Lung Cancer Genetics Study (LCGS) Data Documentation v1
October 2025

This documentation is provided for the 2025-07-18 ILD data deliverable

Sample
- 18 years old or older
- Country is US
- Accepted the 23andMe research consent
- Accepted the LCGS study consent
- Accepted the LCGS study terms of service acknowledgement document

Lung Cancer Genetics Study Baseline Survey
Lung Cancer Genetics Study Baseline Survey [lung_cancer_baseline].docx

Variables
All variables from the LCGS Baseline Survey are included.

Only a subset of variables from the Health Survey, Tobacco (Smoking) Survey, and Environment and Lifestyle History Survey from the 23andMe database are provided.

For example, the variable iqb.hp_list_24_v3 ("Have you ever been diagnosed with, or treated for, any of the following conditions? — Asthma, COPD (chronic obstructive pulmonary disease, including emphysema), Bronchitis, Cystic fibrosis, or any other lung or respiratory condition") is included because it serves as a dependency for the questions iqb.lc_baseline_chronic_bronchitis_dx and iqb.lc_baseline_tuberculosis_dx, which are part of this dataset. However, data for asthma and cystic fibrosis are not included. Variable inclusion decisions were made by study personnel.

Free text is not included in this deliverable.

Codebook
LCGS_Baseline_Survey_Codebook_v1

Codebook Column Descriptions

Column name	Description
Phenotype	Variable name*. Related to observation_name in Lifebit Source Table.
question_dependencies	Dependencies at the question level
group_question_text	Question text for grouped questions**

question_text	Main question text
tooltip_text	Tooltip text. Appears as a "?" button for additional information
continue_cards_text	Text displayed on continuation cards between questions
database_description	Description of the phenotype from the 23andMe database
choices	Answer choice IDs and labels, showing how options appear to respondents. Related to observation_value and observation_value_raw in Lifebit Source Table.
value_encodings	Mapping of answer values to choice IDs. Related to observation_value and observation_value_raw in Lifebit Source Table.
column_type	SQL datatype [BIGINT, BOOLEAN, FLOAT, BIGINT[], TIMESTAMP_NS, VARCHAR, DOUBLE]
question_type	Type of question [radio, integer, checkbox, content_paragraphs, date_year, postal_code autocomplete_multi]***
allowable_min	Minimum allowable answer value
allowable_max	Maximum allowable answer value
source	Name of 23andMe survey or from the 23andMe database****

Lifebit Source Table Column Descriptions

Column name	**Description**
observation_date	Date question answered
observation_name	Relates to Phenotype in “Codebook Column Descriptions” table
observation_value	observation_value_raw without special characters
observation_value_raw	Relates to value_encodings and column_type in “Codebook Column Descriptions” table
participants_id	Unique participant identifier

*** phenotype naming conventions**

- iqb.*
  - phenotype representing a single question with no further processing
- lc_*
  - phenotype derived from questions in the Lung Cancer Genetics Study Survey
- hp_*
  - phenotype derived from questions in the primary Health Survey
- hp_list_

        - Phenotype representing the “parent” question asked before a “child” diagnosis question in the primary Health Survey.
            - For example, *iqb.hp_list_32* asks: “Have you ever been diagnosed with, or treated for, any of the following conditions?

 • Breast cancer

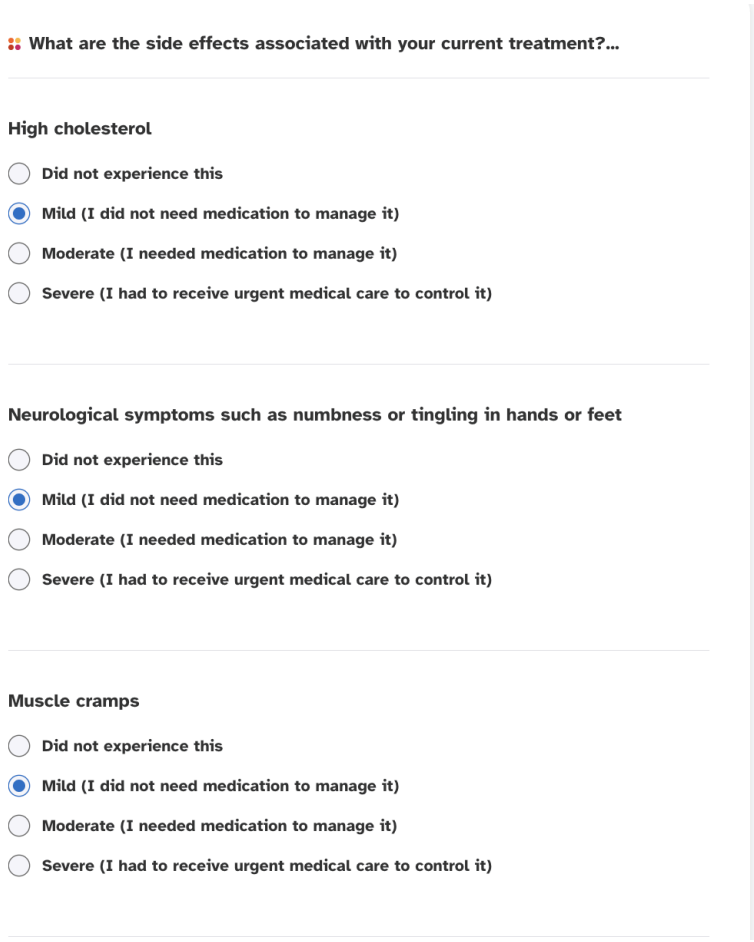
 • Colorectal cancer (colon cancer or rectal cancer)

 • Lung cancer

 • Prostate cancer” [Yes/No/I’m not sure]
            - If *iqb.hp_list_32* = yes OR not_sure, then iqb.hp_breast_cancer_v2, iqb.hp_colorectal_cancer_v2, iqb.hp_lung_cancer_v2 and iqb.hp_prostate_cancer_v2 are asked.
            - To get complete control (“no”) counts where a hp_list_ is listed in question_dependencies, combine parent and child responses.
    - env_*
        - phenotype derived from questions in the Environment and Lifestyle History Survey

**** grouped questions**

- Grouped questions are questions that appear in sequence and share the same question stem. For example, iqb.lc_baseline_muscle_cramps, iqb.lc_baseline_neurological_symptoms and iqb.lc_baseline_high_cholesterol are questions that share the group_question_text “What are the side effects associated with your current treatment?...”. In the 23andMe survey interface, this text remains static as a header, while the question_text “Muscle cramps” and “Neurological symptoms such as numbness or tingling in hands or feet” and “High cholesterol” dynamically appears along with the answer choices underneath this header.



***** question_type = content_paragraphs**

- Phenotypes with a question_type being “content_paragraph” do not have data. These are used to display text in between questions.

****** Source = 23andMe database**

- Phenotypes with a source being “23andMe database” are derived from the greater 23andMe database. These will have a *database_description*. *Question_text* is provided where applicable, but *question_type* or *choices* are not provided.
- lung_cancer_baseline = Lung Cancer Genetics Study Baseline Survey
- health_profile = 23andMe Health History Survey
- tobacco = 23andMe Smoking Survey
- environment_lifestyle_history = 23andMe Environment and Lifestyle History Survey

Other

- **hp_breast_cancer_v2, hp_colorectal_cancer_v2, hp_lung_cancer_v2** and **hp_prostate_cancer_v2.** Respondents may be asked these questions twice within the

Health Survey. hp_breast_cancer_v2, hp_colorectal_cancer_v2, hp_lung_cancer_v2 and hp_prostate_cancer_v2 are asked at the beginning of 23andMe's main Health Survey. We ask about these cancers again later on in the survey, when we ask the parent question hp_any_other_cancer : "Have you ever been diagnosed with, or treated for, any type of cancer not already mentioned?". If answered "yes" then the question, "Have you been diagnosed with any of the following types of cancer?" with "Breast cancer" and "Colorectal cancer" and "Lung cancer" and "Prostate cancer" being answer options (hp_breast_cancer, hp_colorectal_cancer, hp_lung_cancer and hp_prostate_cancer, respectively).

- HTML tags are used in the question_text field. This shows exactly how the question was coded and appears to respondents. For example, A word flanked by "" is bolded in the survey text.

Genotyping array platforms

90% of study participants are on the 23andMe v5 platform. 7.5% are on v4, and the remainder are on v3 or v2.

The v2 platforms were based on the Illumina HumanHap550 BeadChip and contain a total of about 560,000 variants, including about 25,000 custom variants selected by 23andMe. The v3 platform was based on the Illumina OmniExpress BeadChip and contains a total of about 950,000 variants and custom content to improve the overlap with 23andMe's v2 array. The v4 platform was a fully custom array of about 570,000 variants and included a lower redundancy subset of v2 and v3 variants with additional coverage of lower-frequency coding variation. The v5 platform was based on the Illumina Global Screening Array, consisting of approximately 654,000 preselected variants and approximately 50,000 custom content variants. Samples that failed to reach 98.5% genotyping call rate were reanalyzed.