**UMEA BIOBANK DATASETS DOCUMENTATION**



Right now we have four datasets with information obtained through Umea Biobank: GLACIER (glacier\_corr\_170124.csv), VIKING (viking\_vip\_jal\_141020\_DDB.txt and VIKING.csv), EWAS (EWAS\_data.RData) and Multimodality (VIP\_161102.csv). Each of them are located in purple under the folder with their name.

All these datasets contain both baseline and follow-up visits available at the date of obtaining the data. The description of the variables found in these datasets is in two files (VIP\_variable\_description.docx and Diet\_variable\_description.doc). This and additional information can be downloaded from <http://www.biobank.umu.se/biobank/biobank---for-researchers/access/biobank-research---nutrition-database/>

**Cross-sectional or longitudinal analysis**

A single observation per participant should be used unless change over time is of interest. If more than one visit to VHU/VIP is available for a participant, the order is indicated by the visit variable (*besok*). Multiple visits can also be ordered chronologically using the participant-unique id variable (*id*) and year of visit (*pyear*).

NOTE: a few participants lack a formal baseline visit (besok=1), and start with ‘*besok’* equal to two or three.

**Some details to think about when using this data:**

1. Some of the variables will have the following values:

6666= unable to interpret given answer

7777= questionnaire is missing

8888= question is not represented within current questionnaire

9999= question has not been answered

Convert all these values to missing.

1. Blood pressure and lipid (triglycerides and cholesterol) values should be corrected for VIKING, EWAS and Multimodality datasets (not for GLACIER dataset since these corrections were already done for this dataset) as the method for measuring blood pressure and lipids changed after September 1st 2009.

For blood pressure measurements the method changed from measuring participants in a supine position to a sitting position. For lipid measurements the method changed from Reflotron to a clinical chemical analysis at the laboratory. The conversion is needed so values before and after September 1st 2009 (in some datasets a variable efter\_090901 could be used to identified these measurements) could be comparable. See Tables 1 and 2 for conversion algorithms.

1. LDL values can be calculated by using Friedewald formula [1]:

LDL = TC – HDL – (TG/5)] - if lipids are measured in mg/dl

LDL = TC – HDL – (TG/2.2)] - if lipids are measured in mmol/l

1. In all the datasets, blood pressure and lipid values should also be corrected for medication (See variable C5a for blood pressure medication and variable C5e for lipids medication).

* Please correct blood pressure values with a constant listed in Table 3.
* Please correct lipid values with a constant listed in Table 4.
* After applying corrections check that LDL doesn´t have negative values and if it has remove the negative values.

1. Exclusions due to deviating values should be done. There are limits proposed by VHU, see Table 5. Using sd and excluding outliers may be necessary as well.
2. Glucose and lipid values (blods0, blods2, skol, stg, HDL) are observed after a certain number of fasting hours given in the variable fasta. Hence unless only individuals fasting for a certain number of hours are used, fasta should be used as the adjustment variable in the analysis.
3. When working with diet variables:

7.1. Exclusion based on the quality of the FFQ: It is necessary to exclude diet information from participants based on the quality of the completed FFQ using the variable 'Exclude’. For the cleanest dataset possible only individuals with Exclude= 0 should be kept. However as this could lead to the loss of too many participants, individuals with exclude = 0 or 2 can be kept in the dataset, but individuals with exclude= 1 should be excluded in all cases, see also Appendix 1 for more details.

7.2. Exclusion based on biologically implausible energy intake. Two ways of doing this:

- Based on total energy intake: Exclude participants with TEI < 500 kcal/day or > 4500 kcal/day.

- Based on FIL variable: FIL = TEI/estimated basal metabolic rate. Guideline is to exclude diet information from participants with the bottom 5% and top 2.5% of the FIL distribution.

Comment: The second way to exclude based on implausible energy intake is the preferred way to exclude.

7.3. Working with dietary variables, we should be aware that in 1996 the FFQ was reduced from including 84 food items to 66 items, which makes a bit of difference for some of the diet variables. Thus, analyses using dietary variables should be adjusted for FFQ version. This adjustment can be made by using ‘enkver2’ variable which contains three categories. 'short', 'long' and 'apri'. Short = 66-68 item FFQ, long = 84 item FFQ. 'apri' is the same as 'long', the only difference is that long is the newer version that was optically read, whereas apri is the old one that was manually entered. So, enkver2 variable can be used as it is to do the adjustments or combine both ‘long’ and ‘apri’ in one category.

**NOTE: Cholesterol – important background information**

Cholesterol was initially measured based on the initiative of individual primary health care centers/physicians (i.e. not as a result of a general VIP protocol), and as a result of a clinical diagnosis ‘justifying’ that it was measured. When the concept of the ‘metabolic syndrome’, as a sign of increased cardiometabolic risk, gained ground it was decided that a full lipid status profile should be performed on everyone who fulfilled the criteria. This was enforced (at the clinical level) around 2003.

Initially lipids were measured on the Reflotron. With time, the frequency of these measurements increased as a result of increased awareness of the relevance of lipids for clinical risk assessment. On their own initiative, a lot of health care centers started to send the samples to the hospital chemical laboratory for analysis, as the general feeling was it took too much time to calibrate the Reflotron. To harmonize the lipid measurements it was thus decided in 2009 that all lipid measurements were to be done by the hospital chemical laboratory, and that this should be done on all participants (starting 1 Sept 2009).

This implies that the initial group of participants with HDL measured might not be that good a representative of the general population at large …

**Table 1. Conversion algorithms for sitting/supine measured blood pressure**

|  |  |  |  |
| --- | --- | --- | --- |
| 40 yr | Men | Sitting systolic BP | 21.612 + (0.835 x Supine systolic BP) |
|  |  | Supine systolic BP | 24.595 + (0.792 x Sitting systolic BP) |
|  |  |  |  |
|  |  | Sitting diastolic BP | 14.463 + (0.848 x Supine diastolic BP) |
|  |  | Supine diastolic BP | 17.282 + (0.753 x Sitting diastolic BP) |
|  |  |  |  |
|  | Women | Sitting systolic BP | 19.922 + (0.830 x Supine systolic BP) |
|  |  | Supine systolic BP | 8.669 + (0.919 x Sitting systolic BP) |
|  |  |  |  |
|  |  | Sitting diastolic BP | 13.680 + (0.847 x Supine diastolic BP) |
|  |  | Supine diastolic BP | 5.784 + (0.890 x Sitting diastolic BP) |
|  |  |  |  |
|  |  |  |  |
| 50 yr | Men | Sitting systolic BP | 19.748 + (0.861 x Supine systolic BP) |
|  |  | Supine systolic BP | 9.850 + (0.910 x Sitting systolic BP) |
|  |  |  |  |
|  |  | Sitting diastolic BP | 13.390 + (0.878 x Supine diastolic BP) |
|  |  | Supine diastolic BP | 12.363 + (0.812 x Sitting diastolic BP) |
|  |  |  |  |
|  | Women | Sitting systolic BP | 12.723 + (0.906 x Supine systolic BP) |
|  |  | Supine systolic BP | 16.051 + (0.859 x Sitting systolic BP) |
|  |  |  |  |
|  |  | Sitting diastolic BP | 17.675 + (0.800 x Supine diastolic BP) |
|  |  | Supine diastolic BP | 13.566 + (0.798 x Sitting diastolic BP) |
|  |  |  |  |
| 60 yr | Men | Sitting systolic BP | 20.246 + (0.853 x Supine systolic BP) |
|  |  | Supine systolic BP | 7.763 + (0.936 x Sitting systolic BP) |
|  |  |  |  |
|  |  | Sitting diastolic BP | 16.308 + (0.833 x Supine diastolic BP) |
|  |  | Supine diastolic BP | 9.029 + (0.864 x Sitting diastolic BP) |
|  |  |  |  |
|  | Women | Sitting systolic BP | 13.817 + (0.900 x Supine systolic BP) |
|  |  | Supine systolic BP | 9.999 + (0.914 x Sitting systolic BP) |
|  |  |  |  |
|  |  | Sitting diastolic BP | 15.084 + (0.836 x Supine diastolic BP) |
|  |  | Supine diastolic BP | 7.992 + (0.870 x Sitting diastolic BP) |

The conversions are made using the following age ranges: 35-44 years for 40 yr category, 45-54 years for 50 yr category and 55-64 years for 60 yr category.

The algorithms do not specify what to do with participants around 30 or 70 years of age. If the number of participants in these two age groups is low they can be excluded. If the number of participants measured after 2009 is small analyses can be restricted to those measured before 2009. In some circumstances applying the 40 yr category algorithm for individuals around 30 years and the 60 yr category algorithm for participants around 70 yr could be discussed.

**Table 2. Conversion algorithms from Reflotron/laboratory measurement in lipid traits**

TG=Triglyceride

Chol= Cholesterol

|  |  |
| --- | --- |
| S-TG – **Reflotron** | 0.177 + (0.932 x S-TG - **Clin Chemistry**) |
| S-Chol – **Reflotron** | 0.170 + (0.939 x S-Chol - **Clin Chemistry**) |

**Table 3. Correction constants for blood pressure**

|  |  |  |
| --- | --- | --- |
| Anti-hypertensive or blood pressure lowering medication | SBP | + 15 mmHg |
| DBP | + 10 mmHg |

**Table 4. Correction constants for lipid medication**

|  |  |  |
| --- | --- | --- |
| HMG-CoA reductase Inhibitors (statins) | Total cholesterol | + 1.347 mmol/l |
| Triglycerides | + 0.208 mmol/l |
| HDL | -0.060 mmol/l |
| LDL | +1.290 mmol/l |

**Table 5. Limit values proposed by VHU (VIP) for exclusions:**

Langd (height): <130 cm or >210 cm

Vikt (weight): <35 kg

Bmi: <15 or >70

Midja (waist circumference): <60 cm

Skol (total cholesterol): <0.5 or >15

Hdl: <0.15 or>7

Ldl: Not defined

Stg[[1]](#footnote-1): <0.15 or >20

Blods0: <1[[2]](#footnote-2) or >25

Blods2: <1[[3]](#footnote-3) or >35

Sbt: <20 or >300

Dbt: <20 or >250

Additional exclusions in footnotes

**Appendix 1**

Exclude is explained in the NSDD-variable list and indicates level of insufficiently reported diet data:

* 0 = Complete set of portion size indications and < 10% of food frequencies missing
* 1 = > 10% of food frequencies missing
* 2 = the set of portion size indications is not complete

Observe: The exclusions you need to make depend on the research question you are asking, but Exclude = 0 will give you the ‘cleanest’ dataset when working with diet data.

**References**

[1] Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clinical chemistry 18: 499-502

1. 1STG measurements <0.8 should be additionally excluded due to the sensitivity of the Reflotron benchtop analyzer. [↑](#footnote-ref-1)
2. VHU guideline is less than 1, but values <2 should be additionally excluded since they are biologically implausible values, person would be in a coma. [↑](#footnote-ref-2)
3. VHU guideline is less than 1, but values <2 should be additionally excluded since they are biologically implausible values, person would be in a coma. [↑](#footnote-ref-3)