Lifelines data catalogue manual

With the help of this manual you will be able to select the data you would like to use for your research. The catalogue may be a bit overwhelming at first because of the various filters and the amount of data variables, but with this manual you will be able to familiarize yourself with the various functions and make your selection.

Before you start:

- You do not need to make an account or sign in when you are just browsing around. If you want to fill your cart and order your selection, signing in is required.
- If you are new to Lifelines, we recommend that you check out our wiki (just follow the "more info" links in the catalogue) to learn about the basic structure of our cohort and data collection, as well as the meaning of certain terms and codes.
- We recommend that you first spend some time orienting yourself - play around with the filters and functions, locate the variables you are interested in - without actually filling your cart. Once you are ready to order an actual dataset, please start with an empty cart and follow the instructions below ("**ordering**"). Avoid changing the filters while you are selecting the variables for your dataset in your cart.

FAQ:

- **What is the function of the participant and assessment filters on the left?**
  They have several functions:
  - It will give you information about participant numbers (total and per variable per assessment) within certain age, sex, and/or subcohort groups
  - It will make the variable-assessment matrix on the right more manageable by filtering out information that you are not interested in
  - When you order a dataset, it ensures you will only get the participants and assessments that you would like to use for your research (-> easier to manage and compliant with data minimalization).

  If you do not wish to select specific participant groups or assessments (i.e. you "want it all") and this fits with your research proposal, then please do not click any buttons here, just leave it as it is. You can hide the filters with the little arrows.

- **How does the Age filter work?**
  Here you can select the age group(s) you are interested in.
  Note that in earlier releases the 18-64 and 65+ groups were combined, and now they are separated. If you want all adults, please select both (18-64 ánd 65+) buttons.
  Note that the "age at" filter works a bit tricky. It can be used to select participants who fall into a certain age group during a given assessment, as defined by both a) their actual age during the entire assessment and b) the age-dependent questionnaires they received for that assessment. It does not select the (relatively few) "ambiguous" participants (i.e. participants who went from one age group to the next in the middle of an assessment, or who received age-dependent

questionnaires not matching their actual age). If you need these participants as well, then do not use this "age at" filter.
Alternatively, you may use the year-of-birth filter to target the proper age group.
NB: you can recognize the age group at which a certain variable was aimed via the subject part in the variable code (-> chi, ach etc., see list at the end of this document).
NNB: in your final dataset, you will receive the age at the time of measurement (in months) for each participant and each measurement.

- **What is the "sec" assessment?**
  The "sec" assessment is, in reality, not an assessment. It is a collection of variables that cannot be pinpointed to a single assessment, either because it is a derivative based on variables from multiple assessments, or because it is data from an external source that is linked to participants via their home postal code.

- **What do the numbers in the variable-assessment matrix mean?**
  The participant numbers in the variable-assessment matrix reflect the number of participants who participated in the measurement or questionnaire in which the given variable was collected. Note that this includes "missings" (particularly relevant for questionnaire data). In other words, the number tells you how many participants were given the question, but not how many participants actually answered.
  Is it crucial for your research to know in advance how many missings you can expect for a certain variable? Please ask our data managers!

- **Why do variables have these complex codes instead of the old, short, variable names?**
  The "historical" variable names were certainly shorter and simpler, but often not very informative. In addition, the sharp increase in new variables makes it more and more difficult to find a short yet unique and informative name. We therefore developed a new code system that gives you more immediate info on the variable (without having to look it up in a code book) and gives us more room to design new names. Please find an explanation of the code components at the end of this document.
  Note that the historical variable names will still be provided in your dataset as part of the variable metadata, to help you with the transition (for example reusing existing syntaxes, or using the annotated questionnaires which still have the historical names for the time being).

- **What do the little + signs mean?**
  Many of the variables come in groups that we expect will always be ordered together (so-called "subvariables"). Hiding them under the + signs makes the matrix much more manageable. You can reveal these subvariables by clicking the +.
  Notable examples of subvariables are:
    - Questions that were repeated at several time points in the serial COVID-questionnaires
    - Equivalent measurements from the different leads in an ECG
    - Repeated questions within a questionnaire for a series of persons or types, such as child nr. 1-10, or medication type 1-10, etc.

  When you select all the variables in a subsection by clicking the vertical arrow at the top, you will automatically also select all the subvariables hidden under the + signs.

However, when you select a single variable with a + sign, please unfold the subvariables and make sure you have selected them all, including the main variable.

- **Where can I find variables on age, gender, and zip-code?**
  Some Lifelines variables are provided by default in every issued dataset. These variables are not visible in the catalogue. Variables that are provided by default are: a project-specific pseudonym, date of birth (month/year), date of death (month/year), date of inclusion (month/year),way of inclusion, date of an element (month/year), invitation for an element (yes/no), time interval between an element and date of inclusion (months), age on the date of an element (months), gender on the date of an element, and home zip code (PC4) on the date of an element.
  More information on this can be found on the wiki.

- **Where can I find variables on omics data in the catalogue?**
  The online catalogue can be used to select phenotypic data, e.g. questionnaire data, measurements, lab data. Genetic data, like the data from the SNP arrays, but also methylation data, RNAseq and micriobiome data, cannot be selected using the catalogue. You can request these data sets by adding this to your application form (section D. Data Selection, question 23). The genetic data will be made available to you via the UMCG high performance cluster (a Linux environment). The phenotype data selection you submit through the online catalogue, will be made available alongside the genetic data.

  It is possible to select a subcohort of participants for which SNP data is available:
  - **Genotype data:** You can select UGLI if you only want to select participants for which SNPs are available which were assessed using the Infinium Global Screening Array® (GSA) MultiEthnic Disease Version 1.0. If you want participants for which SNPs were assessed using the HumanCytoSNP-12v2 you can select subcohort GWAS. If you want both, select both UGLI and GWAS.
  - **DEEP data:** microbiome, proteomics, RNAseq, methylation, metabolites, telomere length and cytokine measurements were assessed within the add-on study DEEP. If you select the subcohort DEEP, you select the participants for which the above were assessed.
    Please note that genotype data for DEEP participants was assessed by using the HumanCytoSNP-12v2 array. If you want to request SNP data for DEEP participants, please also select GWAS as a subcohort.

  Another possibility is to not select a specific subsohort (UGLI, GWAS or DEEP). In this case you will receive phenotype data from participants for which you have selected certain variables. When you gain access to the data, which will be through the UMCG HPC, you can use the provided linkage files to select participants for which your requested omics data is available. Participants for which this data is not available, can potentially be used as controls.

- **I am interested in a specific gene, how can I find out if SNPs located within this gene are available within the genetic datasets?**
  You can send data management an email (data@lifelines.nl) with information on the specific gene you are interested in. Please include the specific rs numbers in your email. Our contacts at the genetics department of the UMCG can find out for you if the rs numbers are located on the

array, or whether they were assessed within the imputed dataset. Please note that costs may be involved for such requests.

- **I have noticed that a variable appears in more than one subsection. Why is that?**
  This is to accommodate different search strategies. It is, however, the same variable, and when you select it in one subsection it will appear as "selected" in the alternative subsection as well. This is also true for the subsections that appear under two different sections: especially in the case of secondary and linked variables (the ones with a [2]). Again, if you select variables in one of these subsections they will be automatically selected in the alternative subsection as well. You will see these variables under both (sub-)sections in your cart, but you will see these variables only once (under their "primary" (sub-)section) in your dataset.

- **Why does it say: *warning, you may have empty variables in your selection*?**
  This happens when your participant or assessment filter reduces a previously selected variable to 0 participants. You will not receive that variable in your dataset (as there are no data to release).

- **How does the search function work?**
  The search function searches in the entire variable metadata (which includes the Dutch and English descriptions/questions, and answer options if applicable). If you have a certain variable code in mind, you can narrow your search to the codes ("Search in name fields only").
  If you already had a section/subsection selected when you start your search, it will limit the search to that section/subsection. Broaden your search by clicking the link "Search all sections".
  Note that the search function still has some issues that we are trying to resolve. Please do not fully rely on the search function at this moment, but also look for relevant variables in the logical subsections (-> use the wiki to get a better sense of the subsections).

- **How can I find all the COVID variables?**
  Covid-variables are distributed over 16 subsections. Most of them are Covid-specific (i.e. Lifestyle (Covid-19)), but some are not. All subsections are described in our wiki.
  You could go to these subsections one by one and find & select the associated covid-variables (make sure the COVQ-assessment is selected in your assessment filter).
  Alternatively, you could generate the full list of Covid-variables (independent of their subsection) by searching for "covt" in name fields only. This will take a while to process.
  If it matches your proposal, you may select and order all variables at once from this full list. However, most approved proposals only require a subset of (Covid-)variables and you should not select and order variables without a pre-approved purpose.

Ordering a dataset

We recommend that you start the process of building a data selection to order only when:

- you have fully oriented yourself
- you know which participant and assessment filters you need (if any)
- you have a list of all the subsections and variables you need and where to find them

- you have some time to finish the data selection in one session (this is not necessary, you can save your data selection and return to it later).
- you have your OV number (received from the LL Research Office) and filled-in application form ready

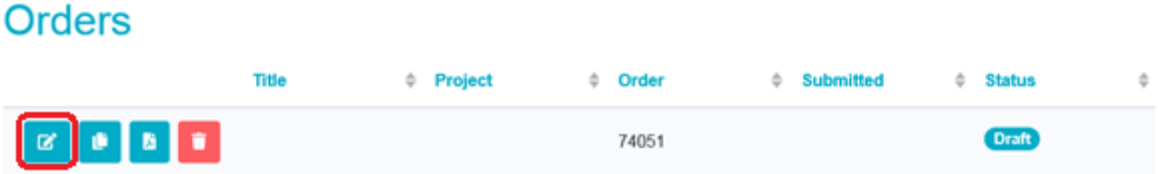
When this is the case, please sign in and start with an empty cart.
Important note: you are automatically signed out when you are working on your data selection for a longer period of time. Therefore, make sure you press save in time (e.g. every hour)!

Step 1: Select participants. If your research requires a certain selection of participants, it is recommended to use this filter at the start of the process. Changing the participant filter after selecting variables may lead to empty variables in your cart.
Note that you will really only receive the selected participants in your dataset!

Step 2. Select assessments. If you need only certain assessments in your research, it is recommended to use this filter now, before you start selecting variables.
Note that you will really only receive the selected assessments in your dataset!

Step 3. Select variables. It is recommended to have a list of required variables and their section/subsection ready once you start your real data selection, to avoid forgetting one.
When you are done, please double check if you properly selected all "subvariables" hidden under the +-signs!!!

Step 4: Save your cart online and as a pdf. We recommend this pdf step to have a "hardcopy" of your cart's contents. Should for some reason your online data selection get lost, this will make it easier to recreate it.
You may now immediately order your cart (find it in the "order" tab), or save it online, take a break and continue it on a different moment. For this you need to locate your saved cart under the "order" tab and edit it.
Please note: when you save your order and continue another day, your cart will seem empty. Your data selection can be found as draft in the tab called 'Order', as shown in the example below. Select 'edit order' to continue with your selection.



Step 5. Order your cart. Click on the "order" button, this submits the cart to Lifelines and a data manager will check its contents against your proposal. You may have already submitted a proposal to the Research Office before, or you may submit it as an attachment with this order.
Note that both components (data selection and application form) must be approved by Lifelines before the data can be released. We will notify you about this process after your data request.

Variable codes

The new variable code contains a logical set of information, as follows:
keyword1_keyword2_subject_type_identifier1_identifier2
For example:
diabetes_medication_ch3_q_1_a

- **Keyword1** describes the main topic of the question. In case of validated questionnaire instruments, keyword1 is the name of the instrument.
  For COVID-variables, keyword1 shows the time point of the variable (i.e. questionnaire 01-12)
- **Keyword2** describes the subtopic of the variable.
- **Subject** states who the variable is about, i.e. the participant itself as an adult or a child, or a family member of the participant.

Subject code	Subject	Age	Reporter
adu	Participant	> 18y	Participant self
ach	Participant	13-17y	Participant self
chi	Participant	0-17y	Parent of participant
ch0	Participant	0-6m	Parent of participant
ch1	Participant	6m-“now”	Parent of participant
ch1a	Participant	6m-3y	Parent of participant
ch2	Participant	4y-“now”	Parent of participant
ch2a	Participant	4-7y	Parent of participant
ch3	Participant	8y-“now”	Parent of participant
ch3a	Participant	8-12y	Parent of participant
ch4	Participant	13y-“now”	Parent of participant
fam	Family member	any	Participant

- **Type** describes the nature of the variable:

q	question
m	measurement
c	code (i.e. coded text fields) or calculation (i.e. sum scores)
l	linked data from an external source
e	evaluation (i.e. an advise or conclusion from a medical expert)
qc	quality controlled copy of an existing variable, protocol for the qc-steps available

- **Identifiers 1** and **2** make sure that variables with shared keywords and subjects/types are still unique. In case of validated or structured instruments, the identifier follows the structure from the original instrument.