1 and besok1==2 and there are **33906** subjects there.

-Then I looked if there are any subjects having besok1==2 and besok1=3 with 9-11 years difference and are not already in the previous subset and there were **584** subjects there, which I added to the subset.

-Then I looked if there are any subjects having besok1==1 and besok1=3 with 9-11 years difference and are not already in the previous subset and there were **8** subjects there, which I added to the subset.

-There were no appropriate subjects or subjects were already in the subset for the besok1==2 and besok1==4, same for besok1==1 and besok1==4 and same for besok1==3 and besok1==4.

So all together there are **34495** subjects that have two visits with a time difference between 9 and 11 years and most of them the first visit is the actual first visit.

I had saved the subset as VIP_170206_cleaned_subset.csv in the VIP_data folder on the server.
I had also created a list of enummers(unique row ids) for the "Visit 1" and "Visit 2" of each subject and saved it as Visit1_Visit2_enummers.csv in the folder two_visits_subset on the server.

VARIABLE SUMMARIES

I had added the variable summaries for the subset only. The table including all these data is called VIP_variable_description_summary(.csv, .ods, .xlsx) and it is on the server, in the folder Documents.
In case any body forgot, the table includes the variable names, description, type(discrete or continuous), summary for all, summary for subset. Summaries are counts for discrete, including number of missing values and for continuous the summaries are min, max, mean, std and the number of missing values.