**Draft: Questions to consider when defining study population for studies based on the BHLL questionnaire**

*When defining the study population for your BHLL study, we suggest that you consider excluding for (at least) these four reasons:*

1. Completeness of the answered questionnaires
2. Residency in Denmark / availability of SEP-data
3. Questionable answers
4. Study-specific exclusions
5. **Completeness of the answered questionnaires**

114,555 people with partly or fully answered the questionnaire. (Besides that, 876 started the questionnaire, but didn’t answer any question. These are excluded per default).

* + 96,423 completed the questionnaire, and 18,132 (16%) only partly completed the questionnaire. (NB! Even in “fully” answered questionnaires there can still be missing values in some of the fields (like waist circumference)).
    - If you are mainly interested in some of the earlier questions, it is possible to identify all who answered until a given question in the questionnaire. Among these 18,132 respondents who only partly finalized their questionnaire:
      * 6,667 Have stopped at question number 12, or earlier
      * 8,776 Have stopped at question number 15, or earlier
      * 9,694 Have stopped at question number 65, or earlier
      * 14,215 Have stopped at question number 75, or earlier

1. **We suggest exclusion of 159 (0.1%) who had not been resident in Denmark at 31dec2020.**

Background: The questionnaire was submitted based on residency in Denmark June 30th, 2021. At DST (project 708491 and 709876) the general sociodemographic data (e.g., cohabitation, residency, region from the bef-files) and socioeconomic measures (e.g. education, occupation and income) are only available on December 31st each year. A very few people can have immigrated within first half of 2021, answered the questionnaire, and then emigrated or died before December 31st, 2021, leaving us without possibility to get socioeconomic data on them. [And possible not full follow up for death?]

We suggest the exclusion of the 159 respondents that was not resident in Denmark at 31dec2020.

(103 of the 159 can be found in the bef-file from 31dec2021, but including those, and not the remaining 56 person, could possibly lead to a bias)

As a positive side effect, limiting the cohort to only those living in Denmark December 31st, 2020, will have an additional benefit that it secures a minimum of 8 months of lookback within LPR, LMDB, LABF etc. (the 8 months from January to September 2021 – or typically a bit longer when the response-date was later).

* In addition, the population can be limited further to people who did not immigrate within one year prior to answering the questionnaire (to ensure at least one year of follow-up). (Another ~49 questionnaires (on top of the 159) will be excluded by this).
  + Requiring 5 years of FU: Another 1059 questionnaires will be excluded.

1. **Questionable answers:**

Handling mismatch of sex and age between questionnaire and register data. (In case of missing age or sex, registry data are used).

* + Age-mismatch: We find 366 responses with a major disagreement between self-reported **age** compared to Danish Civil Registration System (a difference of more than one year, or age missing) among the 114,555 answers. Indicating that the questionnaire might be filled by a wrong person, or a person who on purpose wanted to insert wrong data. (We find 245 among the 96,423 with a completed questionnaire)
    - Insisting on an exact match of age will remove another 4202 participants though! This could be caused by people starting the questionnaire on one date, and finalizing it later, or simply not remembering their exact age or mistyping it. We suggest accepting a difference in one year.
  + 134 with disagreement between self-reported **sex** compared to Danish Civil Registration System or missing self-reported sex. Indicating that the questionnaire might be filled by a wrong person, or a person who on purpose wanted to insert wrong data. But on the other hand, it could also result in excluding a few people perceiving themselves as having another gender than what is registered in Statistics Denmark (though taking the age group 50-65 into account, we expect those numbers to be low).

1. **Studyspecific exclusions**

Will depend on study. Some examples of general and study specific exclusion from previous studies:

* **Example** of exclusions from the study in obesity and self-reported-health:
* Missing BMI (or implausible registrations for weight, height or BMI)
* BMI<18.5 (underweight) [Because this study focused on obesity vs. normal weight]
* Lookback <1 yr
* Missing self-reported health

🡪We end up with 88,399 valid records (among these there were 222 respondents (0.25%) who didn’t finalize the questionnaire).

* **Example** (cohort-profile paper)

“We excluded respondents who answered none of the questions as well as those who provided no information on sex or indicated an age other than 50–65 years.”

* Excluding 92 respondents who didn’t answer the sex-question
* Excluding another 180 who didn’t answer the age question, or where the age wasn’t in the interval 50-66 years old.

🡪 Resulting in a population of 114.283 respondents.

* **Example** (Metabolic Syndrome).

Measure of Waist circumference is one of the components of metabolic syndrome but also has low completeness. We ended up with

1. A main cohort (with valid waist-measurement): 50,736
2. A comparison cohort of persons without a valid waist-measurement for sensitivity analyses: 63,186