

Instructions for using Mica Search Tool of euCanSHare Data Catalogue

www.eucanshare.eu



4th November 2022, Version 1

Tarja Palosaari and Jaakko Reinikainen, Finnish Institute for Health and Welfare (THL)

© 2022 by [the euCanSHare Project](#)



Contents

- [euCanSHare Data Catalogue](#)
- Functionalities of the Search Tool
 - [Basic functionalities](#)
 - [Adding search criteria and presenting results](#)
 - Filtering by study- and/or variable-properties
 - Modifying search queries
 - Presenting search results in various ways
 - [Saving the search query for later use](#)
- [Finding harmonization potential across the studies](#)
- [Example search with solution](#)

euCanSHare Data Catalogue

- Data Catalogue is a discovery tool for cardiovascular research data.
- It provides detailed information on
 - characteristics of the studies, such as study design and data access policies;
 - data collected in the studies, such as variable descriptions;
 - omics and bioimaging data available for the studies.
- Data Catalogue is built using OBiBa's Mica and Opal software applications, including:
 - Cohort Browser that provides a structured description of the participating studies, and
 - Search Tool that allows users to browse information within the Data Catalogue using a powerful search engine.
 - Instructions for using the Search Tool are provided in this tutorial.

More information on catalogue tools

- OBiBa tools are open source software for data management, harmonization, co-analysis and dissemination of the epidemiological research.
- OBiBa tools are described in detail in the websites of the [Maelstrom Research](#) and [OBiBa](#)
- Watch also the video tutorial on using the search of the Maelstrom Catalogue that is built on same tools as the euCanSHare Data Catalogue
 - Video: <https://www.maelstrom-research.org/page/tutorials?topic=search>

euCanSHare Data Catalogue frontpage: <https://mica.eucanshare.bsc.es/>
Mica Search Tool can be accessed by clicking **GO** in "Search for studies and variables"

euCanSHare Catalogue Home Repository

euCanSHare Data Catalogue

We facilitate data discoverability in cardiovascular research with a data catalogue that includes data from multiple European and Canadian cohorts.

Description of Studies **GO** Search for studies and variables **GO**

Number of studies related with different data types, sources and samples

Data Type	Number of Studies
Sociodemographics	28
Lifestyle	30
Biomarkers	31
Omics	13
ECG	13
Imaging	9
Diseases and death	31
Environmental	7
Physical measures	31
Biological samples	31

Cohort Browser

Search Tool

Functionalities of the Search Tool (1/22)

Basic functionalities

Search

The Search tool aims to facilitate the identification of epidemiological studies or harmonization projects collecting or having generated variables of interest to answer specific research questions. The Search tool allows users to browse variables, epidemiological studies (including specific data collection events), and harmonization projects, and to explore the harmonization potential across studies or data collection events.

If you need help to get started, watch the [Search tutorial](#)

Query

Download the search results presented as a CSV file.



No search criteria

Choose the layout for presenting the search results: list or comparison table. (Comparison table works only if some variable-level criteria are selected.)

Lists view

Comparison table

Variables 101 923

Datasets 178

Studies 34

Choose the layout for presenting the search results: list of variables, datasets or studies that met the search criteria used.

All Individual Harmonization

Show 100 entries

Choose the type of the studies included – **Individual** studies are usually of interest.

Data types, sources, and samples available

Individual

A **harmonization study** is a research project harmonizing data across individual studies to answer specific research questions.

Functionalities of the Search Tool (2/22)

Basic functionalities - Results as a list of studies

The screenshot shows the euCanSHare Catalogue search results page. The interface includes a top navigation bar with 'euCanSHare Catalogue', 'Home', and 'Repository' menus. A search bar contains the text 'Query'. Below the search bar, there are buttons for 'Individual' (circled in orange) and 'Harmonization'. A callout box points to the 'Individual' button with the text 'Modifiable search query'. Below the search bar, there are buttons for 'Lists view' and 'Comparison table'. A callout box points to the 'Lists view' button with the text 'List view of studies selected'. Below the search bar, there are buttons for 'Variables 101 351', 'Datasets 175', and 'Studies 31' (circled in orange). A callout box points to the 'Studies 31' button with the text 'See the variable descriptions of the study'. Below the search bar, there are buttons for 'All', 'Individual', and 'Harmonization'. A callout box points to the 'Individual' button with the text 'First of all, click the "Individual" to exclude harmonization studies'. Below the search bar, there is a 'Show 100 entries' dropdown menu. Below the search bar, there is a table with columns for 'Acronym', 'Name', 'Type', 'design', 'Data types, sources, and samples available', 'Participants', 'Datasets', and 'Variables'. The table lists three studies: ATBC, AtheroGene Study, and BHFC. The 'Variables' column for the ATBC study is circled in orange. A callout box points to the 'BHFC' study with the text 'See the study description.'. The footer of the page contains the text '© 2022 by the euCanSHare Project'.

euCanSHare Catalogue Home Repository

Query

Individual x

Modifiable search query

Lists view Comparison table

List view of studies selected

Variables 101 351 Datasets 175 Studies 31

All Individual Harmonization

First of all, click the "Individual" to exclude harmonization studies

Show 100 entries

See the variable descriptions of the study

Individual

Previous 1 Next

Acronym	Name	Type	design	Data types, sources, and samples available	Participants	Datasets	Variables
ATBC	ATBC Study	Individual	Cohort	✓ ✓ ✓ ✓ - - ✓ - ✓ ✓	29000	4	279
AtheroGene	AtheroGene Study	Individual	Cohort	✓ ✓ ✓ ✓ ✓ - ✓ - ✓ ✓	3800	2	77
BHFC	Brno Heart Failure Cohort	Individual	Cohort	✓ - ✓ - ✓ ✓ - ✓ ✓	702	3	89

See the study description.

© 2022 by the euCanSHare Project

Functionalities of the Search Tool (3/22)

Basic functionalities - Icons for data collected in the studies



euCanSHare Catalogue



Home

Repository



Sign in

Clear search

Clear the search query / start a new search.

Data collected, study-level properties

Filter by Studies

- Study properties
- Data topics and sources
- Omics data
- Imaging data
- Biosamples
- Study identification

Filter by Variables

- General classification
- CV rel. diseases variables
- Variable name & label

Individual

Lists view Comparison table

Variables 101 351 Datasets 175 Studies 31

All Individual Harmonization

Show 100 entries

Select the number of rows in a page: 10/20/50/100

Sociodemographics
Lifestyle
Biomarkers
Omics
ECG
Imaging
Diseases and/or death
Environmental
Biosamples
Physical measures

Data types, sources, and samples available

Previous 1 Next

Acronym	Name	Type	Study design	Data types, sources, and samples available										Individual		
				Participants	Datasets	Variables										
ATBC	ATBC Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	29000	4	279
AtheroGene	AtheroGene Study	Individual	Cohort	✓	✓	✓	✓	✓	-	✓	-	✓	✓	3800	2	77
BHFC	Brno Heart Failure Cohort	Individual	Cohort	✓	-	✓	-	✓	✓	✓	-	✓	✓	702	3	89

Functionalities of the Search Tool (4/22)

Basic functionalities - Results as a list of variables

Variables can be selected to be included in the downloadable CSV file.

List view of variables selected

Navigate between the result pages

See the variable description.

Individual

Lists view Comparison table

Variables 101 351 Datasets 175 Studies 31

All Individual Harmonization

Show 100 entries

<input type="checkbox"/>	Name	Label	Value type	Annotations	Type	Study	Population	Data Collection Event	Dataset
<input type="checkbox"/>	SEX	Sex	Integer		Collected	AtheroGene	AtheroGene	Recruitment	AtheroGeneBL
<input type="checkbox"/>	AGE1	Age	Decimal		Collected	AtheroGene	AtheroGene	Recruitment	AtheroGeneBL
<input type="checkbox"/>	HEIGHT	Height	Decimal		Collected	AtheroGene	AtheroGene	Recruitment	AtheroGeneBL
<input type="checkbox"/>	WEIGHT	Weight	Decimal		Collected	AtheroGene	AtheroGene	Recruitment	AtheroGeneBL
<input type="checkbox"/>	BMI	BMI	Decimal		Collected	AtheroGene	AtheroGene	Recruitment	AtheroGeneBL

Previous 1 2 3 4 5 ... 1014 Next

© 2022 by the euCanShare Project

[Link to this search](#)

Functionalities of the Search Tool (5/22)

Basic functionalities - Results as a list of datasets

Clear search

Individual

Filter by Studies

- Study properties
- Data topics and sources
- Omics data
- Imaging data
- Biosamples
- Study identification

Filter by Variables

- General classification
- CV rel. diseases variables
- Variable name & label

Lists view Comparison table

Variables 101351 Datasets 175 Studies 31

All Individual Harmonization

Show 20 entries

Previous 1 2 3 4 5 ... 9 Next

Acronym	Name	Type	Studies	Variables
AtheroGeneBL	Atherogene Study Baseline	mica_dataset.classname.studydataset	1	69
AtheroGeneFU	AtheroGene Study Follow-up	mica_dataset.classname.studydataset	1	8
KORA0101_MORGAM	KORA Cohort 01 Baseline MORGAM data	mica_dataset.classname.studydataset	1	98
KORA01FU_MORGAM	KORA Cohort 01 Follow-up MORGAM data	mica_dataset.classname.studydataset	1	55

See the dataset description and the variables included.

© 2022 by the euCanShare Project

Adding search criteria and presenting results (1/2)

Search criteria can be applied either using properties defined on 1) study-level or 2) variable-level, or both, and search results in all cases can be chosen as list of studies, datasets or variables.

1) To search by using study-level properties

i.e. filtering search results based on data collected in the studies or the features of the participants. “Filter by Studies”

Includes ALL variables in the studies.

Lists view Comparison table

Variables 101 351 Datasets 175

Studies 31

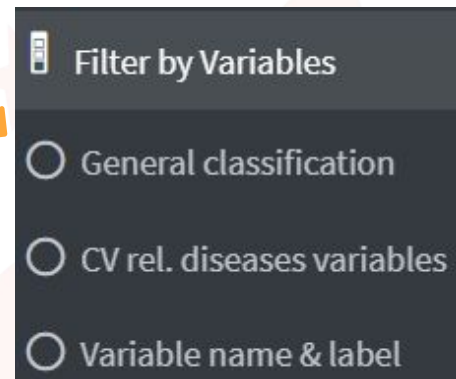
All Individual Harmonization

Note that filtering on study-level affects **only to the studies included** in the results, and NOT the variables by their content.

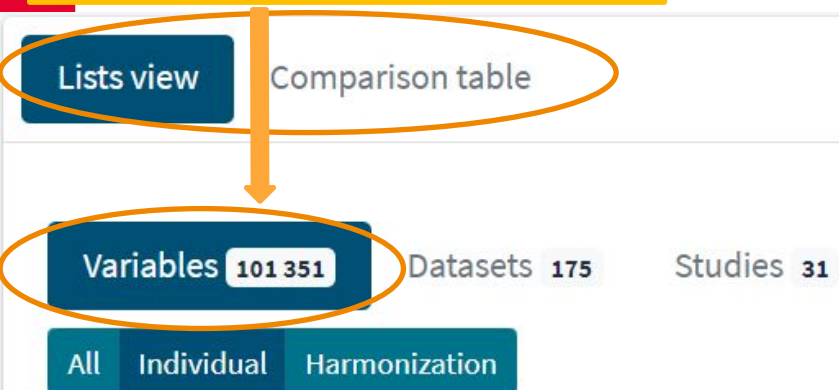
- Filter by Studies
- Study properties
- Data topics and sources
- Omics data
- Imaging data
- Biosamples
- Study identification

Adding search criteria and presenting results (2/2)

2) To search by using variable-level properties i.e. filtering search results by the contents of the variables included in the studies. “Filter by Variables”



Variables will be filtered by their content when variable-level criteria are used.



Filtering on variable-level affects in addition to studies included, **also to the list of the variables by their content**. Comparison table view can also be selected after filtering by variables' content.

Functionalities of the Search Tool (6/22)

1) Adding search criteria in study-level - Filter by studies

E.g. next slide shows contents of “Study properties” and “Data topics and sources”

Use search criteria for study-level properties

The screenshot shows a search tool interface with a sidebar on the left containing various filter categories. The main area displays search results for 'Individual' studies. A callout box highlights the 'Filter by Studies' option in the sidebar. Another callout box points to the 'Individual' filter button in the search results. A third callout box points to the 'Studies 31' button. A fourth callout box points to the 'Study identification' filter option in the sidebar. The table below shows the results of the search, with columns for Acronym, Name, Type, Study design, and various data types.

Acronym	Name	Type	Study design	Participants	Datasets	Variable
ATBC	ATBC Study	Individual	Cohort	29000	4	279
AtheroGene	AtheroGene Study	Individual	Cohort	3800	2	77
BHFC	Brno Heart Failure Cohort	Individual	Cohort	702	3	89

Selected search criteria are shown here and the query can be modified.

Only list of studies is of interest if filtered just by study properties.

Use “Study identification” to select the specific study/studies by the acronym of the study.

Functionalities of the Search Tool (7/22)

1) Adding search criteria in study-level - Study properties

Study properties

Study properties as defined in the catalogue.

Select all subcategories at once

Here selected are “cohort” from “study design” and “general population” from “sources of recruitment”

Study design Select All Clear Selection

The design of an observational or experimental study (e.g. cohort, case control).

Cohort Case-control Case only
 Cross-sectional Clinical trial Other

Observational study that involves the analysis of data collected from a population at one specific point in time. More

Selection criteria - Country of residence Select All

Participant's country of residence.

<input type="checkbox"/> Andorra	<input type="checkbox"/> United Arab Emirates	<input type="checkbox"/> Afghanistan
<input type="checkbox"/> Antigua and Barbuda	<input type="checkbox"/> Anguilla	<input type="checkbox"/> Albania
<input type="checkbox"/> Armenia	<input type="checkbox"/> Angola	<input type="checkbox"/> Antarctica
<input type="checkbox"/> Argentina	<input type="checkbox"/> American Samoa	<input type="checkbox"/> Austria
<input type="checkbox"/> Australia	<input type="checkbox"/> Aruba	<input type="checkbox"/> Åland Islands
<input type="checkbox"/> Azerbaijan	<input type="checkbox"/> Bosnia and Herzegovina	<input type="checkbox"/> Barbados
<input type="checkbox"/> Bangladesh	<input type="checkbox"/> Belgium	<input type="checkbox"/> Burkina Faso
<input type="checkbox"/> Bulgaria	<input type="checkbox"/> Bahrain	<input type="checkbox"/> Burundi
<input type="checkbox"/> Benin		

More: See all categories if the list is long More

Pointing by mouse gives the description of the category.

Then click:

Display results

Selection criteria - Sex Select All

Participant's sex

Men only Women only

More

Sources of recruitment Select All Clear Selection

The population(s) from which individuals are recruited to participate in the study.

General population Specific population Participants from existing studies

Functionalities of the Search Tool (8/22)

1) Adding search criteria in study-level - Data topics and sources

Here selected are “imaging” and “biomarkers” from “data types”

Click:

Display results

Data sources

Select All

Source from which information is generated/extracted

- Questionnaires
- Biosamples
- Cognitive measures
- Administrative databases
- Physical measures
- Others

▼ More

Data types

Select All Clear Selection

Data types

- Imaging
- Lifestyle data
- Diseases and/or death
- Omics
- Sociodemographics
- Environmental
- ECC
- Biomarkers

▼ More

Functionalities of the Search Tool (9/22)

Results by studies when some study-level properties are selected.

Modifying the search query



By clicking small arrows in the query you can modify the criteria. Note that OR “|” is the default operator when selecting many categories within one section; AND-operator can not be applied here.

AND/OR can be selected

Imaging available for 6 and biomarkers for 26 studies (26 in total)

Operation	Description
any	Classified to any non-empty category.
none	Classification must be missing.
in (default)	Classified to at least one of the selected categories.
not in	The negation (opposite) of "in".

Acronym	Name	Type	Study design	Data types, sources, and samples available												
ATBC	ATBC Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	4932	6	751
Brianza	Brianza MONICA Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	2171	-	-
Caerphilly	Caerphilly Prospective Study	Individual	Cohort	✓	✓	✓	✓	✓	-	✓	-	✓	✓	2171	-	-

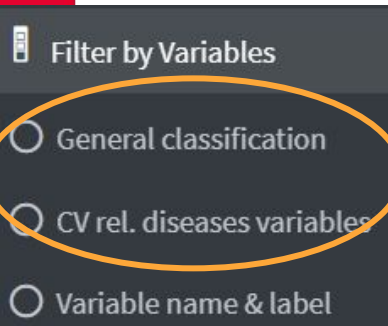
Filter by variables -options (1/2)

- General classification
 - “Areas of Information” classification for the topics of the variables, developed by Maelstrom Research
 - A variable is classified into some subcategory of e.g. “Lifestyle and behaviours”, “Diseases” or “Laboratory measures”
- Cardiovascular (CV) related disease variables
 - This classification includes some specific CV related diseases, that are subcategories of broader definition of the diseases, developed by euCanShare

These classifications make it possible and easy to search for variables in the specified areas of interest across the studies.

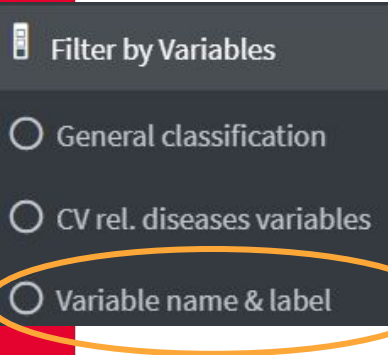
- Studies must have added the variable descriptions and also classified each variable into these categories, to be able to find that study and its variables when using variable-level filtering.

These two variable classifications are the recommended way to perform a search using the variable properties.



Filter by variables -options (2/2)

- Variable name & label
 - Search by matching the text in the variable name or label text
 - Using this search may miss some or all relevant variables of the studies, due to e.g.
 - language or abbreviations used, or
 - insufficient information given in the variable label (the information can be written in the broader description of the variable, and that is not used in the search).
 - This search option can be used in the situations when the variables of the study are already familiar to the user.



Functionalities of the Search Tool (10/22)

2) Adding search criteria in variable-level - Filter by variables

E.g. next slides shows some categories of the “General classification” and “Cardiovascular (CV) related disease variables”

Use search criteria for variable-level properties

Clear search

- Studies
- Properties
- and sources
- Imaging data
- Biosamples
- Study identification
- Filter by Variables**
- General classification
- CV rel. diseases variables
- Variable name & label

Query

Individual AND Cohort AND General population AND Biomarkers | Imaging

Lists view Comparison table

Variables 77 041 Datasets 161 Studies 26

All Individual Harmonization

Show 100 entries

NOTE that studies must have added the variable descriptions to the catalogue and the variables have to be classified into the specific categories (listed in these links), so that it is possible to find that study and its variables by using these variable classes.

Acronym	Name	Type	Study design	Data types, sources, and samples available										Individual		
				Participants	Datasets	Variables	Genetics	Imaging	Microbiology	Proteomics	Metabolomics	Cellular	Behavioral	Environmental	Other	
ATBC	ATBC Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	29000	4	279
Brianza	Brianza MONICA Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	4932	6	751
Caerphilly	Caerphilly Prospective Study	Individual	Cohort	✓	✓	✓	✓	✓	-	✓	-	✓	✓	2171	-	-

Functionalities of the Search Tool (11/22)

2) Adding search criteria in variable-level - General classification (part of it)

Here selected are "tobacco" from "lifestyle and behaviours" and "diseases of the circulatory system (I00-I9)" from "diseases"

Click:

Display results

<p>Diseases Select All Clear Selection</p> <p>Information about past and current disease as categorized in ICD-10.</p> <ul style="list-style-type: none"><input type="checkbox"/> Certain infectious and parasitic diseases (A00-B99)<input type="checkbox"/> Endocrine, nutritional and metabolic diseases (E00-E90)<input type="checkbox"/> Diseases of the eye and adnexa (H00-H59)<input type="checkbox"/> Diseases of the<input type="checkbox"/> Neoplasms (C00-D48)<input type="checkbox"/> Mental and behavioural disorders (F00-F99)<input type="checkbox"/> Diseases of the ear and mastoid process (H60-H95)<input type="checkbox"/> Diseases of the<input type="checkbox"/> Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89)<input type="checkbox"/> Diseases of the nervous system (G00-G99)<input checked="" type="checkbox"/> Diseases of the circulatory system (I00-I99)<input type="checkbox"/> Diseases of the skin	<p>Socio-demographic and economic characteristics Select All</p> <p>Information about socio-demographic and economic characteristics.</p> <ul style="list-style-type: none"><input type="checkbox"/> Age/birthdate<input type="checkbox"/> Marital/partner status<input type="checkbox"/> Residence<input type="checkbox"/> Sex/gender<input type="checkbox"/> Family and household structure<input type="checkbox"/> Birthplace<input type="checkbox"/> Language<input type="checkbox"/> Other socio-demographic and economic characteristics<input type="checkbox"/> Twin<input type="checkbox"/> Education<input type="checkbox"/> Citizenship and immigrant status<input type="checkbox"/> Labour force and retirement <p>More</p>	<p>Lifestyle and behaviours Select All Clear Selection</p> <p>Information about past and current lifestyle, behaviours and activities.</p> <ul style="list-style-type: none"><input checked="" type="checkbox"/> Tobacco<input type="checkbox"/> Nutrition<input type="checkbox"/> Transportation<input type="checkbox"/> Sexual behaviours and orientation<input type="checkbox"/> Misbehaviour and criminality<input type="checkbox"/> Alcohol<input type="checkbox"/> Breastfeeding<input type="checkbox"/> Personal hygiene<input type="checkbox"/> Leisure activities<input type="checkbox"/> Other and unspecified lifestyle information<input type="checkbox"/> Drugs<input type="checkbox"/> Physical activity<input type="checkbox"/> Sleep<input type="checkbox"/> Technological devices <p>More</p>
	<p>Reproductive health history Select All</p> <p>Birth and current or past reproductive health history of</p> <ul style="list-style-type: none"><input type="checkbox"/> Contraception<input type="checkbox"/> Other reproductive health-related information<input type="checkbox"/> Pregnancy, delivery and birth <p>More</p>	<p>Perception of health, quality of life, development and functional limitations Select All</p> <p>Information about perception of general health, quality of life, child development and decline of functional capacities.</p> <ul style="list-style-type: none"><input type="checkbox"/> Perception of health<input type="checkbox"/> Functional limitations<input type="checkbox"/> Quality of life<input type="checkbox"/> Use of assistive devices<input type="checkbox"/> Life course development<input type="checkbox"/> Other perception of health, quality of life and functional limitation-related information

Functionalities of the Search Tool (12/22)

2) Adding search criteria in variable-level - Cardiovascular related disease variables

Here selected is
"heart failure (I50)"
from
"cardiovascular
related diseases"

Click:

Display results

Cardiovascular related diseases

Select All Clear Selection

Diabetes mellitus
(E10-E14)

Disorders of
lipoprotein
metabolism and
other lipidaemias
(E78)

Hypertensive
diseases (I10-I15)

Ischaemic heart
diseases (I20-I25)

Valve disorders (I34-
I37)

Conduction
disorders and
cardiac arrhythmias
(I44-I49)

Heart failure (I50)

Diseases of the
circulatory system
falling into multiple
categories

Cerebrovascular
diseases (I60-I69)

Functionalities of the Search Tool (13/22)

Results by studies when some study- and variable-level properties selected

At first, check the query used!

Here is an unintended error: AND-operator

Criteria used in variable-level:

Criteria used in study-level:

Query

Criteria used in variable-level: Tobacco OR Diseases of the circulatory system (I00-I99) AND Heart failure (I50)

Criteria used in study-level: Individual AND Cohort AND General population OR Biomarkers | Imaging

Lists view Comparison table

Variables 278 Datasets 59 Studies 11

All Individual Harmonization

Show 100 entries

Acronym	Name	Type	Study design	Data types, sources, and samples available										Individual		
				Participants	Datasets	Variables										
Catalonia	Catalonia MONICA Study	Individual	Cohort	✓	✓	✓	-	-	-	✓	-	✓	✓	5505	3	20
DAN-MONICA	DAN-MONICA Study	Individual	Cohort	✓	✓	✓	-	-	-	✓	-	✓	✓	7582	8	43

AND / OR selection:

NOTE that a variable is usually classified into one category only, so AND-operator may not be sensible for filtering by variables.

When adding a criteria from “cardiovascular related disease variables”, AND is used mistakenly as the default operator. **Change it to OR.**

“Heart failure” happen to be a subcategory for the “diseases of the circulatory system”, so the result set here is not totally empty, even if AND-operator is used.

Functionalities of the Search Tool (14/22)

Results by studies when some study- and variable-level properties selected

Query



Tobacco x OR Diseases of the circulatory system (I00-I99) x OR Heart failure (I50) x
Individual x AND Cohort x AND General population x AND Biomarkers | Imaging x

Corrected query:

OR-operator is used for all variables categories

Lists view Comparison table

Note that the variables of some studies were not yet (in October 2022) classified to the cardiovascular related disease classes.

Variables 8 482

Datasets 147

Studies 23

23 studies include 8482 variables on some of the selected variable classes.

All Individual Harmonization

Show 100 entries

Previous 1 Next

Acronym	Name	Type	Study design	Data types, sources, and samples available										Individual		
				👤	📄	🖋️	🧬	🫀	🏠	☠️	🌐	🔧	🔄	Participants	Datasets	Variables
ATBC	ATBC Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	29000	4	148
Brianza	Brianza MONICA Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	4932	6	348
Catalonia	Catalonia MONICA Study	Individual	Cohort	✓	✓	✓	-	-	-	✓	-	✓	✓	5505	4	282

Functionalities of the Search Tool (15/22)

Results by variables when some study- and variable-level properties selected

Tobacco OR Diseases of the circulatory system (I00-I99) OR Heart failure (I50)
Individual AND Cohort AND General population AND Biomarkers | Imaging

Lists view Comparison table

Variables 8482 Datasets 147 Studies 23

List view of variables selected

All Individual Harmonization

Show 20 entries

Variables in the specific classifications

Previous 1 2 3 4 5 ... 425 Next

<input type="checkbox"/>	Name	Label	Value type	Annotations	Type	Study	Population	Data Collection Event	Dataset
<input type="checkbox"/>	CIGS	MORGAM variable: "Do you smoke cigarettes now?"	Integer	Tobacco	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM
<input type="checkbox"/>	NUMCIGS	MORGAM variable: "On average how many cigarettes do you now smoke a day?"	Integer	Tobacco	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM
<input type="checkbox"/>	EVERCIG	MORGAM variable: "Did you ever smoke cigarettes regularly in the past?"	Integer	Tobacco	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM

[Link to this search](#)

Functionalities of the Search Tool (16/22)

Results as comparison table of variable classes, by the studies

Some variable-level criteria need to be selected to see the comparison table.

Tobacco x OR Diseases of the circulatory system (I00-I99) x OR Heart failure (I50) x
Individual x AND Cohort x AND General population x AND Biomarkers | Imaging x

Lists view Comparison table

To see the time when the data are collected: Data collection events

Study

Comparison table by studies selected

Data Collection Event Filter

All Individual Harmonization

Study	Cardiovascular related diseases x	Lifestyle and behaviours x	Diseases x
	Heart failure (I50) x	Tobacco x	Diseases of the circulatory system (I00-I99) x
KORA	10	67	313
Brianza	0 Null cell	42	306
Catalonia	20	30	252
DAN-MONICA	43	66	455
ESTHER	2	12	108
Estonia	3	15	117

Click the number of variables to see the list of variables in this class for the study.

23 studies listed

Functionalities of the Search Tool (17/22)

Results as comparison table of variable classes, by the data collection events

Tobacco x OR Diseases of the circulatory system (I00-I99) x OR Heart failure (I50) x
Individual x AND Cohort x AND General population x AND Biomarkers | Imaging x

Lists view **Comparison table**

Data collection event selected

Study Dataset
All Individual Harmonization

To filter the search results to include those rows that includes variables in all classes, i.e. remove those having null cell ("full coverage").

Data Collection Event Filter

Population/Data Collection Event (DCE)

Study Population Data Collection Event Cardiovascular related diseases x Lifestyle and behaviours x Diseases x
Heart failure (I50) x Tobacco x Diseases of the circulatory system (I00-I99) x

278 1 213 7 269

KORA KORA (Augsburg) Study Cohort 01 (Survey S1) KORA S1 (Augsburg) Study Cohort 01 Baseline 1984-10 to 1985-05 2 16 17

KORA S1 (Augsburg) Study Cohort 01 mortality and disease outcome follow-up 1984-10 to 2009-12 0 Null cells 0 27

KORA (Augsburg) Study Cohort 02 (Survey S2) KORA S2 (Augsburg) Study Cohort 02 Baseline 1989-10 to 1990-06 2 15

[Link to this search](#) 17

Functionalities of the Search Tool (18/22)

Results as full coverage comparison table, by the studies

Tobacco x OR Diseases of the circulatory system (I00-I99) x OR Heart failure (I50) x
Individual x AND Cohort x AND General population x AND Biomarkers | Imaging x AND Acronym:... x

Lists view **Comparison table**

Study Dataset

All Individual Harmonization

Filtering “full coverage” selected and “data collection event” unselected

Data Collection Event
Full coverage
Subdomains with Variables **Filter**

Null cells removed, 11 studies listed

	Cardiovascular related diseases x	Lifestyle and behaviours x	Diseases x
Study	Heart failure (I50) x	Tobacco x	Diseases of the circulatory system (I00-I99) x
	278	758	4 494
KORA	10	67	313
Catalonia	20	30	252
DAN-MONICA	43	66	455
ESTHER	2	12	108
Estonia	3	15	117
FINRISK	77	96	929

Functionalities of the Search Tool (19/22)

Removing the full coverage restrictions, results by studies

Tobacco x OR Diseases of the circulatory system (I00-I99) x OR Heart failure (I50) x
Individual x AND Cohort x AND General population x AND Biomarkers | Imaging x AND Acronym... x

Lists view Comparison table

Variables 5 252 Datasets 80 Studies 11

All Individual Harmonization

Show 100 entries

Previous 1 Next

Acronym	Name	Type	Study design	Data types, sources, and samples available										Individual		Harmonization		
				👤	🏭	🖋️	🕒	❤️	🏠	☠️	🌐	🔍	🔄	Participants	Datasets	Variables	Datasets	Variables
Catalonia	Catalonia MONICA Study	Individual	Cohort	✓	✓	✓	-	-	-	✓	-	✓	✓	5505	4	282	-	-
DAN-MONICA	DAN-MONICA Study	Individual	Cohort	✓	✓	✓	-	-	-	✓	-	✓	✓	7582	8	521	-	-
ESTHER	ESTHER Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	✓	✓	✓	9949	2	120	-	-
Estonia	Estonian Biobank	Individual	Cohort	✓	✓	✓	✓	-	-	✓	✓	✓	✓	210000	2	132	-	-
FINRISK	FINRISK Study	Individual	Cohort	✓	✓	✓	✓	-	-	✓	-	✓	✓	75000	12	1,025	-	-

1) Select list view and results by studies.

2) Remove the restriction to the specific studies by clicking X.

Functionalities of the Search Tool (20/22)

Adding search criteria in variable-level - Searching text in the variable label

Variable name & label

Here text “daily” is search to match the text given in the variable label because variables concerning daily smoking is of interest.

Name	Label
Variable name.	Variable label.
<input type="text"/>	<input type="text" value="daily"/>

Click:

Functionalities of the Search Tool (21/22)

Results by variables when also matching text with variable label

Tobacco x OR Diseases of the circulatory system (I00-I99) x OR Heart failure (I50) x AND Label:match(daily) x
Individual x AND Cohort x AND General population x AND Biomarkers | Imaging x

Lists view Comparison table

List view of variables selected

AND-operator should be used here, as daily smoking variables are of interest (i.e. categorized to tobacco AND label contains "daily").

Variables 287 Datasets 80 Studies 23
All Individual Harmonization

Show 100 entries

Previous 1 2 3 N

"MAXCIGS" is shown later as an example of the variable description.

Name	Label	Value type	Annotations	Type	Study	Population	Data Collection Event	Dataset
<input type="checkbox"/> MAXCIGS	MORGAM variable: "What is the highest average daily number of cigarettes you have ever smoked for as long as a year?"	Integer	Tobacco	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM
<input type="checkbox"/> DSMOKER	MORGAM variable: Daily cigarette smoker	Integer	Tobacco	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM
<input type="checkbox"/> STOPAGE	MORGAM variable: Age when the person stopped smoking cigarettes daily	Integer	Tobacco	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM

Functionalities of the Search Tool (22/22)

Saving the search query for later use

Copy-paste the URL in the address bar and save the URL for later use.
That link can be used to do the the exactly same search again.

[https://mica.eucanshare.bsc.es/search#lists?type=variables&query=variable\(limit\(0,100\),and\(or\(or\(in\(Mlstr_area.Lifestyle_behaviours,\(Tobacco\)\),in\(Mlstr_area.Diseases.\(Circulatory_sys_dis\)\)\),in\(Cardiovascular_and_related_diseases.Cardiovascular_related_diseases,\(Heart_failure\)\)\),match\(daily,Mica_variable.label\)\)\),study\(limit\(0,100\),and\(and\(in\(Mica_study.className,Study\),in\(Mica_study.methods-design,\(cohort_study\)\)\),in\(Mica_study.populations-recruitment-dataSources,\(general_population\)\)\),in\(Mica_study.populations-dataCollectionEvents-dataTypes,\(biomarkers.imaging\)\)\)\)\)\)](https://mica.eucanshare.bsc.es/search#lists?type=variables&query=variable(limit(0,100),and(or(or(in(Mlstr_area.Lifestyle_behaviours,(Tobacco)),in(Mlstr_area.Diseases.(Circulatory_sys_dis))),in(Cardiovascular_and_related_diseases.Cardiovascular_related_diseases,(Heart_failure))),match(daily,Mica_variable.label))),study(limit(0,100),and(and(in(Mica_study.className,Study),in(Mica_study.methods-design,(cohort_study))),in(Mica_study.populations-recruitment-dataSources,(general_population))),in(Mica_study.populations-dataCollectionEvents-dataTypes,(biomarkers.imaging)))))))

are Catalogue Home Repository

Query

Tobacco OR Diseases of the circulatory system (I00-I99) OR Heart failure (I50) AND Label:match(daily)
Individual AND Cohort AND General population AND Biomarkers | Imaging

Lists view Comparison table

Variables 287 Datasets 80

Studies 23

All Individual Harmonization

For example the search query done in the previous slide:

[https://mica.eucanshare.bsc.es/search#lists?type=variables&query=variable\(limit\(0,100\),and\(or\(or\(in\(Mlstr_area.Lifestyle_behaviours,\(Tobacco\)\),in\(Mlstr_area.Diseases.\(Circulatory_sys_dis\)\)\),in\(Cardiovascular_and_related_diseases.Cardiovascular_related_diseases,\(Heart_failure\)\)\),match\(daily,Mica_variable.label\)\)\),study\(limit\(0,100\),and\(and\(in\(Mica_study.className,Study\),in\(Mica_study.methods-design,\(cohort_study\)\)\),in\(Mica_study.populations-recruitment-dataSources,\(general_population\)\)\),in\(Mica_study.populations-dataCollectionEvents-dataTypes,\(biomarkers.imaging\)\)\)\)\)\)](https://mica.eucanshare.bsc.es/search#lists?type=variables&query=variable(limit(0,100),and(or(or(in(Mlstr_area.Lifestyle_behaviours,(Tobacco)),in(Mlstr_area.Diseases.(Circulatory_sys_dis))),in(Cardiovascular_and_related_diseases.Cardiovascular_related_diseases,(Heart_failure))),match(daily,Mica_variable.label))),study(limit(0,100),and(and(in(Mica_study.className,Study),in(Mica_study.methods-design,(cohort_study))),in(Mica_study.populations-recruitment-dataSources,(general_population))),in(Mica_study.populations-dataCollectionEvents-dataTypes,(biomarkers.imaging)))))))

Finding the harmonization potential across the studies

- The variable descriptions from different studies can be compared in the catalogue to find the possible definition of the harmonized variable.
- The variable descriptions of the catalogue includes
 - Variable definition, type and unit
 - Category values and labels for the categorical variables
 - In some cases, the number of available and missing values

Finding the harmonization potential (1/5)

Viewing the variable description of the specific data collection event (1/2)

Here presented is the variable description of the "MAXCIGS" collected in the baseline of the KORA S1 Survey.

MAXCIGS Variable name

MORGAM variable: "What is the highest average daily number of cigarettes you have ever smoked for as long as a year?" Variable label

["See the specific description of the variable in the MORGAM website."](#) Possible broader description of the variable

Overview

Value type Integer
Nature Continuous
Entity type Participant

Possible category names and labels (MAXCIGS is continuous variable, and only missing categories are defined)

Categories

Name	Label	Missing
888	irrelevant if EVERCIG = 2	✓
999	insufficient data	✓

Definition

Dataset KORA0101_MORGAM
Study KORA
Population KORA (Augsburg) Study Cohort 01 (Survey S1)
Data Collection Event KORA S1 (Augsburg) Study Cohort 01 Baseline

Classes in which the variable is classified i.e. annotations

Annotations

Source Questionnaire
Target Participant
Lifestyle and behaviours Tobacco

Annotations: "Questionnaire" = information was collected in a questionnaire, "Participant" = information is about the study participant, "Tobacco" = information is about the consumption tobacco of in any form

Finding the harmonization potential (2/5)

Viewing the variable description of the specific data collection in the study (2/2)

Summary statistics of the variable show the availability information for the variable in the specific dataset (here MAXCIGS in KORA S1 Baseline), **if the dummy data on the missingness are provided.**

Summary Statistics

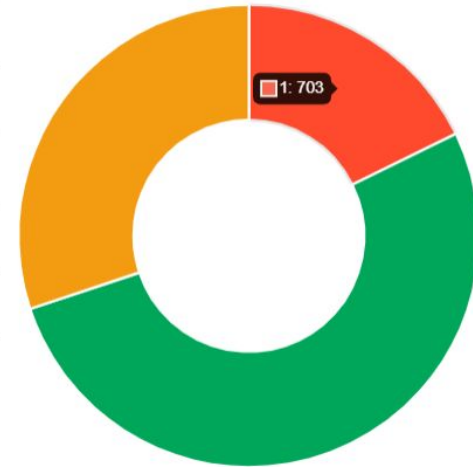
N	3,969
N with values	2,774
N missings	1,195

Number of observations in the dataset

Number of missing values

Proportions of the missing categories as pie chart

1 2 3



“MAXCIGS” variable is relevant only for the current or ex-smokers, and “replaceable to valid” for ever non-smokers.

Missing categories defined for the study and the frequencies for the available and missing values

Frequencies

Value	Frequency	% with values	% missings	%	Missing
1 Valid value	703	25.34		17.71	
2 Missing, but replaceable to valid	2,071	74.66		52.18	
3 Missing (not by design)	1,195		100	30.11	✓
4 Missing by design	0		0	0	✓
5 Not applicable	0		0	0	✓
N/A Empty values	0		0	0	✓

Finding the harmonization potential across the studies

- On next slides, a simple example set of variables shows the ways that data catalogue can be used for finding:
 - harmonization potential across the studies/cohorts
 - common definition for the harmonized variable
- Small number of similar variables on high blood pressure diagnoses across studies are selected (selected from [this list](#)):
 - “HIBP” of KORA S1 and ESTHER Baselines (harmonized MORGAM variable)
 - “hcbphigh” of CAHHM CPTP cohort Enrollment
 - “s0_hypert” of SHIP-START-0
 - variables “6150_0_0”, “6150_0_1”, “6150_0_2”, and “6150_0_3” of UK Biobank Recruitment

Finding the harmonization potential (3/5)

Downloading the results as a CSV file

- 1) Tick the desired variables
- 2) Click "Download"
- 3) Open CSV

Query

Lists view Comparison table

Variables 14 Datasets 14 Studies 5

All Individual Harmonization

Show 100 entries

Previous 1 Next

<input type="checkbox"/>	Name	Label	Value type	Annotations	Type	Study	Population	Data Collection Event	Dataset
<input checked="" type="checkbox"/>	HIBP	i MORGAM variable: "Have you ever been told by a doctor or other health worker that you have high blood pressure?"	Integer	i Diseases of the circulatory system (I00-I99)	Collected	KORA	KORA (Augsburg) Study Cohort 01 (Survey S1)	KORA S1 (Augsburg) Study Cohort 01 Baseline	KORA0101_MORGAM
<input type="checkbox"/>	HIBP	i MORGAM variable: "Have you ever been told by a doctor or other	Integer	i Diseases of the circulatory system (I00-I99)	Collected	KORA	KORA (Augsburg) Study Cohort 02 (Survey S2)	KORA S2 (Augsburg) Study Cohort 02	KORA0201_MORGAM

1)

2)

3)

 Variables (10).csv

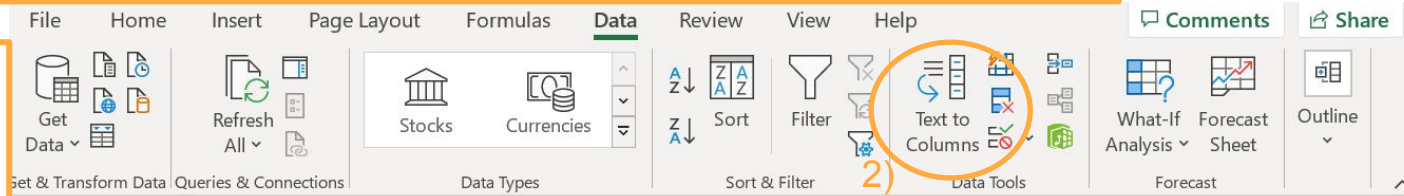
CSV file (comma ", " separated table of the search results) is downloaded. The file can be opened using Excel.



Finding the harmonization potential (4/5)

Opening the results and comparing variables

- 1) In Excel, select column A
- 2) In "Data" tab click "Text To Columns" and follow the converting steps (delimited by comma ",") ->
- 3) Splitted to the columns.



	A	B	C	D	E	
1	Name	Label	Unit	Type of	Categories	Typ
2	HIBP	MORGAM variable: "Have you ever been told by a doctor or other health worker that you have high blood pressure?"		Integer	1: yes 2: no 9: insufficient data	
3	hcbphigh	Have you ever been told by a health care professional that your blood pressure was high (excluding during pregnancy)?		Textual		
4	HIBP	MORGAM variable: "Have you ever been told by a doctor or other health worker that you have high blood pressure?"		Integer	1: yes 2: no 9: insufficient data	
5	s0_hypert	Have you ever been diagnosed with high blood pressure by a doctor?		Integer	1: ja 2: nein 8: wei? nicht 9: refused to answer	
6	6150_0_0	Touchscreen Health and medical history Medical conditions: Vascular heart problems diagnosed by doctor		Textual	1: Heart attack 2: Angina 3: Stroke 4: High blood pressure -7: None of the above -3: Prefer not to answer	Col

Note that using "Text To Columns" for Categories delimited by "|" (copy first column E to the most right of the table), classes can be separated further into different cells.

All variables 6150_0_0 - 6150_0_3 have the same description: ACE touchscreen question: Has a doctor ever told you that you have had any of the following conditions? (You can select more than one answer).

Finding the harmonization potential (5/5)

Comparing variables across studies

The definition of [this CAHMM variable](#) needs more investigation as the category values are not given in the catalogue. Here it is assumed that yes & no categories are included.



	A	B	C	D	K	L	M	N
1	Name	Label	Unit	Type of	Categories			
2	HIBP	MORGAM variable: "Have you ever been told by a doctor or other health worker that you have high blood pressure?"		Integer	1: yes	2: no	9: insufficient data	
	hcbphigh	Have you ever been told by a health care professional that your blood pressure was high (excluding during pregnancy)?		Textual	?: yes?	?: no?		
4	HIBP	MORGAM variable: "Have you ever been told by a doctor or other health worker that you have high blood pressure?"		Integer	1: yes	2: no	9: insufficient data	
5	s0_hypert	Have you ever been diagnosed with high blood pressure by a doctor?		Integer	1: yes	2: no	8: Don't know	9: Refused
6	6150_0_3	Touchscreen Health and medical history Medical conditions: 4: high blood pressure diagnosed by doctor		Textual	?: yes?	?: no?		-3: Prefer not to answer

The definition of the UKBB variable needs more investigation, but here it is assumed that [this 4th variable](#) indicates the answers to the "4 - High blood pressure" option in the questionnaire.



"Has a doctor ever told you that you have had any of the following conditions?" (You can select more than one answer).

A possible definition for a harmonized variable:

"Have you ever been told by a doctor or other health professional that you have high blood pressure?"

- 1: yes
- 2: no
- 9: missing

Example search

Question: 1) Which cohort studies have collected variables on all these categories: cerebrovascular diseases, alcohol, medication/supplement intake, and quality of life?

2) What are the names of the quality of life variables in these studies?

Solution: Filtering by both studies and variables is needed and the results must be shown as the comparison table for further filtering off the null cells of the variable categories.

Solution for example, step 1/5 – Start a new search: <https://mica.eucanshare.bsc.es/search#>
Click "Individual" and go to...

Individual

Filter by Studies

Study properties

...Study properties
Select "cohort"

Study properties

Study properties as defined in the catalogue.

Study design

Select All Clear Selection

The design of an observational or experimental study (e.g. cohort, case control).

- Cohort Case-control Case only
 Cross-sectional Clinical trial Other

Click "Display results" and go to...

Display results

Filter by Variables

General classification

CV rel. diseases variables

...Cardiovascular related disease variables
Select "cerebrovascular diseases (I60-I69)"

Cardiovascular related diseases

Select All Clear Selection

Display results

- | | | |
|---|--|---|
| <input type="checkbox"/> Diabetes mellitus (E10-E14) | <input type="checkbox"/> Disorders of lipoprotein metabolism and other lipidaemias (E78) | <input type="checkbox"/> Hypertensive diseases (I10-I15) |
| <input type="checkbox"/> Ischaemic heart diseases (I20-I25) | <input type="checkbox"/> Valve disorders (I34-I37) | <input type="checkbox"/> Conduction disorders and cardiac arrhythmias (I44-I49) |
| <input type="checkbox"/> Heart failure (I50) | <input type="checkbox"/> Diseases of the circulatory system falling into multiple categories | <input checked="" type="checkbox"/> Cerebrovascular diseases (I60-I69) |

Solution for example, step 2/5 - Go to "general classification"

Filter by Variables

General classification

Find and select "alcohol", "medication and supplement intake" and "quality of life"

Lifestyle and behaviours

Select All Clear Selection

Information about past and current lifestyle, behaviours and activities.

- Tobacco
- Alcohol
- Nutrition
- Breastfeeding
- Transportation
- Personal hygiene
- Sexual behaviours and orientation
- Leisure activities
- Misbehaviour and criminality
- Other and unspecified information
- Drugs
- Physical activity
- Sleep
- Technological

Perception of health, quality of life, development and functional limitations

Select All Clear Selection

Information about perception of general health, quality of life, child development and decline of functional capacities.

- Perception of health
- Quality of life
- Functional limitations
- Use of assistive devices
- Life course development
- Other perception of health, quality of life and functional limitation-related information

Medication and supplements

Select All Clear Selection

Information about medication (whether prescribed or over the counter), including drugs and supplements (e.g. vitamins, plant extracts) used to treat or prevent diseases or to alleviate symptoms of diseases.

- Medication and supplement intake
- Posology and protocol of administration
- Other and unspecified pharmacological interventions

More

Then click:

Display results

Solution for example, step 3/5 - Check the query

OR-operator is needed here as 1 variable can't be classified to all these classes at the same time.

Lists view Comparison table

Study Dataset

All Individual Harmonization

Variables 3 527 Datasets 158 Studies 25

Select Filter: "full coverage"

Full coverage
Subdomains with Variables

Study	Cardiovascular related diseases ×	Lifestyle and behaviours ×	Perception of health, quality of life, development and functional limitations ×	Medication and supplements ×
Study	Cerebrovascular diseases (I60-I69) ×	Alcohol ×	Quality of life ×	Medication and supplement intake ×
	1 608	531	19	1 369
KORA	62	20	0	15
Brianza	81	12	0	14
CAHHM	0	148	12	290
Catalonia	54	2	0	11

25 studies have variables in some of these classes. Then select comparison table to further filtering off the studies that don't have all these variable classes collected.

Null cells

Solution for example, step 4/5 - See the full coverage results

Search filters: Cerebrovascular diseases (I60-I69) x OR Alcohol x OR Medication and supplement intake x OR Quality of life x Individual x AND Cohort x AND UKBB x

Lists view

Comparison table

1 study, UKBB, has variables in all these classes.

Study Dataset Data Collection Event Filter

All Individual Harmonization

Study	Cardiovascular related diseases x	Lifestyle and behaviours x	Perception of health, quality of life, development and functional limitations x	Medication and supplements x
	Cerebrovascular diseases (I60-I69) x	Alcohol x	Quality of life x	Medication and supplement intake x
	20	109	7	677
UKBB	20	109	7	677

Click to see the list of quality of life variables in the UKBB study

Solution for example, step 5/5 - See the list of variables

Lists view Comparison table

Variables 7 Datasets 3 Studies 1

7 quality of life variables

All Individual Harmonization

Show 20 entries

Previous 1 Next

<input type="checkbox"/>	Name	Label	Value type	Annotations	Type	Study	Population	Data Collection Event	Dataset
<input type="checkbox"/>	26413_0_0	Baseline characteristics Indices of Multiple Deprivation: Health score (England)	Decimal	Quality of life	Collected	UKBB	UK Biobank	UK Biobank - Recruitment	UKBB_Baseline_assessment
<input type="checkbox"/>	26420_0_0	Baseline characteristics Indices of Multiple Deprivation: Health score (Wales)	Decimal	Quality of life	Collected	UKBB	UK Biobank	UK Biobank - Recruitment	UKBB_Baseline_assessment
<input type="checkbox"/>	26430_0_0	Baseline characteristics Indices of Multiple Deprivation: Health score (Scotland)	Decimal	Quality of life	Collected	UKBB	UK Biobank	UK Biobank - Recruitment	UKBB_Baseline_assessment
<input type="checkbox"/>	20458_1_0	Mental health Happiness and subjective wellbeing: General happiness	Text	Quality of life	Collected	UKBB	UK Biobank	UK Biobank - Mental health questionnaire	UKBB_Mental_health_enhancement_data
<input type="checkbox"/>	20459_1_0	Mental health Happiness and subjective wellbeing: General	Text	Quality of life	Collected	UKBB	UK Biobank	UK Biobank - Mental health questionnaire	UKBB_Mental_health_enhancement_data

All rows are not presented here

Answers to the example search

1) Which cohort studies have collected variables on all these categories: cerebrovascular diseases, alcohol, medication/supplement intake, and quality of life?

Answer: UK Biobank (UKBB)

[Link to this answer](#)

2) What are the names of the quality of life variables in these studies?

Answer: UKBB variable names are 26413_0_0, 26420_0_0, 26430_0_0, 20458_1_0, 20459_1_0, 120070_1_0, and 120097_1_0.

[Link to this answer](#)