Content

1. Introduction

The Lifelines data capture process is an ongoing process; as a result the Lifelines dataset increases each day. The current catalogue is continuously updated to include the most up-to-date data. This document describes the Lifelines data and subjects included in the catalogue and aims to help the researcher to correctly interpret and use the data.

2. Selection of subjects

2.1 Inclusion criteria

In general Lifelines participants have been included in the catalogue and data release if they meet the following criteria:

1. The participant has given an informed consent.
2. The participant has completed at least one visit or questionnaire from baseline assessment.
3. In case of a paper questionnaire, the questionnaire data has been manually verified by the use of verification software.
4. In case of a digital questionnaire, the questionnaire data has been verified by the use of built-in checks during data entry by the participant.

2.2 Exclusion criteria

In general Lifelines participants have been excluded from the catalogue and data release if they indicate to stop participation of the Lifelines project and request Lifelines to remove all previously obtained data.

Please note: If you are an experienced Lifelines user, you may notice that the total number of participants for whom certain data is available sometimes deviates from the previous data release. In most cases the number of participants is increased, mainly because we have added the latest data not yet in the 2017 release. Occasionally, the number of participants for which a given element or variable is available is decreased. This may occur if they do not meet our stricter quality control, for example if we are not able to verify a participant's identity based on the date of birth and gender provided in an assessment. Should this be the case in your data set and you have questions about this, please do not hesitate to ask them. In the future we might add or delete participants based on new developments. As a result, the amount of cases per release will be slightly different.

3. Basic structure of data files

3.1 Data files

In previous releases, the context in which data was collected was named "VMID", "ENCOUNTERCODE", and "FORM_ID". Although a similar structure remains, we have made some improvements and use new names, as follows:

- An **assessment** is the complete collection of data for a given project. General assessments are the "waves" of Lifelines, i.e. 1A, 2B etc. Additional assessments are research projects such as DEEP and Imalife. You can find a full list of general and additional assessments (and their short code name) in the Lifelines WIKI.
- Each assessment has at least one **element** (and often more). An element is a part of an assessment for which participants are separately invited and that can be pinpointed to one specific date. For example, general assessment 1A (baseline) for adults has 4 elements: visit 1, visit 2, questionnaire 1 and questionnaire 2.
- Covid (COVQ) has (for now) 12 elements: one for each consecutive sub-questionnaire.
- Each element has at least one **variant** (and often more). A variant contains a set of variables that is collected under a unique, coherent protocol. Significant changes to that protocol (i.e. the addition/removal/modification of variables, the machine used for the measurement, collection via paper or digital questionnaires, new selection criteria) results in a new variant. For questionnaires, a variant is equivalent to the old "FORM_IDs". For physical measurements, the variants are a new concept.
- Variant names are informative codes that will tell you about the context and uniqueness of the variant, as follows: *assessment_element_number_description_agegroup_version* For example: 1a_q_1_paper_18-65_index
- For each participant we provide the variant under which a given variable within an assessment was collected (note that one participant can provide a specific variable only once per assessment). Please contact us if you would like to see specific questionnaire variants.

3.2 Secondary data

Some of our data does not fit in the assessment-element-variant model. This is the case when a variable is not collected within a specific element, but is either a sum score of variables that were collected in different elements, or it is not associated with any element at all (for example environmental data dependent on participants' postal codes). On Gearshift in the umcg-lifelines group you will find information on secondary data at the following location:

`/groups/umcg-lifelines/prm03/documentation/`

Additionally, on our WIKI page there is an page which lists all secondary and linked data. On both locations we provide you with information on the development of the secondary or linked data and required references when using this data. In case any information is missing please contact data@lifelines.nl and we will add the information as soon as possible.

3.3 Released files

The Lifelines phenotype dataset consists of multiple data files, one for each element. The data will be released as comma-separated (.csv) format.
The provided comma-separated files contain different information:

1. **Results:** The actual results, including age/sex/postal code of the participant at the time of a specific element.
2. **Enumeration meta data:** Coded answers (only applicable for part of the variables).
3. **Variable meta data:** Information on the variables, including Dutch and English labels, datatype, and historical name if applicable.

Please note: the variables in the data files are alphabetically ordered and not, for example, in the order of the questionnaire.

Besides the phenotype data, two additional files are released:

- **Global summary:** This file contains more information for each participant (replacement 'Participant' file). It includes data of birth, date of inclusion, their way of inclusion, (if applicable) date of death (all in month/year), age per element, and the time intervals between an element and the inclusion date (in months). For some elements, the list of invites is provided (i.e. for each participant in the GS it is stated whether they were invited or not, see 3.4).
- **Quality_issues:** This file contains information on found quality issues in the data. A quality issue can be added on all different levels: e.g. participant, variable, or variant. As soon as an quality issue is found, the issue will be added to this file.
  Please note: these additions will be included in new orders only.

3.4 Invites

In the Global Summary, you will find columns containing information about which participants in your selection were invited for certain elements in your selection. This information is intended to be used for bias analysis (responders vs. non-responders).

- For elements under general assessments 1A, 1B, 1C, and 2A, the entire cohort can be considered as invited (within the given selection criteria such as age).
- For all elements from all other general/additional assessments, the list of invites will be added to your global summary if you ordered a variable from that element.
- The list of actual responders to a given element is equal to the list of all participants in the table containing the results for that element. Note that providers of missing data (missing tokens, see section 3.5) are still responders!
- Please don't forget to order the required variables from the assessments that you want to use to compare responders vs. non-responders in your bias analysis.

- You may see some unexpected or unexplained changes in the list of invites from one element to the next. Most of these discrepancies can be explained by the following two occurrences:
    1. The list of participants that comply with the selection criteria has changed between two elements, for example because:
        - Participants have changed their status (i.e. unavailable, pregnant, etc.)
        - Participants have died
        - Participants have reached the minimum or maximum acceptable age
        - The selection criteria have changed (e.g. became stricter or less strict)
    2. A problem with an email address has occurred or has been solved, i.e. certain participants can from now on or no longer be reached.

If you have reason to believe that a discrepancy cannot be explained by these occurrences, please contact the Lifelines Data Managers (data@lifelines.nl) to investigate this in more detail.

3.5 Missings

In previous releases, the fixed tables made it difficult to indicate the precise meaning of empty cells. In the new data structure, we are able to provide some background information, as follows:

- Each variant consists of a known list of variables. If a participant participated in a given variant, but failed to provide data for a certain variable in that variant, the resulting empty cell is automatically filled with a “missing token” (see below). In other words: we did ask the question, but the participant did not give an answer.
- It is also possible that a participant did not provide data for a variable, because he/she never participated in any variant (within a given element/assessment) in which that variable was assessed. In that case, the empty cell remains empty. In other words, that participant-variant-variable combination simply does not exist.
- Depending on your individual matrix of participants and variables, you may see a combination of missing tokens and empty cells in a variable column.
- Since the meaning and consequences of missing data depends on the protocol used, we have prepared different missing tokens (with more to come in the future):
    - $4 = missing data point in an interview
    - $5 = missing data point in a physical measurement
    - $6 = missing data point in a paper questionnaire
    - $7 = missing data point in a digital questionnaire

Please note that the values are encapsulated by quotes, which might affect the data processing in for example R and Python. For your convenience, a R script has been prepared which can be used to replace the default missing values (such as $6) into any other values you would like to have and that can easily be read by R for multiple files. This script is located on the following location on Gearshift:

```
/groups/umcg-
lifelines/prm03/releases/pheno_lifelines/v2/scripts/replace_missings/
```

4. Structure variable names

All variables have been transformed into a more informative code. The “historical” variable names will remain part of the variable metadata and are provided in your dataset, to help with the rebuilding of old syntaxes etc. . You can find these “historical” variable names in the metadata in the folder variables).

4.1 Variable codes

The new variable code is 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.
- **Keyword2** describes the subtopic of the variable.
- **Subject** states who the variable is about, i.e. the participant itself, 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. open text to limited options) or calculation (i.e. variable 1 x variable 2)
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

- **Identifiers 1** and **2** make sure that each variable is unique. In case of validated or structured instruments, the identifier follows the original structure.

4.2 Standard variables

Each data file contains the following variables:

Variable name	Label	Intention
PROJECT_PSEUDO_ID	Project-specific pseudonym	This is the unique study subject identifier. Note: Please use this column to merge individual data files within your dataset and to communicate questions or request regarding individual subjects with Lifelines, for example for a sample selection request. Important notes: When you have access to multiple Lifelines datasets the PROJECT_PSEUDO_ID values differ between datasets. Please consult the README in the release folder for information on merging your phenotype datasets using PROJECT_PSEUDO_ID to genetic datasets (e.g. gsa_genotypes)
VARIANT_ID	Variant of the element	Indication of a specific questionnaire or measurement version, as described in section 3.1 Data files.
DATE	Date of the element (month/year)	
AGE	Age of participant at the time of element (in years)	
GENDER	Gender of participant at the time of element (male/female)	
ZIP_CODE	Home address zip code of participant at the time of element (level-4)	

Please note: secondary and linked data do not include all the variables listed above.

5. Questionnaire information

An overview of all variants for specific questionnaires can be requested by emailing data@lifelines.nl.

5.1 Revisions of questionnaires

From the start of the data collection in 2006, the questionnaires have been revised at some points to improve the collected data. Also, with the introduction of the digital questionnaires a different set-up of questions was introduced. For example, a tick box is used for a lot of questions on health data items in the hardcopy version of the questionnaire. When a person did not tick the answer box, the value in the dataset for that data item is missing. In the digital questionnaire this is changed into yes/no questions to avoid missing values. For these yes/no questions 1=yes and 2=no.
In addition, for the baseline assessment, the older participants (65 years and older) received a slightly adjusted questionnaire compared to the participants aged between 18-65 years. The developed protocols give insight into the revisions/changes. Contact us if you would like to see these protocols.

Please be aware that for above stated reasons, the number of responses for a specific data item can be quite low, for instance when a question has been included in the latest questionnaire version only.

5.2 Different versions of questionnaires due to the food-frequency questionnaire (FFQ)

For three questionnaires (1B, 1C, and 2A questionnaire 1) there are three different versions of the same questionnaire:

- Version A
- Version B
- Version C

The only difference in the three versions (A, B, C) of these questionnaires is the content of the Food-Frequency Questionnaire (FFQ). All participants complete one of the three parts (A, B, or C) as part of the 1B, 1C, and 2A1 questionnaire.

Individual participants will fill out the three parts of the FFQ in different order (ABC, BAC, CAB, etcetera) and at different moments in time, but all participants will eventually receive all three parts of the FFQ. The actual part of the FFQ which has been completed by the participants for a specific questionnaire is indicated by the “VARIANT_ID”. For more information on this topic, please have a look at the FFQ WIKI page.

5.3 Proxy interview during baseline assessment

For persons who performed the measurements conducted during the baseline visit, but who were not able to complete the baseline questionnaire on paper, because their Dutch was insufficient or they had a low score on the cognition test (the MMSE test), a family member was interviewed (also known as the proxy interview). The data from the proxy interviews can be found looking at the VARIANT_ID. Most questions from the proxy interview were similar to the questions in the regular questionnaire.
No proxy interviews were performed during follow-up questionnaires and the second assessment. The persons were asked to complete the regular questionnaires, with help of someone else if necessary.

6. Privacy sensitive information

6.1 Personal data

Lifelines receives the date of birth, date of death, gender and postal code directly from a linkage with the municipal administration (in Dutch: “BasisRegistratie Personen”; BRP). However, due to privacy reasons, the dataset does not contain the date of birth, date of death, and the entire postal code.

A number of derived data items, like month/year of birth and death as well as the four digits of participant’s postal code (zip code) can be found in your data files.

6.2 Open-ended questions

Several questionnaires contain open-ended questions. These questions allow participants to give a free-form answer. When answering such a question a participant can (unconsciously) fill out identifiable information. As parents completed the questionnaires of their children, there is, for example, a significant chance that they filled out their child’s name in their answer. To prevent spreading of such sensitive information, part of the open-ended questions are not released by default.

7. Data quality considerations

7.1 Questionnaire data

Data from the hardcopy Lifelines questionnaires have undergone a manual verification process, in which verification software was used to verify the data. Purpose of this process was to correct potential scanning errors in the data. Data from the electronic Lifelines questionnaires have undergone an online validation process, by the use of built-in verification checks during data entry by the participants.

7.2 Measurement data

Part of the measurement data have been evaluated by field experts, for example the evaluation of the lung function tests by trained pulmonologists and the ECG by trained cardiologists.
More information on the measurements and the evaluation of the measurement data can be requested at data@lifelines.nl.

7.3 Meta data

The English variable labels and value labels that are used in the data files have been translated by a translation agency based on the original Dutch labels as used in the questionnaires. This is the case for almost all variables. For variables from verified instruments the original English label has been used.

7.4 Population characteristics

The data item "SOURCE" in the dataset shows whether the subject was recruited via the general practitioner, via a family member of via self-registration. The way the subject was included in the study may be associated with other subject characteristics like age.

Details on the recruitment strategy, visits and measurements can be found in the Lifelines protocols and the Lifelines publications . please contact us at data@lifelines.nl if you would like to see these documents.

8. Additional questionnaires

8.1 COVID-19 questionnaires

The COVID-19 data have some distinguishing aspects compared to our regular assessments. Below there are some pointers for using the data:

- Each element (i.e. part of the assessment with 1 invitation and 1 response date) has its own separate table, and therefore each COVID-19 questionnaire (1 to 12) has its own separate data table. When combining different tables, you can easily recognize the questionnaire number in the variable name (e.g. *covt01…*, *covt02…*)
- Participants are included if they completed at least one COVID-19 questionnaire. As a result, participants…
    - …may have participated in one COVID-19 questionnaire only.
    - …are not necessarily present in the first COVID-19 questionnaire, but could have started with the second or a later questionnaire.
- There are differences between the invites for the first COVID-19 questionnaires and later COVID-19 questionnaires, as all participants who didn't complete any of the first seven questionnaires didn't receive any invites for COVID-19 questionnaires onwards.
- Participants received invitations for the first twelve COVID-19 questionnaires in a short-time frame. One consequence of this is that one questionnaire could still be completed by participants while the invitation for the next one had been emailed already; possible overlap between questionnaires concerning the date of completion. If timing is important for your research, please use the *'…responsedate…'* variable for your analyses.

9. More information

After reading the available documentation, you might still have questions concerning your dataset. If this is the case, please do not hesitate to contact the Lifelines Data Managers (mailto:data@lifelines.nl).