Effecting rigorous data harmonization and documentation to understand data heterogeneity and quality

Tina W. Wey & Isabel Fortier

Maelstrom Research

Building Multi-Source Databases for Comparative Analyses

Warsaw, Poland, 17 December 2019

maelstrom

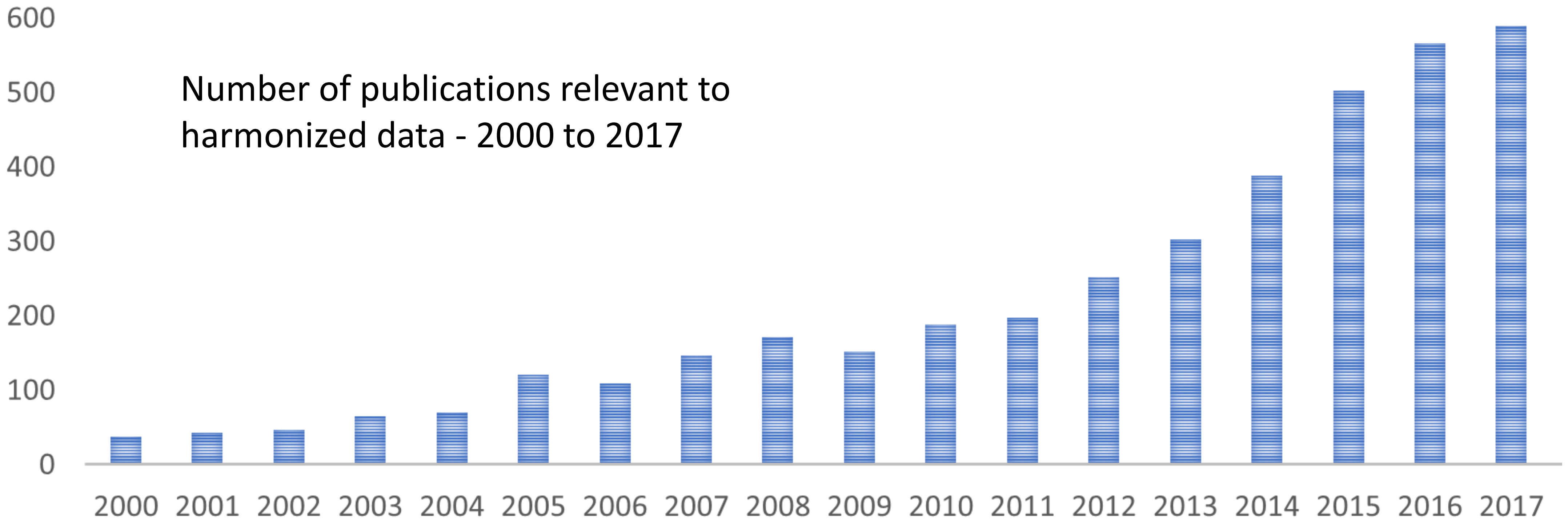
Centre universitaire de santé McGill
Institut de recherche

McGill University Health Centre
Research Institute

Increasing need for harmonized data in epidemiological research

Driven by need to obtain **larger sample sizes** and statistical power; conduct **comparative research** across studies/populations; **extend** the scientific **impact** of individual studies/data sources.

Offers benefits: enabling **timely access** to available data and samples, increasing **potential to share** data across studies, and promoting a **collaborative approach**



Study A

Study B

Study C

Maelstrom Research

Facilitate collaborative epidemiological research through rigorous data documentation, harmonization, integration, and co-analysis

Who we are:

Hosted at
Research Institute of the McGill University Health Centre in Montreal, Canada

International research program
partnering with over 15 international networks and research consortia

Multi-disciplinary team
epidemiologists, data analysts, and computer scientists

Activities:

Methodological guidelines/support
for data cataloguing, harmonization, integration, and co-analysis

Web-based catalogues and harmonization platforms
searchable metadata catalogues and platforms to generate common-format variables for co-analysis

Open-source software
for data cataloguing, harmonization, integration, and co-analysis

mælstrom

Methodological guidelines and open-source software to support data collection, management, dissemination, harmonization and co-analysis

DataSHaPER onyx opal mica DataSHIELD Secure Bioscience Collaboration

A central study catalogue to foster usage of available data

National and international platforms harmonizing, integrating and co-analysing data across studies

PLOS | ONE

RESEARCH ARTICLE

Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit

Julie Bergeron[1], Dan

1 Research Institute of th and Public Health Institute of the Sainte-Justine Univ

IJEA

International Journal of Epidemiology, 2016, 1–13
doi: 10.1093/ije/dyw075
Original Article

Original Article

Maelstrom Research guidelines for rigorous retrospective data harmonization

Isabel
Lauren
P Stolk
Peter C

IJEA

International Journal of Epidemiology, 2017, 1372–1378
doi: 10.1093/ije/dyx180
Advance Access Publication Date: 2 September 2017
Software Application Profile

Software Application Profile

Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination

Dany Doiron,[1–3*†]
Vincent Ferre

IJEA

International Journal of Epidemiology, 2014, 1929–1944
doi: 10.1093/ije/dyu188
Advance Access Publication Date: 26 September 2014
Original article

Data Matters

DataSHIELD: taking the analysis to the data, not the data to the analysis

Amadou Gaye,[1] Yannick Marcon,[2] Julia Isaeva,[3] Philippe LaFlamme,[2] Andrew Turner,[1] Elinor M Jones,[4] Joel Minion,[1] Andrew W Boyd,[1] Christopher J Newby,[5] Marja-Liisa Nuotio,[6,7] Rebecca Wilson,[1] Oliver Butters,[1] Barnaby Murtagh,[8] Ipek Demir,[9] Dany Doiron,[2] Lisette Giepmans,[10] Susan E Wallace,[8] Isabelle Budin-Ljøsne,[3] Carsten Oliver Schmidt,[11] Paolo Boffetta,[12] Mathieu Boniol,[12] Maria Bota,[12] Kim W Carter,[13] Nick deKlerk,[13] Chris Dibben,[14] Richard W Francis,[13] Tero Hiekkalinna,[6,7] Kristian Hveem,[15] Kirsti Kvaløy,[15] Sean Millar,[16] Ivan J Perry,[16] Annette Peters,[17] Catherine M Phillips,[16] Frank Popham,[18] Gillian Raab,[14] Eva Reischl,[17] Nuala Sheehan,[8] Melanie Waldenberger,[17] Markus Perola,[6,7,19] Edwin van den Heuvel,[20] John Macleod,[1] Bartha M Knoppers,[21] Ronald P Stolk,[10,22] Isabel Fortier,[2] Jennifer R Harris,[3] Bruce HR Woffenbuttel,[22,23] Madeleine J Murtagh,[24†] Vincent Ferretti[2,25†] and Paul R Burton[2,24†*]

Fortier, Isabel, et al. "Maelstrom Research guidelines for rigorous retrospective data harmonization." *International journal of epidemiology* (2016): dyw075.

Maelstrom harmonization guidelines

0 Define the research question(s)

1 Assemble information and select studies

1. Document individual study designs, methods and content
2. Select participating studies

2 Define variables and evaluate harmonization potential

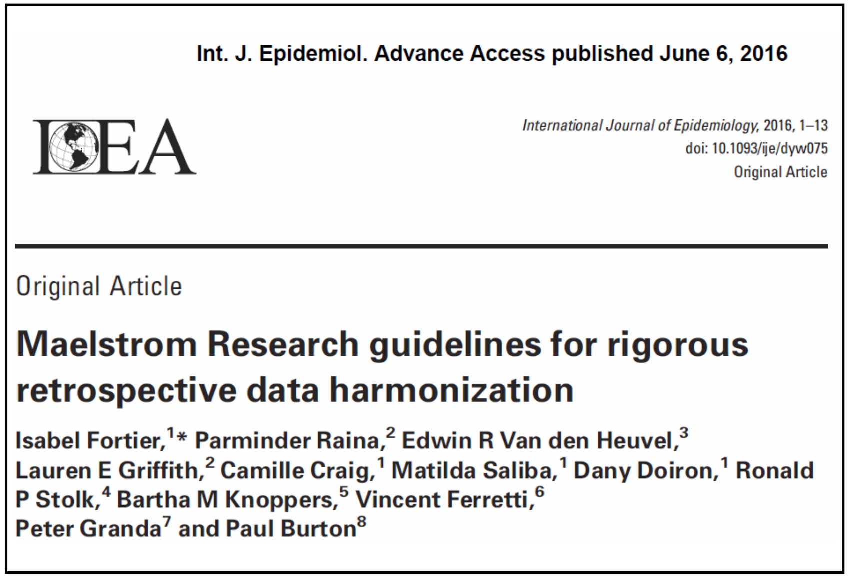
1. Select and define the core variables to be harmonized
2. Determine the potential to create the core variables using the study-specific data items

3 Process data

1. Ensure access to adequate study-specific data items and establish the overall data processing infrastructure
2. Process study-specific data under a common format to generate the harmonized datasets

4 Estimate quality of the harmonization dataset(s) generated

5 Disseminate and preserve final harmonization products



Int. J. Epidemiol. Advance Access published June 6, 2016

IEA

International Journal of Epidemiology, 2016, 1–13
doi: 10.1093/ije/dyw075
Original Article

Original Article

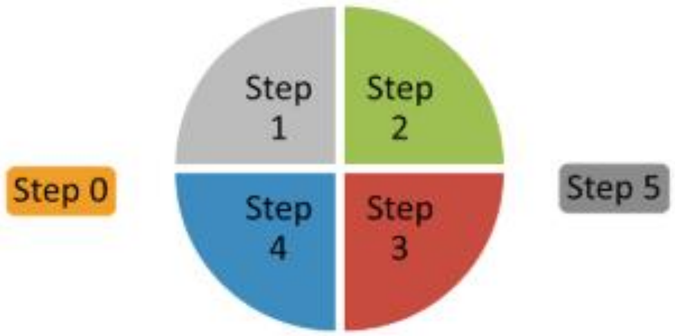
Maelstrom Research guidelines for rigorous retrospective data harmonization

Isabel Fortier,[1]* Parminder Raina,[2] Edwin R Van den Heuvel,[3] Lauren E Griffith,[2] Camille Craig,[1] Matilda Saliba,[1] Dany Doiron,[1] Ronald P Stolk,[4] Bartha M Knoppers,[5] Vincent Ferretti,[6] Peter Granda[7] and Paul Burton[8]

A systematic but adaptable process

Iterative, dynamic process of consideration, evaluation, discussion, validation

Documentation and assessment of source data heterogeneity to understand harmonized output



Iterative Harmonization Steps

Step 0: Define the research questions, objectives and protocol

Step 1: Assemble information and select studies

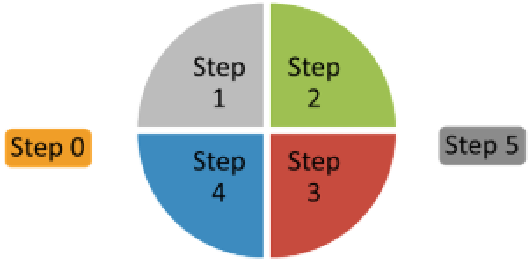
Step 2: Define variables and evaluate harmonization potential

Step 3: Process data

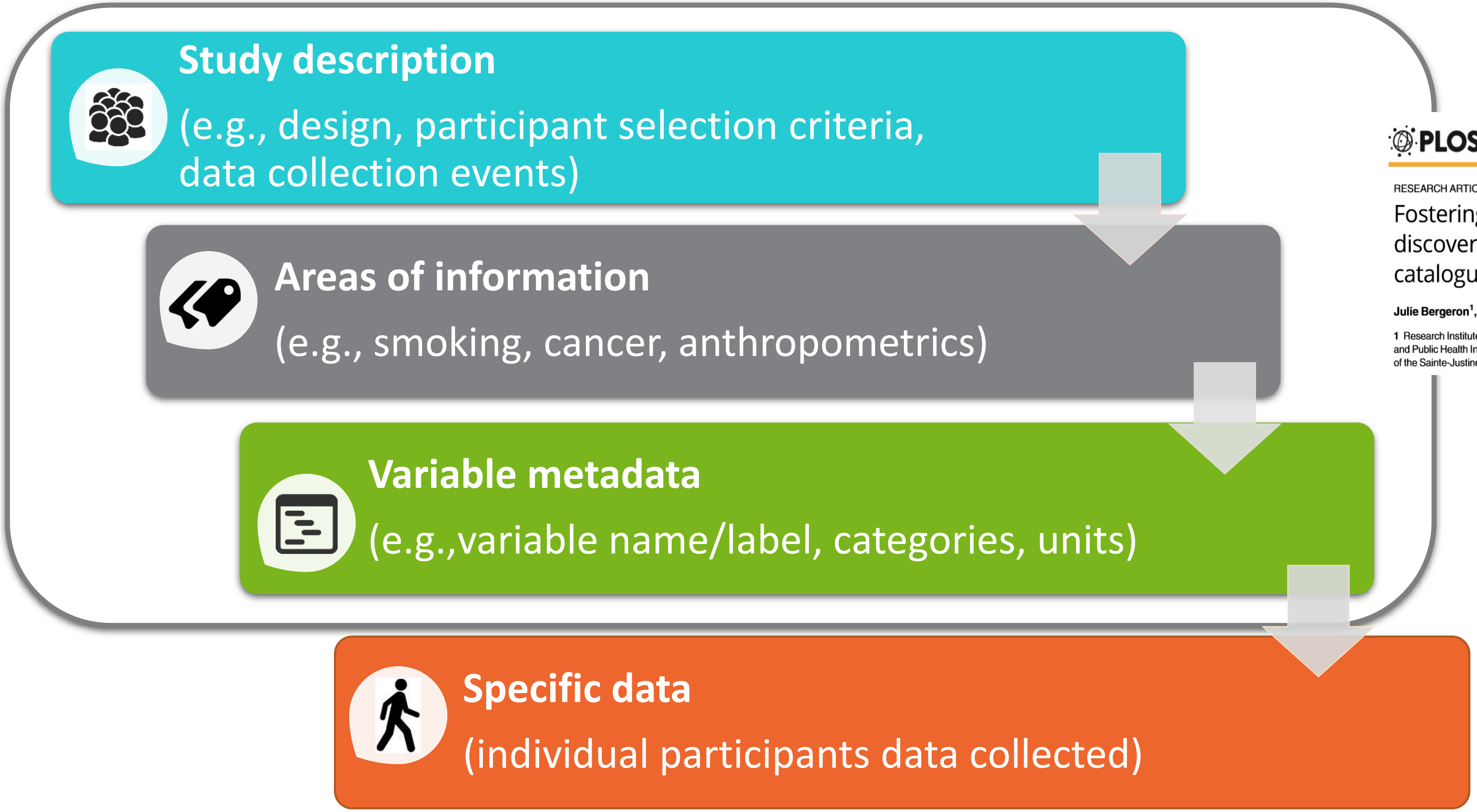
Step 4: Estimate quality of the harmonized dataset(s) generated

Step 5: Disseminate and preserve final harmonization products

Assemble information and select studies: Cohort metadata catalogue



Iterative Harmonization Steps



Study description

(e.g., design, participant selection criteria, data collection events)

Areas of information

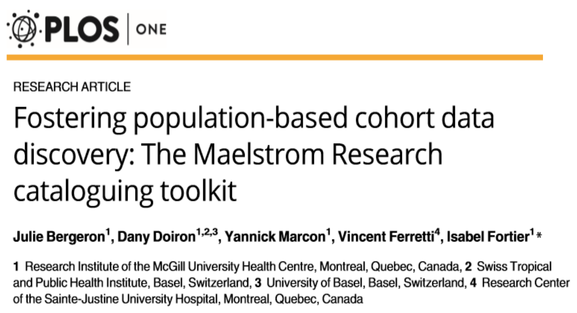
(e.g., smoking, cancer, anthropometrics)

Variable metadata

(e.g.,variable name/label, categories, units)

Specific data

(individual participants data collected)



PLOS | ONE

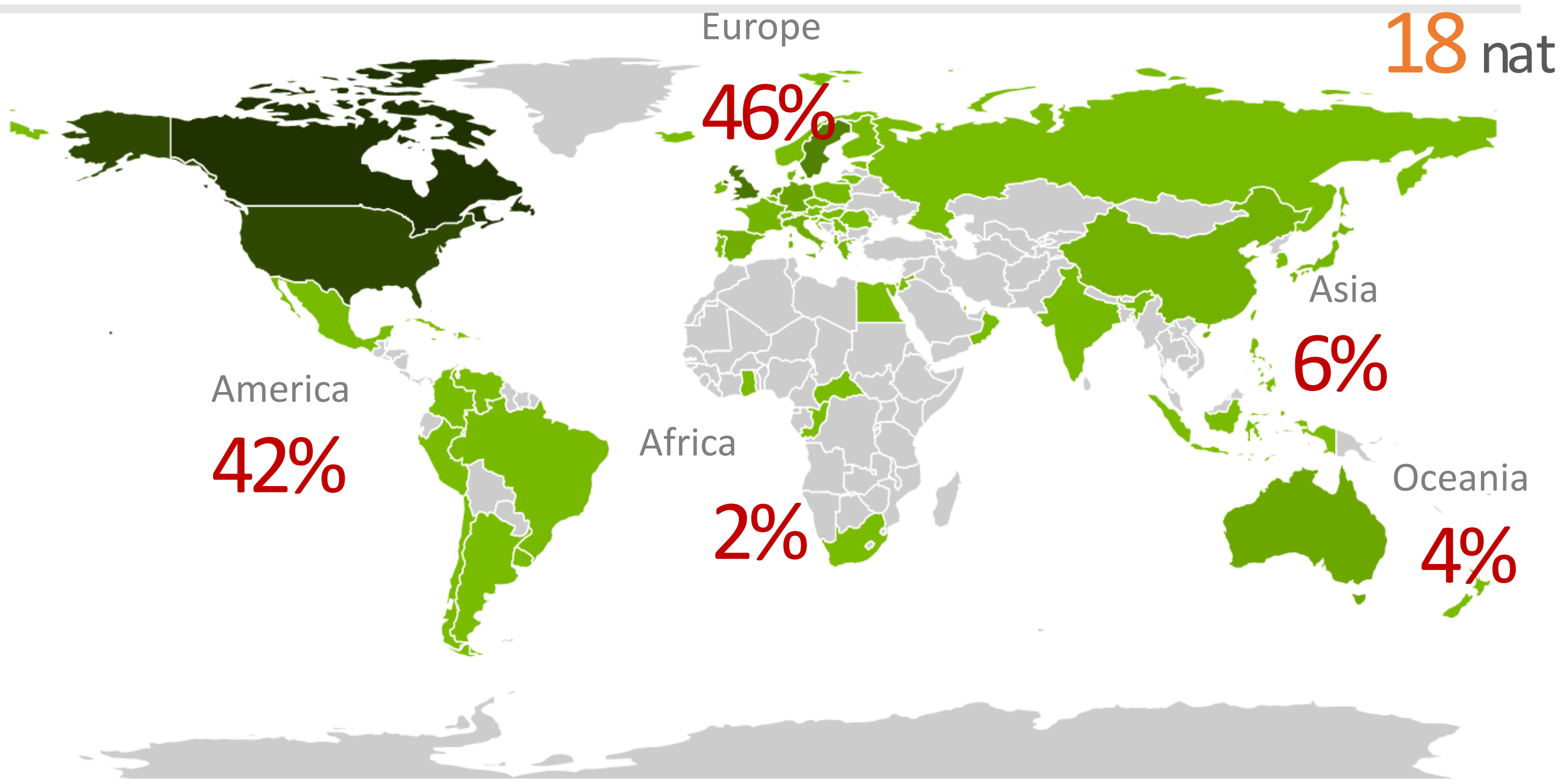
RESEARCH ARTICLE

Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit

Julie Bergeron[1], Dany Doiron[1,2,3], Yannick Marcon[1], Vincent Ferretti[4], Isabel Fortier[1]*

1 Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada, 2 Swiss Tropical and Public Health Institute, Basel, Switzerland, 3 University of Basel, Basel, Switzerland, 4 Research Center of the Sainte-Justine University Hospital, Montreal, Quebec, Canada

The Maelstrom Research metadata catalogue



18 national and international networks

204 studies (**122** with variables)

933,144 variables

6,349,772 cohort participants

NEAR

ReACH Research Advancement through Cohort Cataloguing and Harmonization

BioSHaRE.eu

Cohorts.se

MINDMAP Promoting mental well-being and healthy ageing in cities

Harmonized Approach to Rigorous Measurement Of the EnviroNmental EtiologY of pedMS

CHPT Cross-cohort harmonization project for tomorrow

Canadian Partnership for Tomorrow Project

ATHLOS

BRAIN - Broad and Deep Analyses in Neurodegeneration

PHQE: Quebec Europe data harmonization platform

SPIRIT - Sino-Quebec Perinatal Initiative in Research and Information Technology

Studies, including...

CONSTANCES

epic

NTNU HUNT Research Centre

clsa élcv Canadian Longitudinal Study on Aging Étude longitudinale canadienne sur le vieillissement

Canadian Partnership for Tomorrow Project

LifeLines

biobank uk

Nurses' Health Study

Illustrative harmonization projects

Canadian Partnership for Tomorrow Project

Environmental, lifestyle and genetic factors related to the development and progression of cancer and chronic diseases;
Prospective design; 5 Canadian provinces

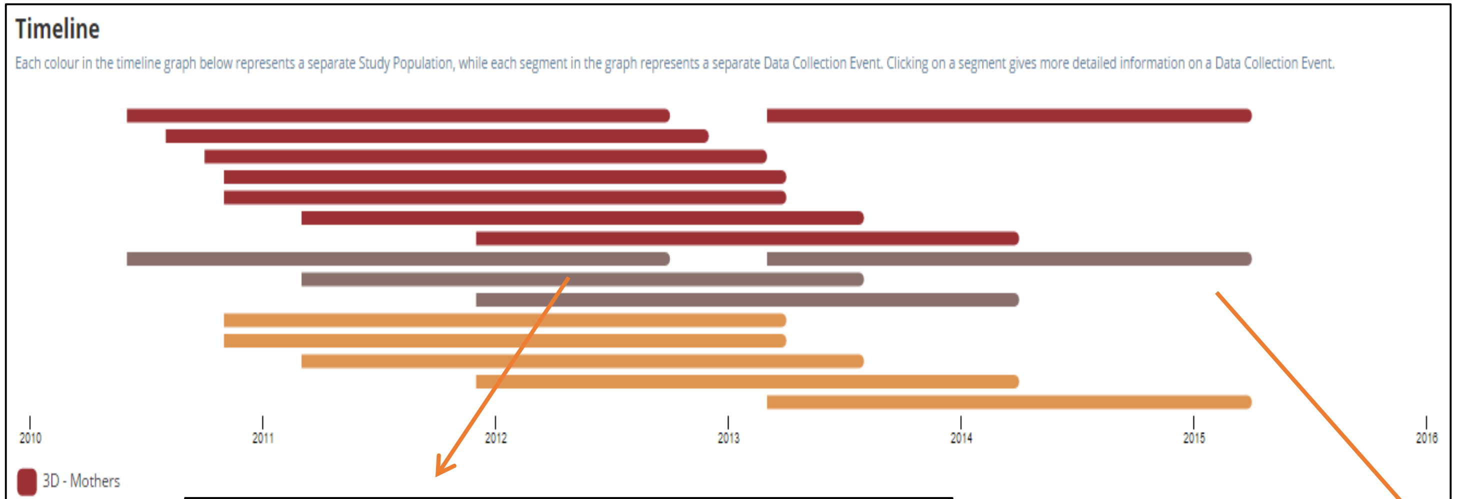
MINDMAP
Promoting mental well-being and healthy ageing in cities

Urban environments and promotion of mental wellbeing and cognitive function of older individuals;
Retrospective design; 7 European countries, Russia, and Canada

ReACH
Research Advancement through Cohort Cataloguing and Harmonization

Canadian pregnancy and birth cohorts data and biological samples to study Developmental Origins of Health and Disease (DOHaD);
Retrospective design; Canadian

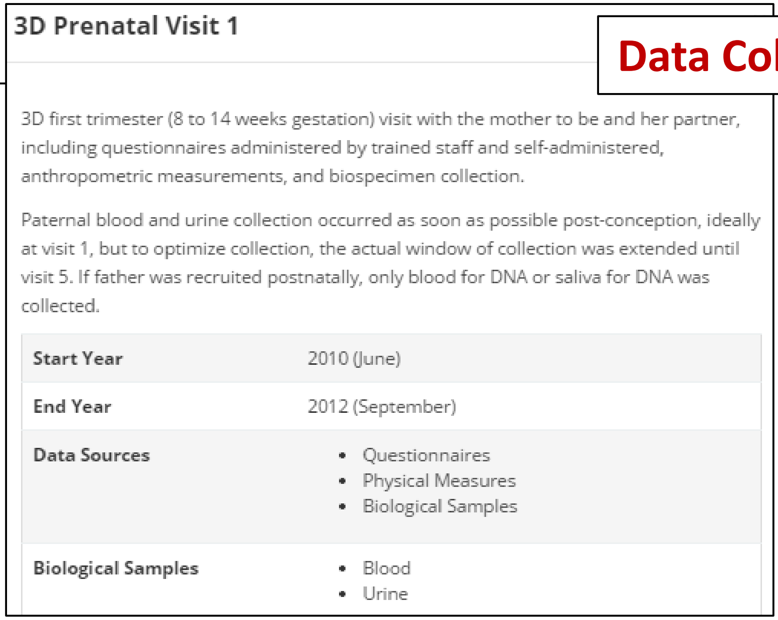
Study Description



Timeline

Each colour in the timeline graph below represents a separate Study Population, while each segment in the graph represents a separate Data Collection Event. Clicking on a segment gives more detailed information on a Data Collection Event.

🟥 3D - Mothers
🟫 3D - Fathers/Partners
🟧 3D - Infants



3D Prenatal Visit 1

3D first trimester (8 to 14 weeks gestation) visit with the mother to be and her partner, including questionnaires administered by trained staff and self-administered, anthropometric measurements, and biospecimen collection.

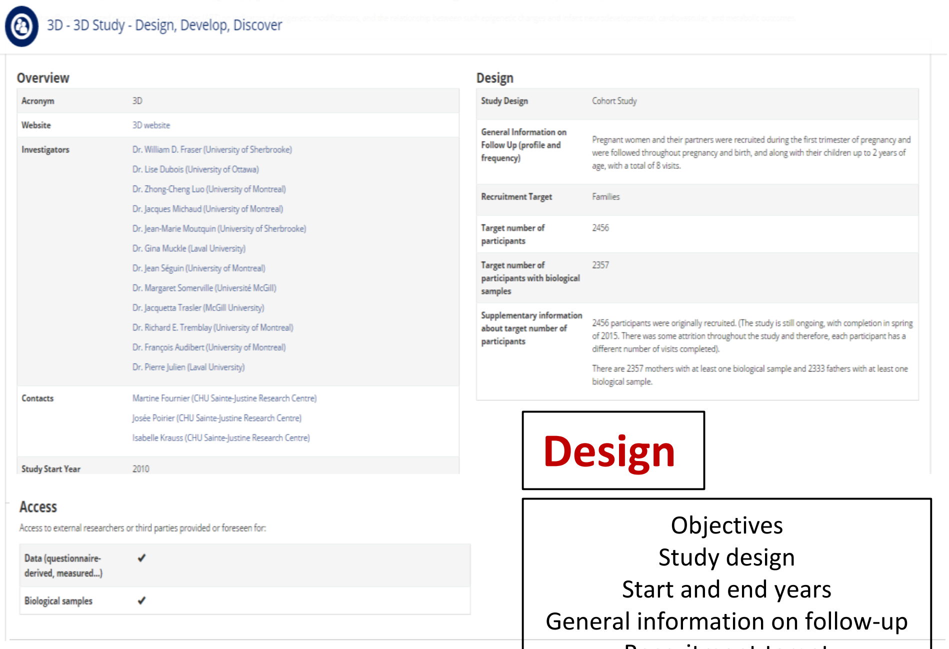
Paternal blood and urine collection occurred as soon as possible post-conception, ideally at visit 1, but to optimize collection, the actual window of collection was extended until visit 5. If father was recruited postnatally, only blood for DNA or saliva for DNA was collected.

Start Year	2010 (June)
End Year	2012 (September)
Data Sources	• Questionnaires • Physical Measures • Biological Samples
Biological Samples	• Blood • Urine

Data Collection Event

Description of the event
Start and end years
Data sources
Biological samples

ReACH
Research Advancement through Cohort Cataloguing and Harmonization



3D - 3D Study - Design, Develop, Discover

Overview

Acronym	3D
Website	3D website
Investigators	Dr. William D. Fraser (University of Sherbrooke) Dr. Lise Dubois (University of Ottawa) Dr. Zhong-Cheng Luo (University of Montreal) Dr. Jacques Michaud (University of Montreal) Dr. Jean-Marie Moutquin (University of Sherbrooke) Dr. Gina Muckle (Laval University) Dr. Jean Séguin (University of Montreal) Dr. Margaret Somerville (Université McGill) Dr. Jacquetta Trasler (McGill University) Dr. Richard E. Tremblay (University of Montreal) Dr. François Audibert (University of Montreal) Dr. Pierre Julien (Laval University)
Contacts	Martine Fournier (CHU Sainte-Justine Research Centre) Josée Poirier (CHU Sainte-Justine Research Centre) Isabelle Krauss (CHU Sainte-Justine Research Centre)
Study Start Year	2010

Access

Access to external researchers or third parties provided or foreseen for:

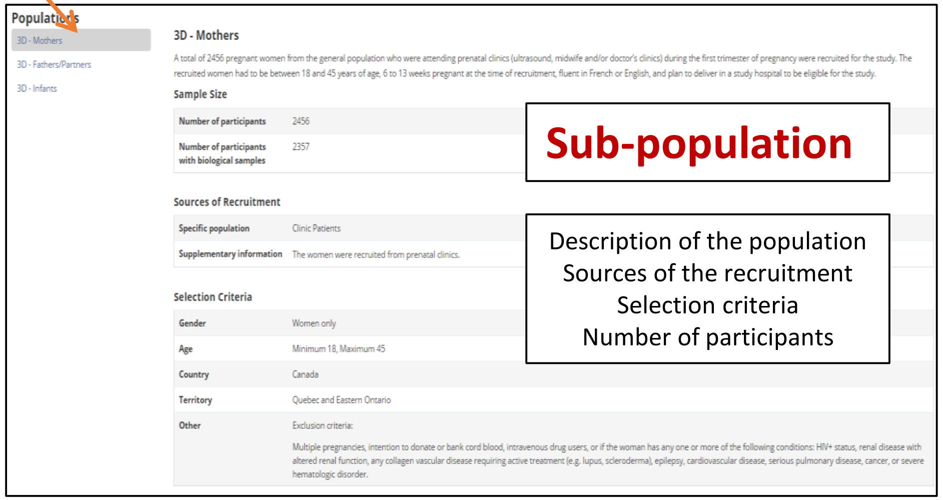
Data (questionnaire-derived, measured...)	✓
Biological samples	✓

Design

Study Design	Cohort Study
General Information on Follow Up (profile and frequency)	Pregnant women and their partners were recruited during the first trimester of pregnancy and were followed throughout pregnancy and birth, and along with their children up to 2 years of age, with a total of 8 visits.
Recruitment Target	Families
Target number of participants	2456
Target number of participants with biological samples	2357
Supplementary information about target number of participants	2456 participants were originally recruited. (The study is still ongoing, with completion in spring of 2015. There was some attrition throughout the study and therefore, each participant has a different number of visits completed). There are 2357 mothers with at least one biological sample and 2333 fathers with at least one biological sample.

Design

Objectives
Study design
Start and end years
General information on follow-up
Recruitment target
Number of participants



Populations

3D - Mothers
3D - Fathers/Partners
3D - Infants

3D - Mothers

A total of 2456 pregnant women from the general population who were attending prenatal clinics (ultrasound, midwife and/or doctor's clinics) during the first trimester of pregnancy were recruited for the study. The recruited women had to be between 18 and 45 years of age, 6 to 13 weeks pregnant at the time of recruitment, fluent in French or English, and plan to deliver in a study hospital to be eligible for the study.

Sample Size

Number of participants	2456
Number of participants with biological samples	2357

Sources of Recruitment

Specific population	Clinic Patients
Supplementary information	The women were recruited from prenatal clinics.

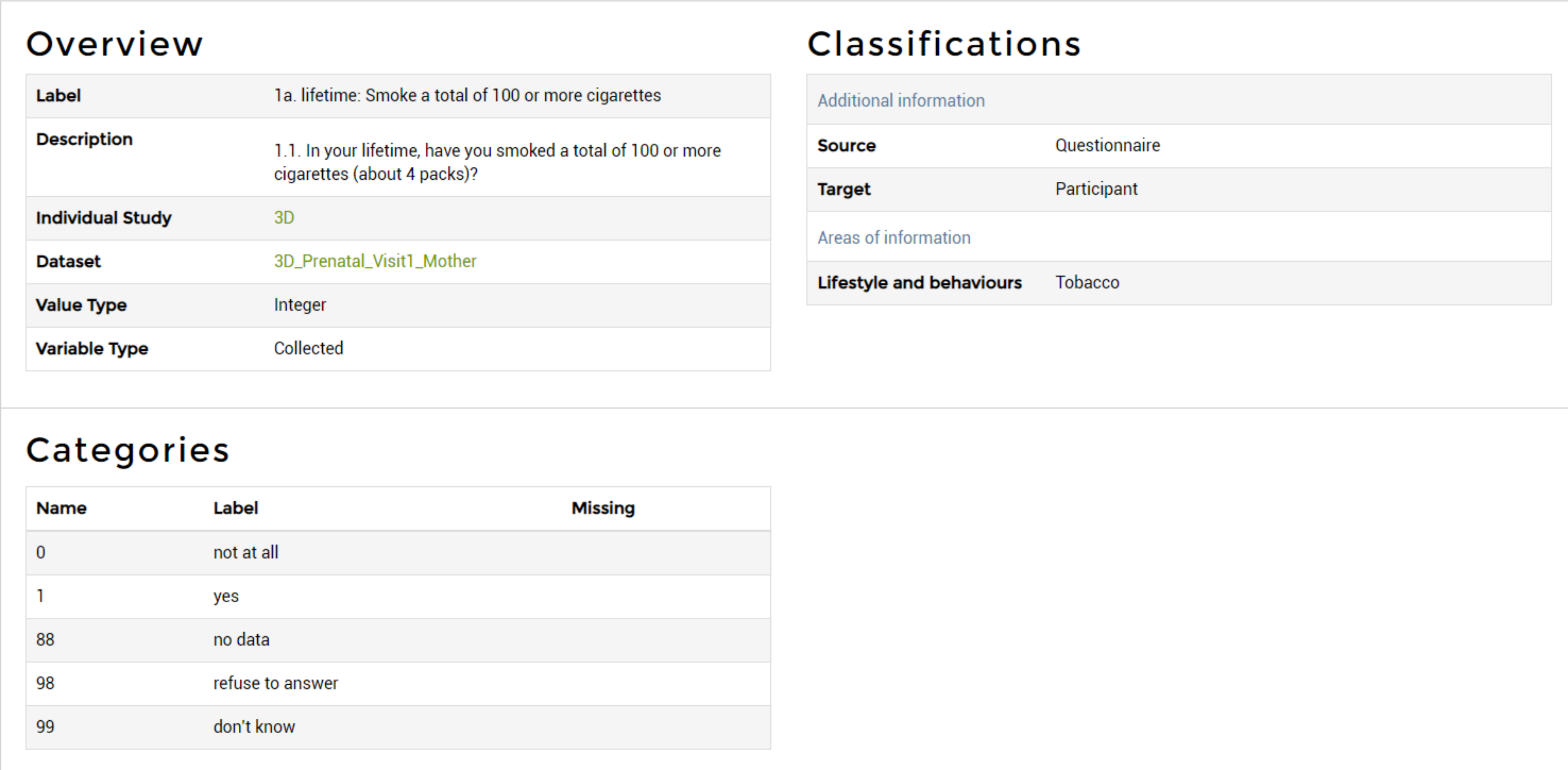
Selection Criteria

Gender	Women only
Age	Minimum 18, Maximum 45
Country	Canada
Territory	Quebec and Eastern Ontario
Other	Exclusion criteria: Multiple pregnancies, intention to donate or bank cord blood, intravenous drug users, or if the woman has any one or more of the following conditions: HIV+ status, renal disease with altered renal function, any collagen vascular disease requiring active treatment (e.g. lupus, scleroderma), epilepsy, cardiovascular disease, serious pulmonary disease, cancer, or severe hematologic disorder.

Sub-population

Description of the population
Sources of the recruitment
Selection criteria
Number of participants

Detailed study-specific source variable information



Overview

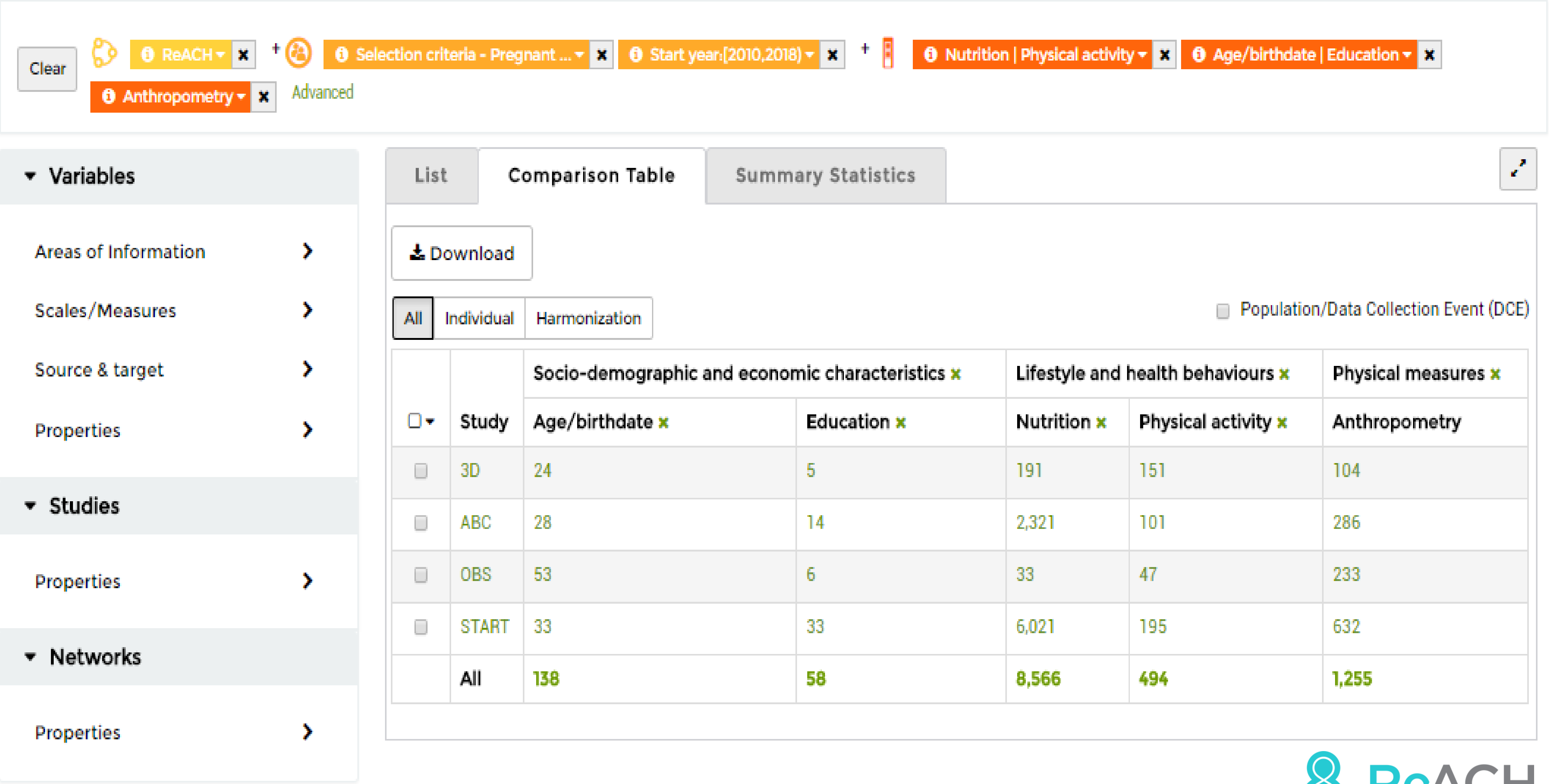
Label	1a. lifetime: Smoke a total of 100 or more cigarettes
Description	1.1. In your lifetime, have you smoked a total of 100 or more cigarettes (about 4 packs)?
Individual Study	3D
Dataset	3D_Prenatal_Visit1_Mother
Value Type	Integer
Variable Type	Collected

Classifications

Additional information	
Source	Questionnaire
Target	Participant
Areas of information	
Lifestyle and behaviours	Tobacco

Categories

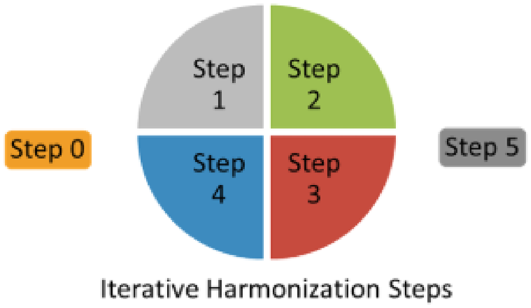
Name	**Label**	**Missing**
0	not at all	
1	yes	
88	no data	
98	refuse to answer	
99	don't know	



		Socio-demographic and economic characteristics		Lifestyle and health behaviours		Physical measures
☐	Study	Age/birthdate	Education	Nutrition	Physical activity	Anthropometry
☐	3D	24	5	191	151	104
☐	ABC	28	14	2,321	101	286
☐	OBS	53	6	33	47	233
☐	START	33	33	6,021	195	632
	All	138	58	8,566	494	1,255

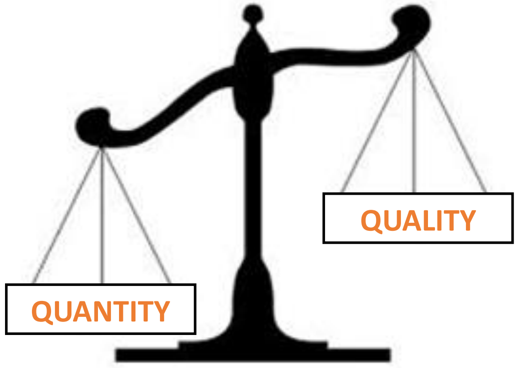
ReACH
Research Advancement through Cohort Cataloguing and Harmonization

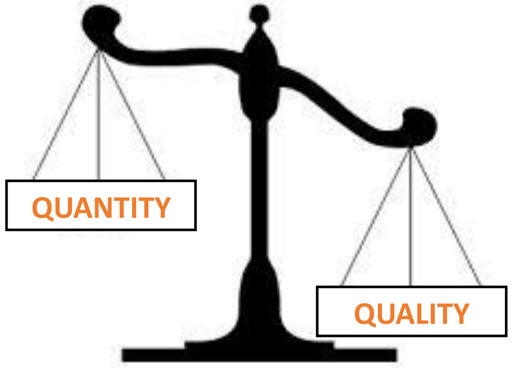
Define core variables (DataSchema)



__Quantity__ = number of studies to include

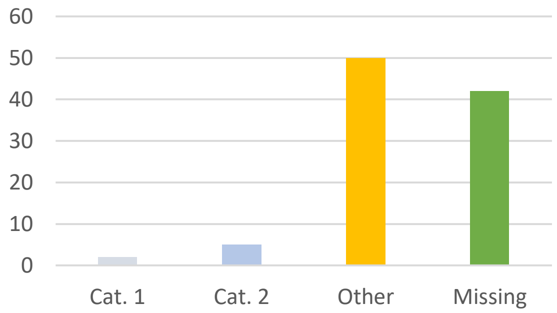
__Quality__ = scientific relevance/precision

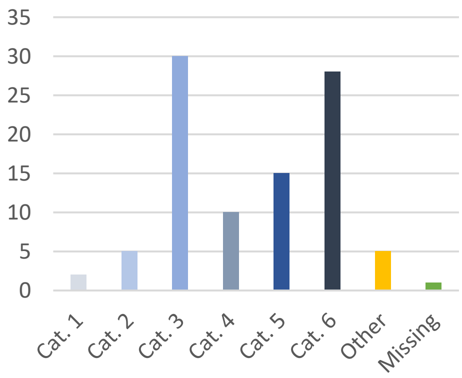




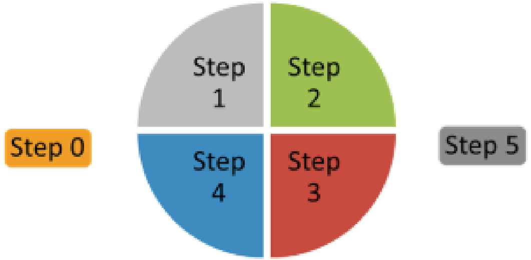
	Study 1	Study 2	Study 3	Study 4	Study 5
Variable X	✓	✓	✓	✓	✓

	Study 1	Study 2	Study 3	Study 4	Study 5
Variable X	✓	✗	✓	✗	✗





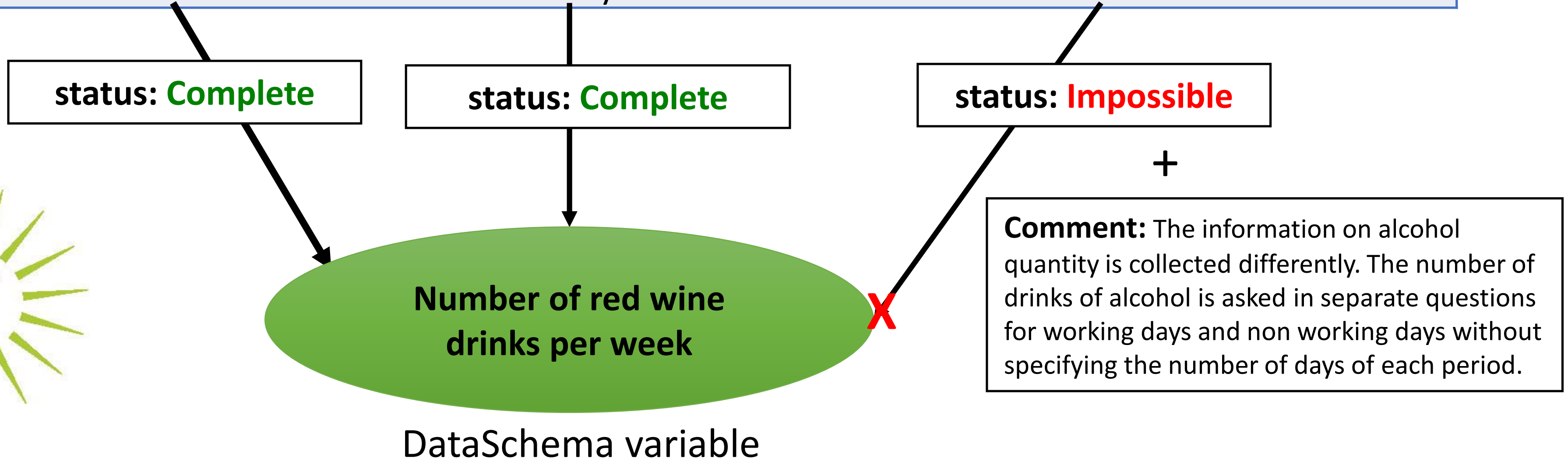
Evaluate harmonization potential



Iterative Harmonization Steps

Variable: Number of red wine drinks

Study A	Study B	Study C
Period: Week (7 days)	**Period:** Weekdays (Sunday to Thursday) Weekend days (Friday to Saturday)	**Period :** Weekday day (working day) Weekend day (non-working day)
Unit: Drinks/week	**Unit:** Per weekdays Per weekend days	**Unit:** Per day



status: Complete

status: Complete

status: Impossible

+

Comment: The information on alcohol quantity is collected differently. The number of drinks of alcohol is asked in separate questions for working days and non working days without specifying the number of days of each period.

Number of red wine drinks per week

DataSchema variable

Target variable: Frequency of Binge Drinking During Pregnancy

Study A

3 collections

1st collection : 8 – 14 weeks

Question: Since you have become pregnant, how often did you have 5 or more drinks on one occasion?

Response: #days of week OR #days of month OR #days since beginning of your pregnancy

2nd collection: 20 - 24 weeks

Question: Since your last visit, how often did you have 5 or more drinks on one occasion?

Response: #days of week OR #days of month OR #days since beginning of your pregnancy

3rd collection: 32 - 35 weeks

Question: Since your last visit, how often did you have 5 or more drinks on one occasion?

Response: #days of week OR #days of month OR #days since beginning of your pregnancy

Study B

2 collections

1st collection : 12 – 16 weeks

Question: Please specify the number of times per month you have four or more drinks at the same sitting or occasion (during this pregnancy)?

Response: >= 1times/month. Please specify number:___ | < 1/month | None

2nd collection : 28 – 32 weeks

Question: Over the past 3 months, how often did you have four or more drinks at the same sitting or occasion?

Response: 6 to 7 times a week | 4 to 5 times a week | 2 to 3 times a week | once a week | 2 to 3 times a month | about once a month | 6 to 11 times a year | 1 to 5 times a year | never

Study C

1 collection

1st collection : 21 – 39 weeks

Question: During this pregnancy, how many times have you consumed at least 5 or more drinks of alcohol in a day?

Response: continuous

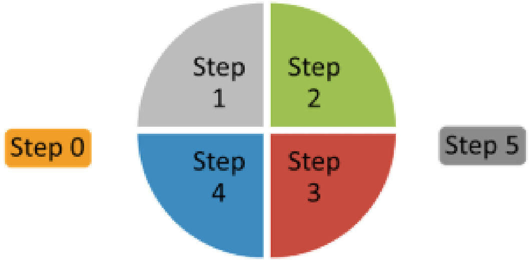
Challenges

- Timing
- Wording of questions
- Wording of categories
- Responses options
- Data collection events

ReACH

Research Advancement through Cohort Cataloguing and Harmonization

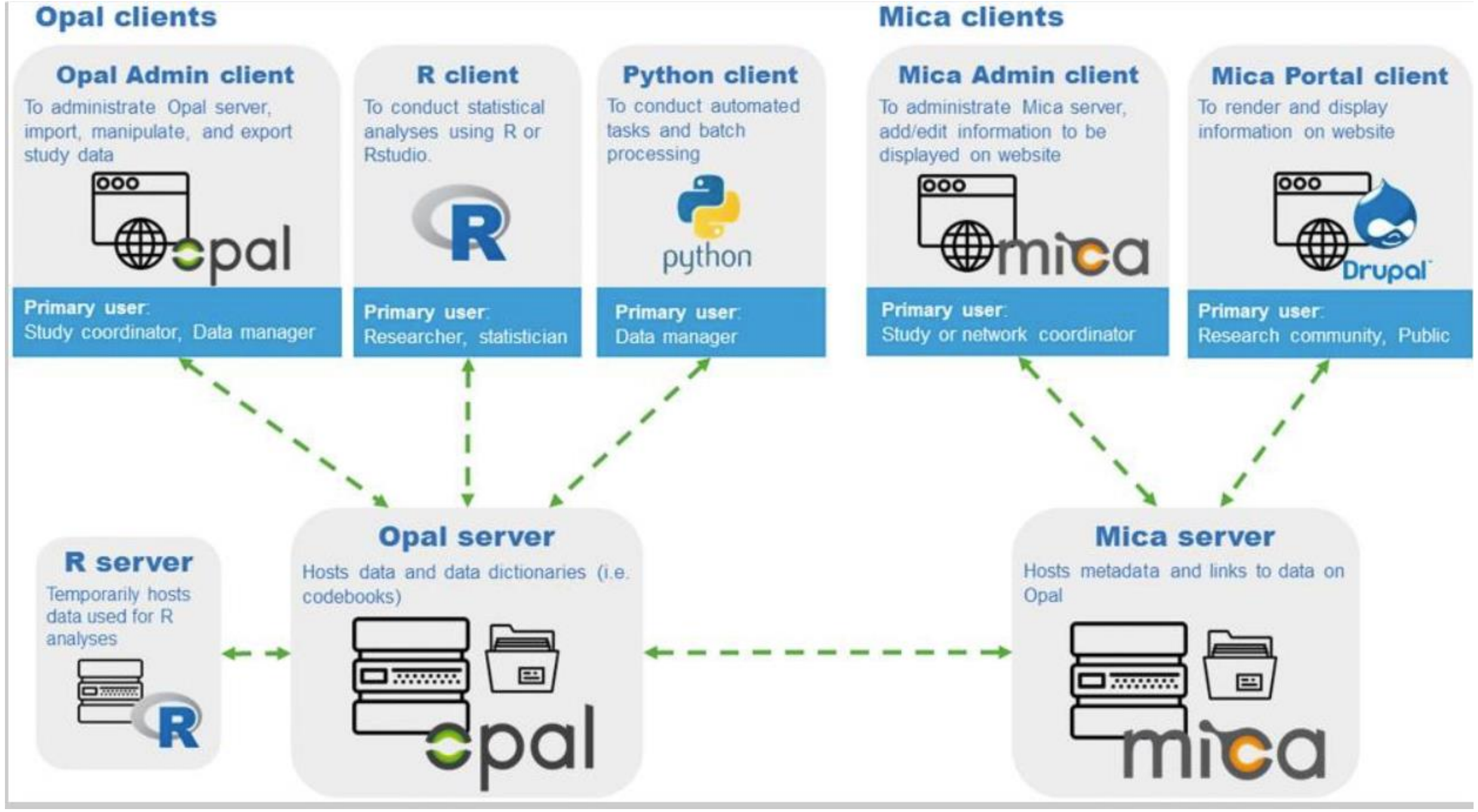
Process data: Access to data



Iterative Harmonization Steps

opal Data repository application integrating and storing data from multiple sources

RStudio Open-source software for data analysis



Opal clients

Opal Admin client
To administrate Opal server, import, manipulate, and export study data
Primary user: Study coordinator, Data manager

R client
To conduct statistical analyses using R or Rstudio.
Primary user: Researcher, statistician

Python client
To conduct automated tasks and batch processing
Primary user: Data manager

Mica clients

Mica Admin client
To administrate Mica server, add/edit information to be displayed on website
Primary user: Study or network coordinator

Mica Portal client
To render and display information on website
Primary user: Research community, Public

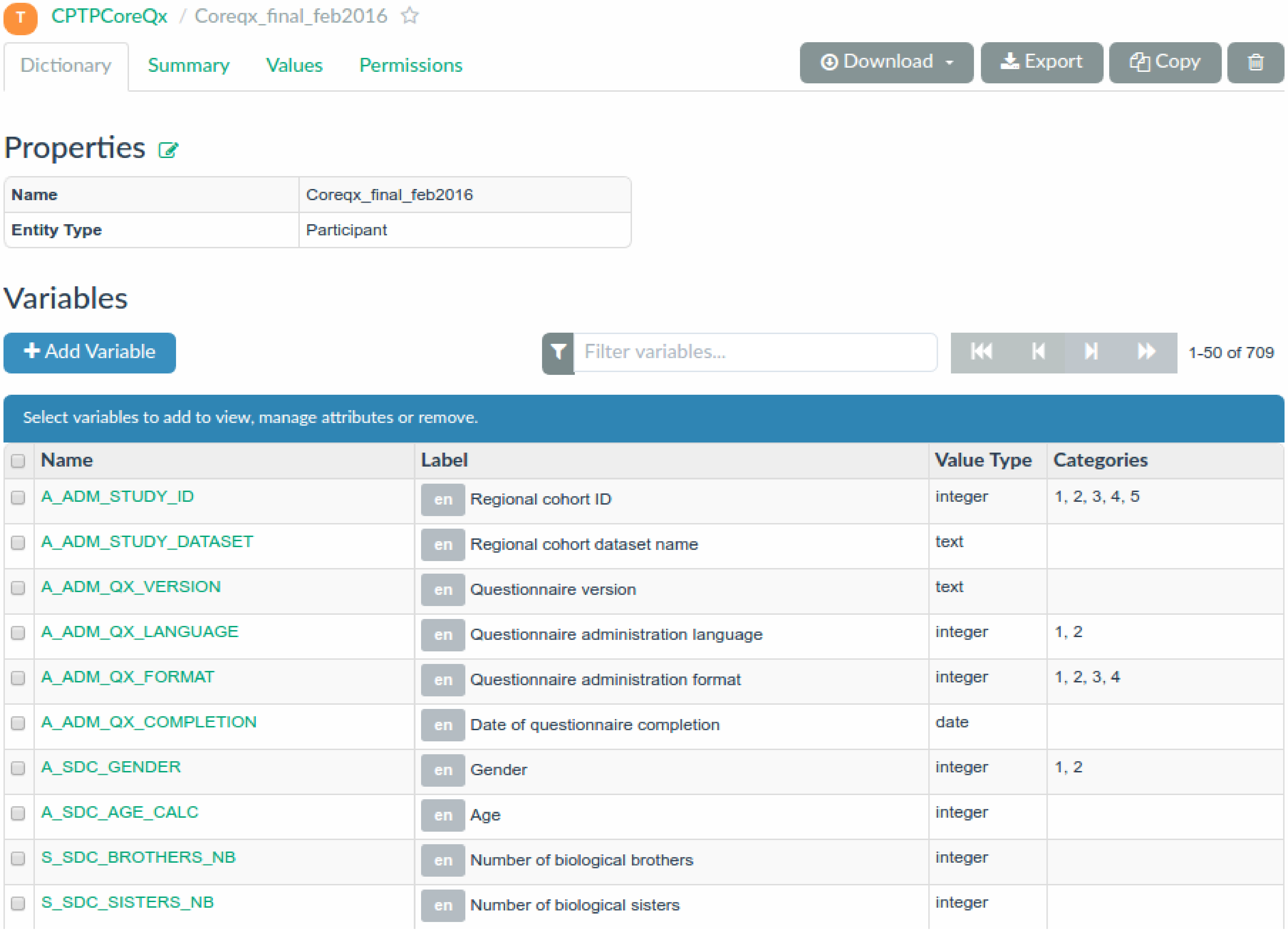
R server
Temporarily hosts data used for R analyses

Opal server
Hosts data and data dictionaries (i.e. codebooks)

Mica server
Hosts metadata and links to data on Opal

opal

Table



Properties

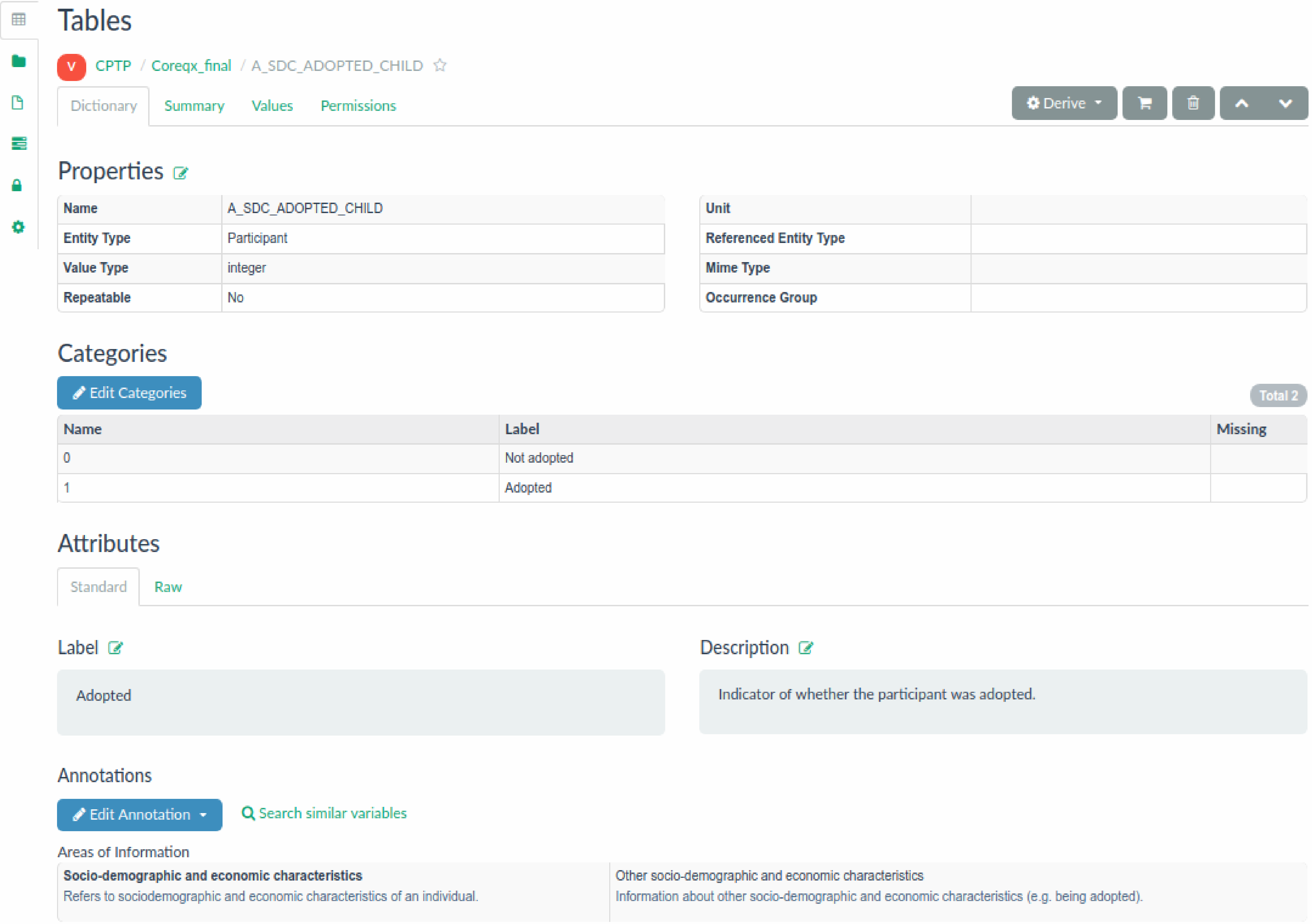
Name	Coreqx_final_feb2016
Entity Type	Participant

Variables

Select variables to add to view, manage attributes or remove.

□	Name	Label	Value Type	Categories
□	A_ADM_STUDY_ID	en Regional cohort ID	integer	1, 2, 3, 4, 5
□	A_ADM_STUDY_DATASET	en Regional cohort dataset name	text	
□	A_ADM_QX_VERSION	en Questionnaire version	text	
□	A_ADM_QX_LANGUAGE	en Questionnaire administration language	integer	1, 2
□	A_ADM_QX_FORMAT	en Questionnaire administration format	integer	1, 2, 3, 4
□	A_ADM_QX_COMPLETION	en Date of questionnaire completion	date	
□	A_SDC_GENDER	en Gender	integer	1, 2
□	A_SDC_AGE_CALC	en Age	integer	
□	S_SDC_BROTHERS_NB	en Number of biological brothers	integer	
□	S_SDC_SISTERS_NB	en Number of biological sisters	integer	

Variable



Tables

V CPTP / Coreqx_final / A_SDC_ADOPTED_CHILD ☆

Properties

Name	A_SDC_ADOPTED_CHILD
Entity Type	Participant
Value Type	integer
Repeatable	No

Unit	
Referenced Entity Type	
Mime Type	
Occurrence Group	

Categories

Name	Label	Missing
0	Not adopted	
1	Adopted	

Attributes

Label

Adopted

Description

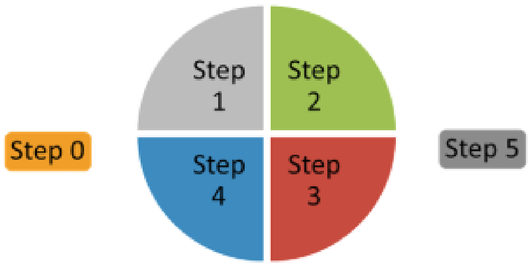
Indicator of whether the participant was adopted.

Annotations

Areas of Information

Socio-demographic and economic characteristics Refers to sociodemographic and economic characteristics of an individual.	Other socio-demographic and economic characteristics Information about other socio-demographic and economic characteristics (e.g. being adopted).

Assess study-specific source data



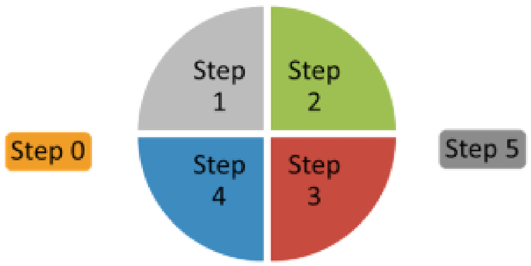
Iterative Harmonization Steps

- Verify:
  - Data format and compatibility with Opal
  - Entity IDs, duplicate IDs, IDs missing values
  - Inclusion criteria
  - Variable list, metadata, format
  - Univariate checks
  - Multivariate checks for cross-variable coherence
  - Document issues, summary reports, communication with cohorts
  - …

opal

R Studio

Generate core variables



Iterative Harmonization Steps

Study specific variables

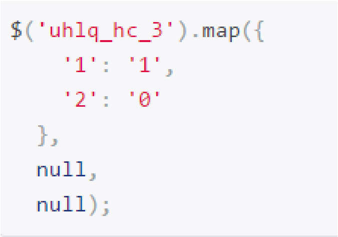
Case 1: Ever had sigmoidoscopy or colonoscopy

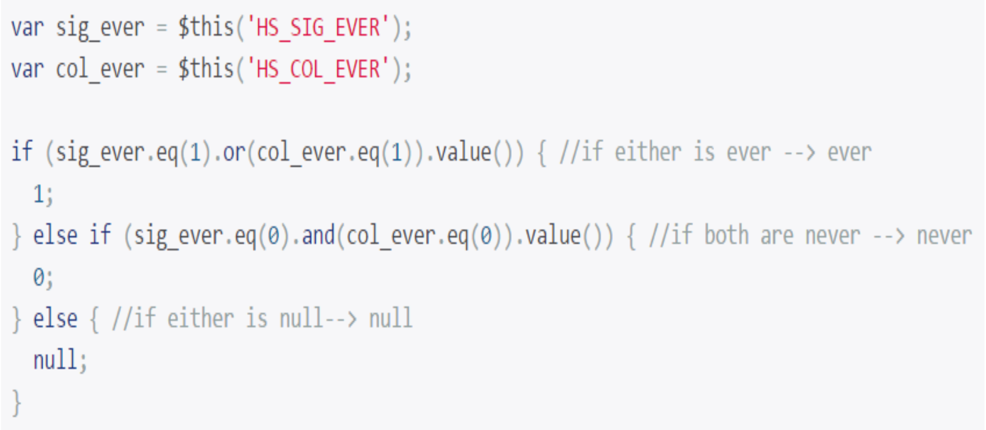
Case 2: Ever had sigmoidoscopy
Ever had colonoscopy

DataSchema variable

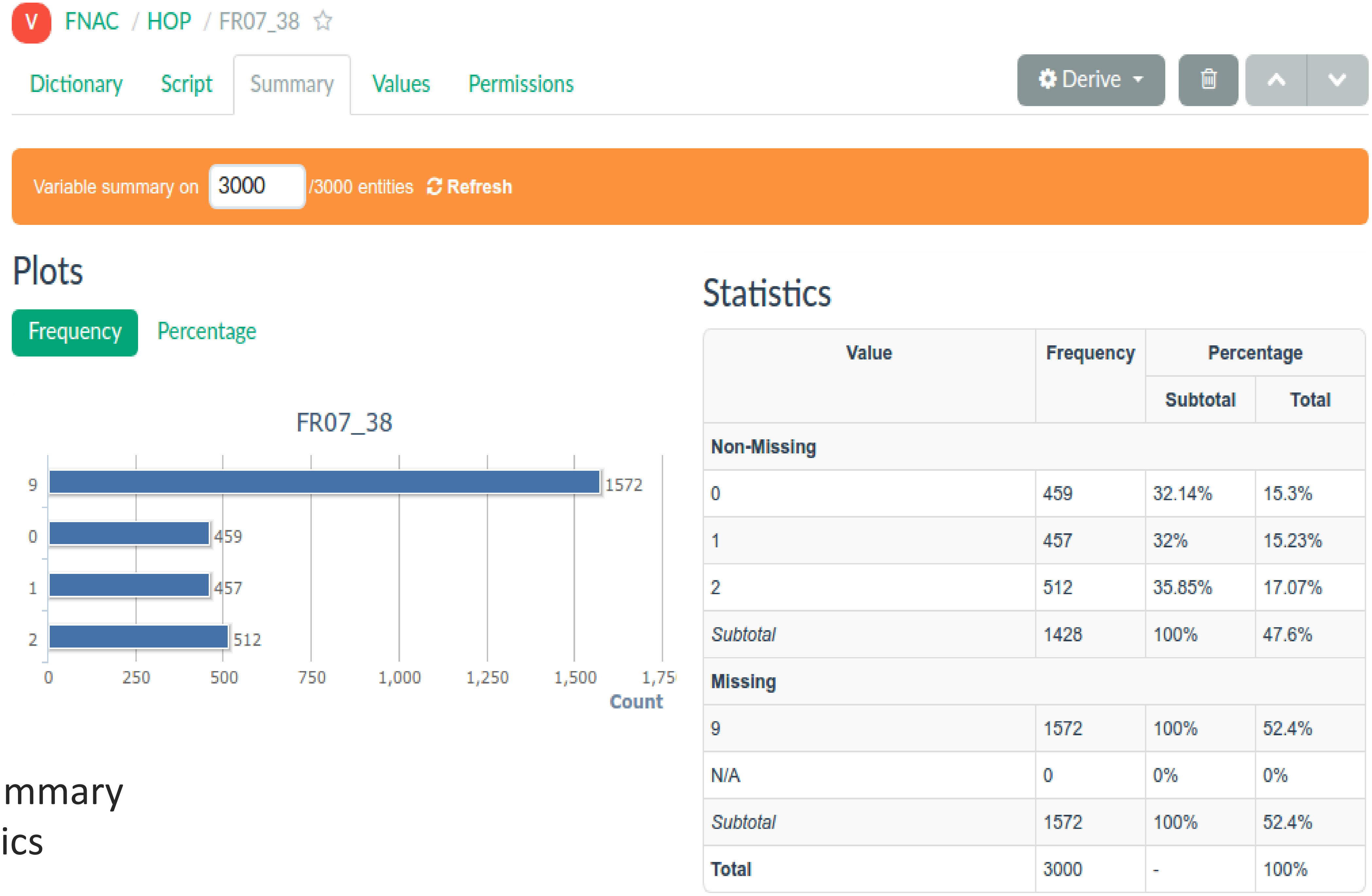
Ever had sigmoidoscopy or colonoscopy

Case	Rule	Script

1	Direct mapping from source variable	
		`$('uhlq_hc_3').map({`
		`    '1': '1',`
		`    '2': '0'`
		`  },`
		`  null,`
		`  null);`

2	Sourced from DataSchema variables
	If HS_SIG_EVER = 1 OR HS_COL_EVER = 1 --> code to 1
	If HS_SIG_EVER = 0 AND HS_COL_EVER = 0 --> code to 0
		
		`var sig_ever = $this('HS_SIG_EVER');`
		`var col_ever = $this('HS_COL_EVER');`
		
		`if (sig_ever.eq(1).or(col_ever.eq(1)).value()) { //if either is ever --> ever`
		`  1;`
		`} else if (sig_ever.eq(0).and(col_ever.eq(0)).value()) { //if both are never --> never`
		`  0;`
		`} else { //if either is null--> null`
		`  null;`
		`}`

opal



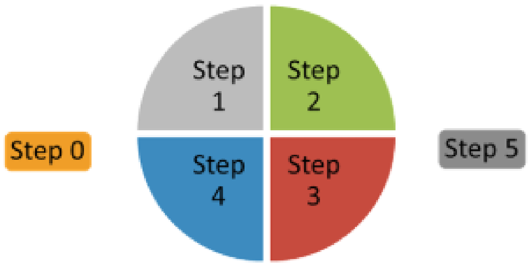
Variable summary on 3000 /3000 entities Refresh

Statistics

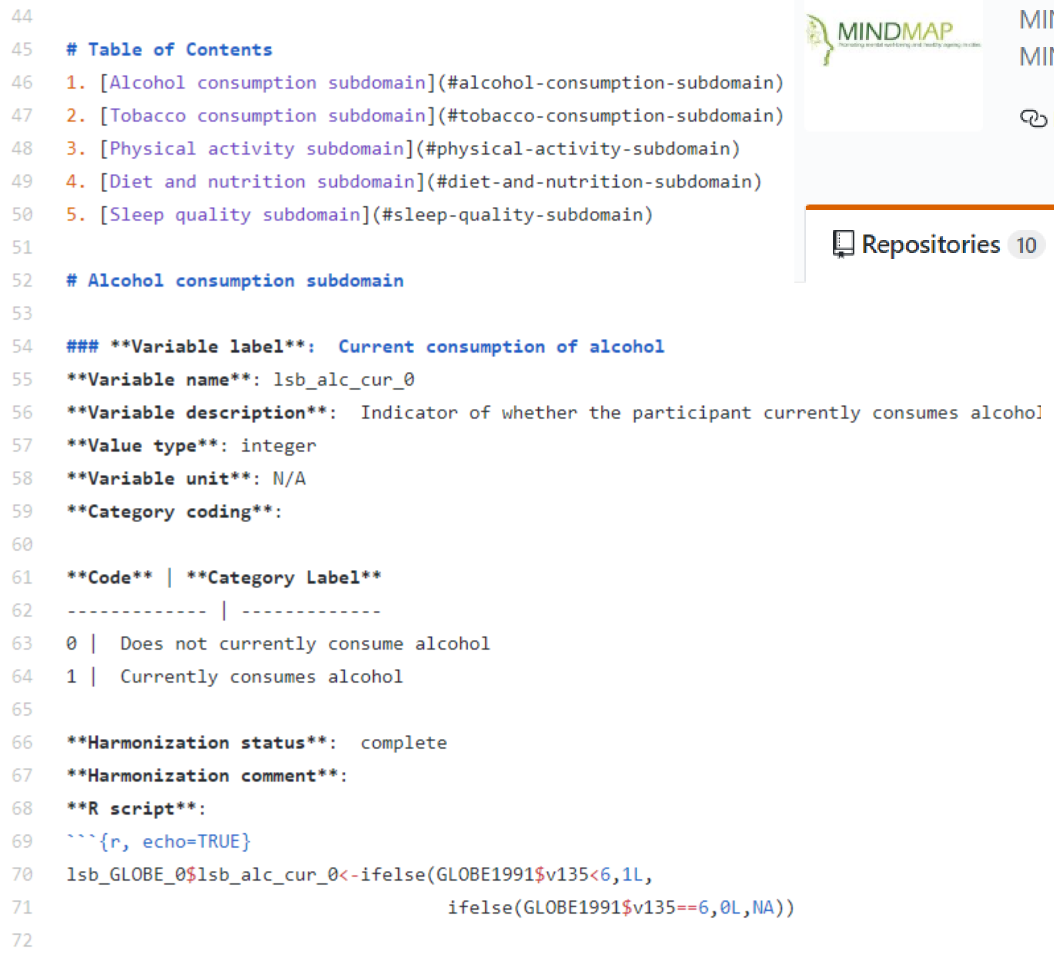
Value	Frequency	Percentage	
		Subtotal	Total
Non-Missing			
0	459	32.14%	15.3%
1	457	32%	15.23%
2	512	35.85%	17.07%
Subtotal	1428	100%	47.6%
Missing			
9	1572	100%	52.4%
N/A	0	0%	0%
Subtotal	1572	100%	52.4%
Total	3000	-	100%

Variable summary statistics

Generate core variables



Iterative Harmonization Steps



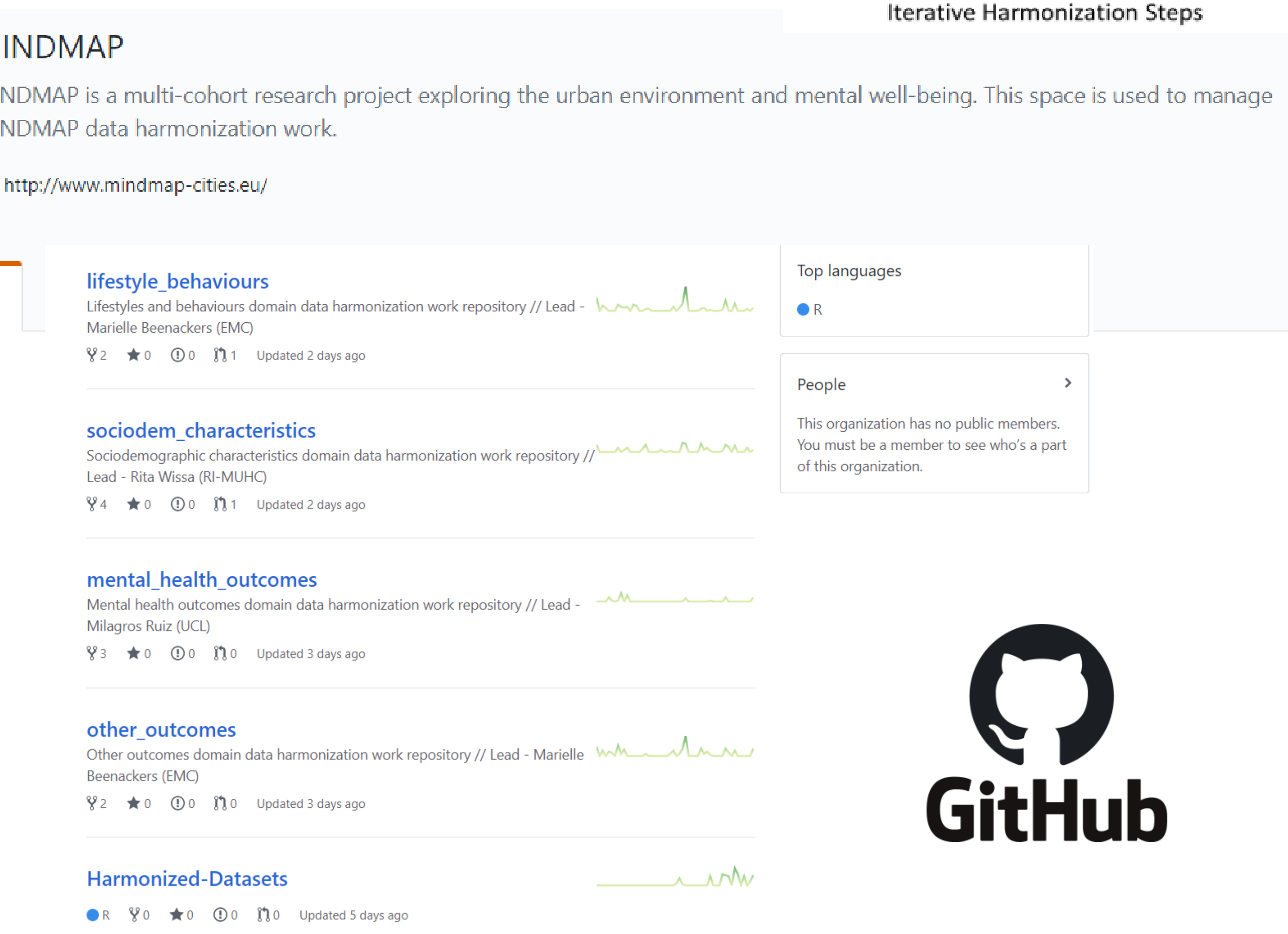
~~~
# Table of Contents
1. [Alcohol consumption subdomain](#alcohol-consumption-subdomain)
2. [Tobacco consumption subdomain](#tobacco-consumption-subdomain)
3. [Physical activity subdomain](#physical-activity-subdomain)
4. [Diet and nutrition subdomain](#diet-and-nutrition-subdomain)
5. [Sleep quality subdomain](#sleep-quality-subdomain)

# Alcohol consumption subdomain

### **Variable label**:  Current consumption of alcohol
**Variable name**: lsb_alc_cur_0
**Variable description**:  Indicator of whether the participant currently consumes alcohol
**Value type**: integer
**Variable unit**: N/A
**Category coding**:

**Code** | **Category Label**
------------- | -------------
0 |  Does not currently consume alcohol
1 |  Currently consumes alcohol

**Harmonization status**:  complete
**Harmonization comment**:
**R script**:
```{r, echo=TRUE}
lsb_GLOBE_0$lsb_alc_cur_0<-ifelse(GLOBE1991$v135<6,1L,
                                   ifelse(GLOBE1991$v135==6,0L,NA))
~~~



MINDMAP

MINDMAP is a multi-cohort research project exploring the urban environment and mental well-being. This space is used to manage MINDMAP data harmonization work.

http://www.mindmap-cities.eu/

Repositories 10

lifestyle_behaviours
Lifestyles and behaviours domain data harmonization work repository // Lead - Marielle Beenackers (EMC)
⑂ 2 ★ 0 ⓘ 0 ⇅ 1 Updated 2 days ago

sociodem_characteristics
Sociodemographic characteristics domain data harmonization work repository // Lead - Rita Wissa (RI-MUHC)
⑂ 4 ★ 0 ⓘ 0 ⇅ 1 Updated 2 days ago

mental_health_outcomes
Mental health outcomes domain data harmonization work repository // Lead - Milagros Ruiz (UCL)
⑂ 3 ★ 0 ⓘ 0 ⇅ 0 Updated 3 days ago

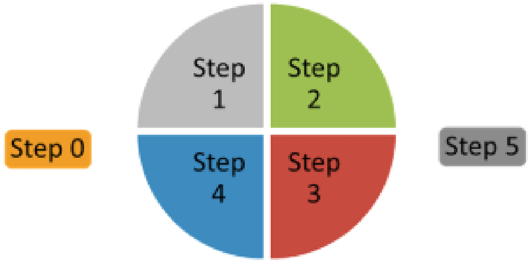
other_outcomes
Other outcomes domain data harmonization work repository // Lead - Marielle Beenackers (EMC)
⑂ 2 ★ 0 ⓘ 0 ⇅ 0 Updated 3 days ago

Harmonized-Datasets
● R ⑂ 0 ★ 0 ⓘ 0 ⇅ 0 Updated 5 days ago

Top languages
● R

People ›
This organization has no public members. You must be a member to see who's a part of this organization.

Estimate quality of harmonized dataset



Iterative Harmonization Steps

- For each study-specific harmonized dataset:
  - Validate harmonization process (algorithms, scripts)
  - Validate data content and consistency
  - Distributions and missing values
  - Consistency with DataSchema (format, categories)
  - Harmonization completion statuses
  - Multivariate checks for cross-variable coherence
  - Document issues, summary reports, communication with cohorts
  - …

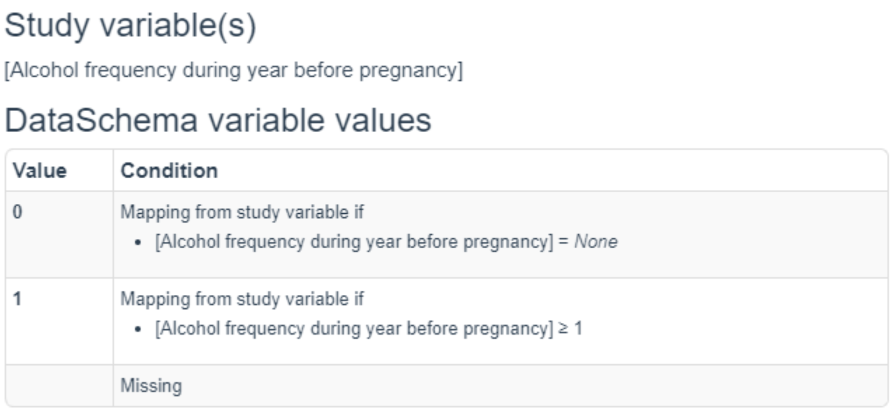
opal

RStudio®

Alcohol consumption of mother 1 year prior to pregnancy : Y/N ?

Study A

Category	Freq.
Every day	43
4-6 / week	185
2-3 / week	503
1/ week	380
2-3 /month	278
1 / month	185
< 1 /month	325
Never	461
Missing	5
Total	2 365



Study variable(s)

[Alcohol frequency during year before pregnancy]

DataSchema variable values

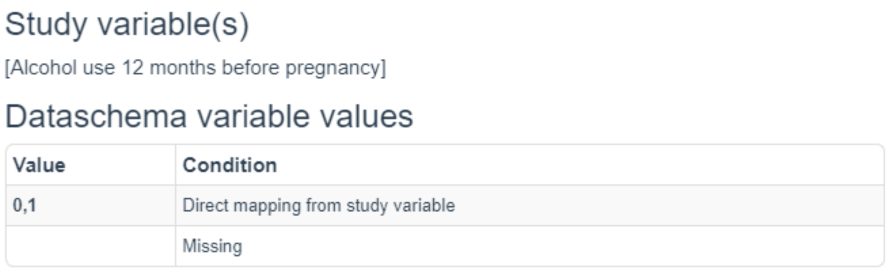
Value	Condition
0	Mapping from study variable if • [Alcohol frequency during year before pregnancy] = *None*
1	Mapping from study variable if • [Alcohol frequency during year before pregnancy] ≥ 1
	Missing

Study-specific harmonized variable

Category	Freq.
Yes	1 899
No	461
Missing	5
Total	2 365

Study B

Category	Freq.
Yes	2 728
No	590
Missing	23
Total	3 341



Study variable(s)

[Alcohol use 12 months before pregnancy]

Dataschema variable values

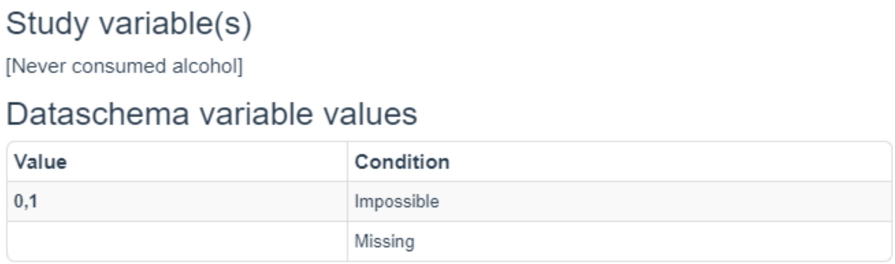
Value	Condition
0,1	Direct mapping from study variable
	Missing

Study-specific harmonized variable

Category	Freq.
Yes	2 728
No	590
Missing	23
Total	3 341

Study C

Category	Freq.
Yes	261
No	1835
Missing	90
Total	2 187



Study variable(s)

[Never consumed alcohol]

Dataschema variable values

Value	Condition
0,1	Impossible
	Missing

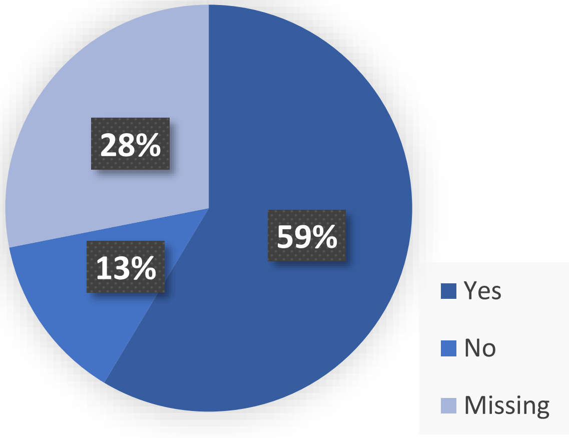
Study-specific harmonized variable

Category	Freq.
Yes	0
No	0
Missing	2 187
Total	2 187

ReACH

Research Advancement through Cohort Cataloguing and Harmonization

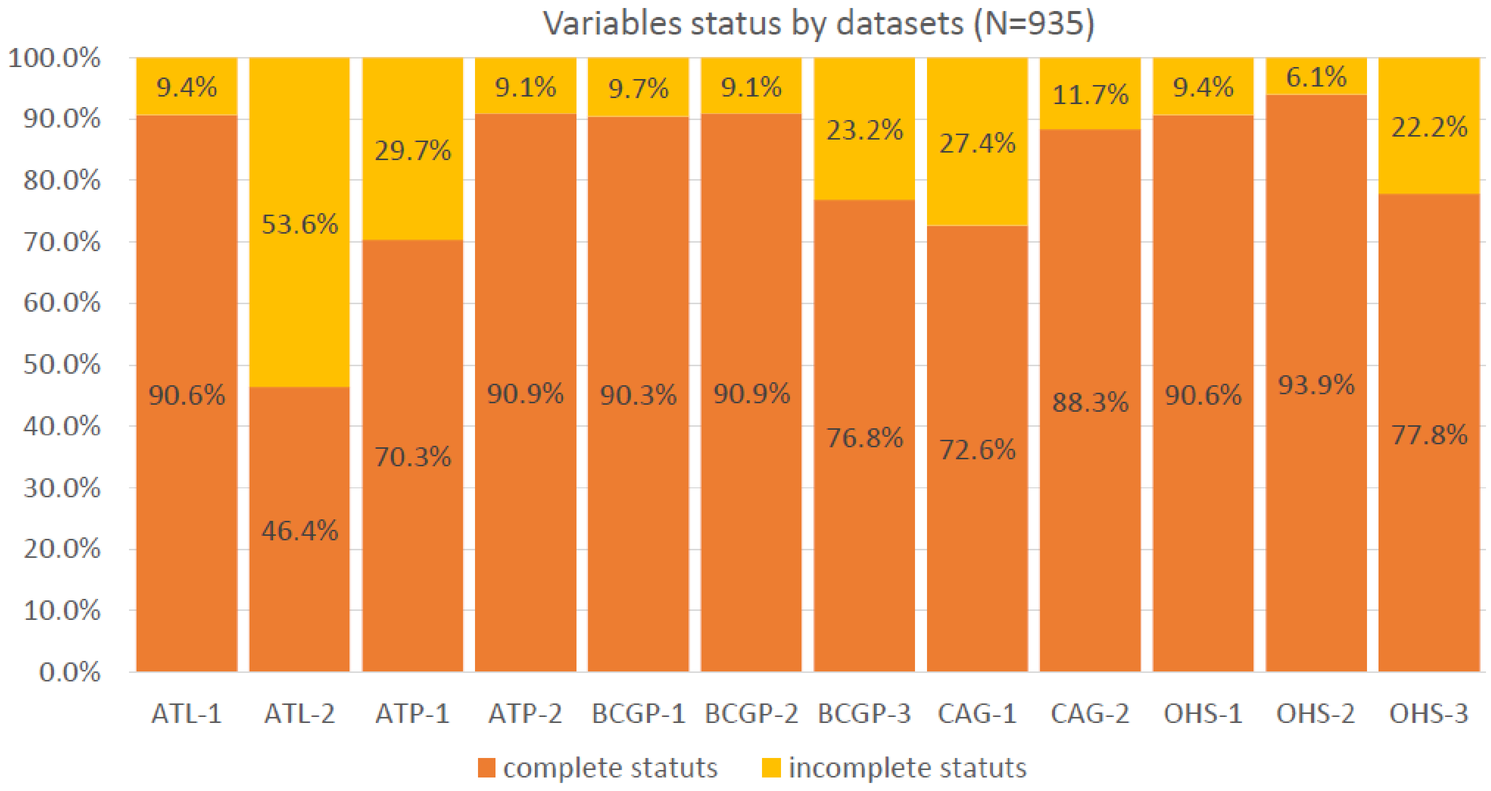
Harmonized variable



Harmonization potential across studies: CPTP Health and Risk Factor Questionnaire

Canadian Partnership for Tomorrow Project

Health and Risk Factor Questionnaire – Harmonized variables

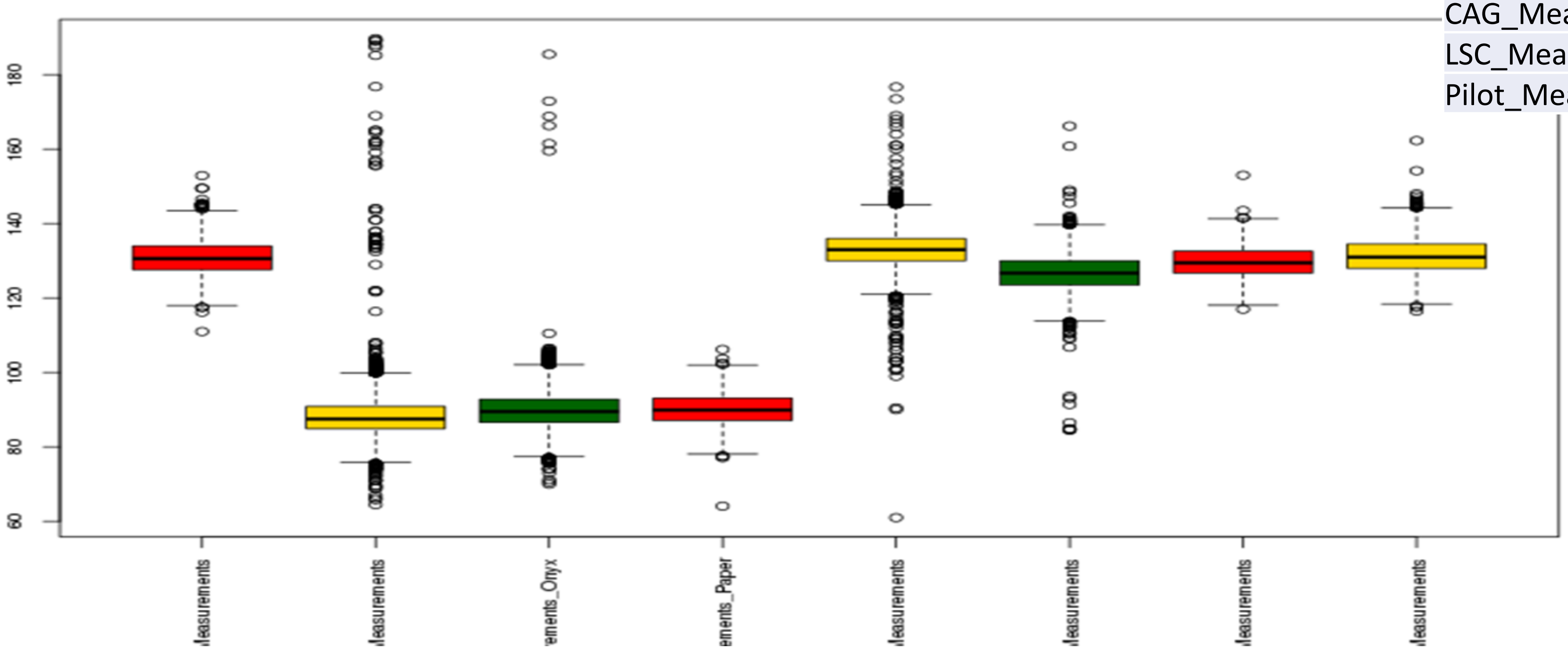


Estimate quality of harmonized dataset

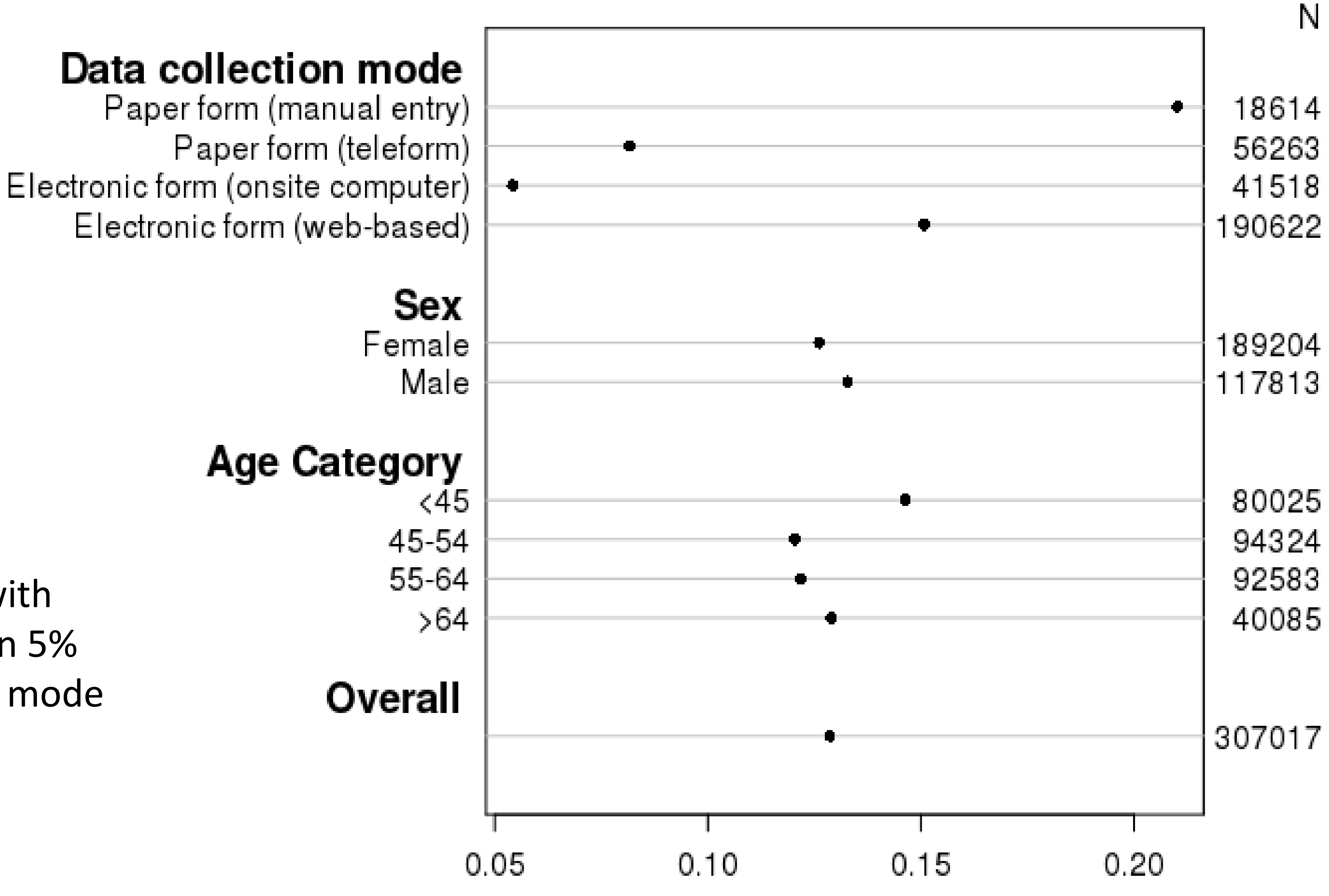
Understand the potential and limitations of the harmonized dataset

	N	Mean
ATL_Measurements	4872	130,88
ATL_Measurements	22703	88,01
ATP_Measurements	29347	89,8
ATP_Measurements	1149	90,2
BCGP_Measurements	16363	133,21
CAG_Measurements	19992	126,83
LSC_Measurements	649	129,78
Pilot_Measurements	7970	131,3

Sitting height



Missing values: CPTP data collection mode



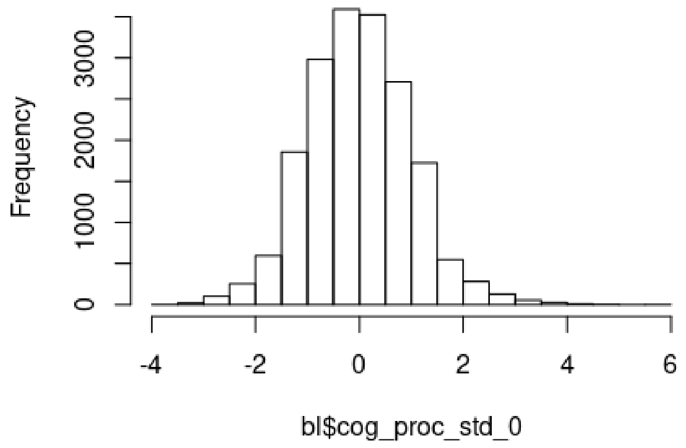
Canadian Partnership for Tomorrow Project

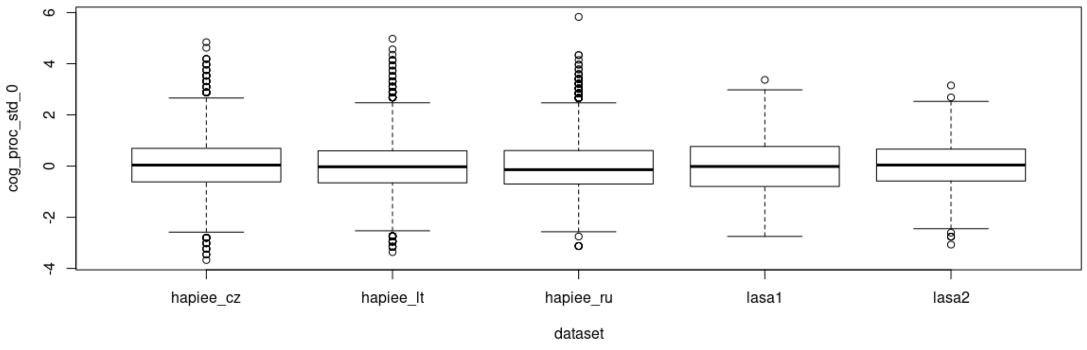
Proportion of participants with missing values for more than 5% according to data collection mode and basic demographic characteristics

Proportion missing values

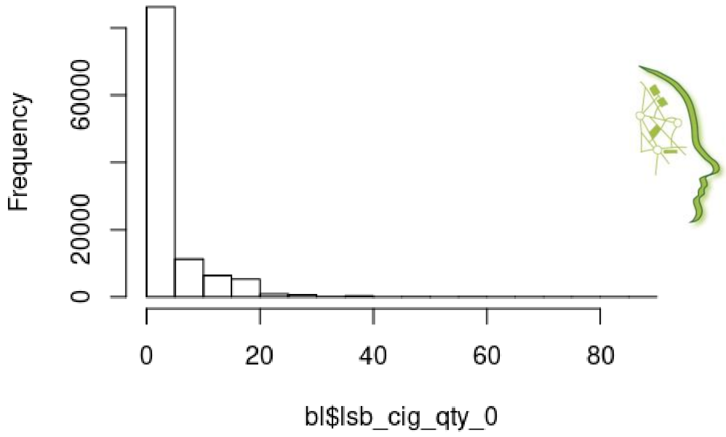
Harmonization process: MINDMAP variable profiles

Cognitive processing (*z*-score standardized)

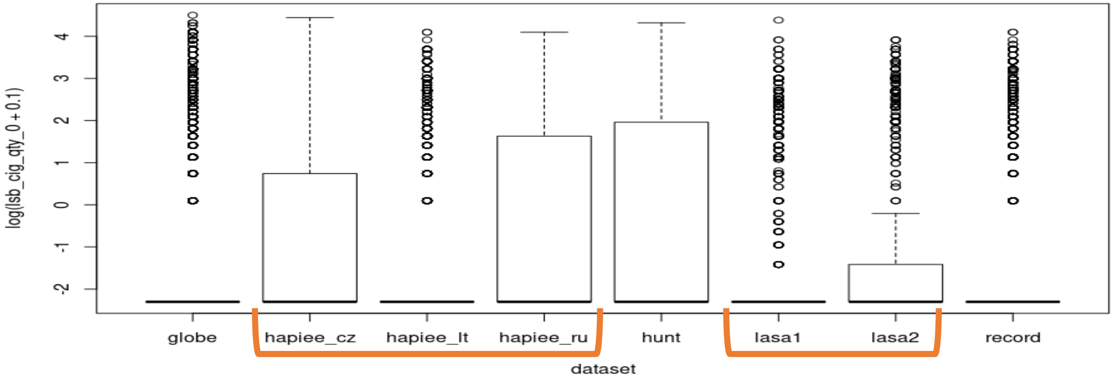




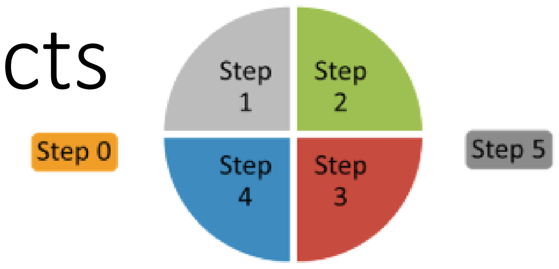
Average # cigarettes smoked



MINDMAP
Promoting mental well-being and healthy ageing in cities

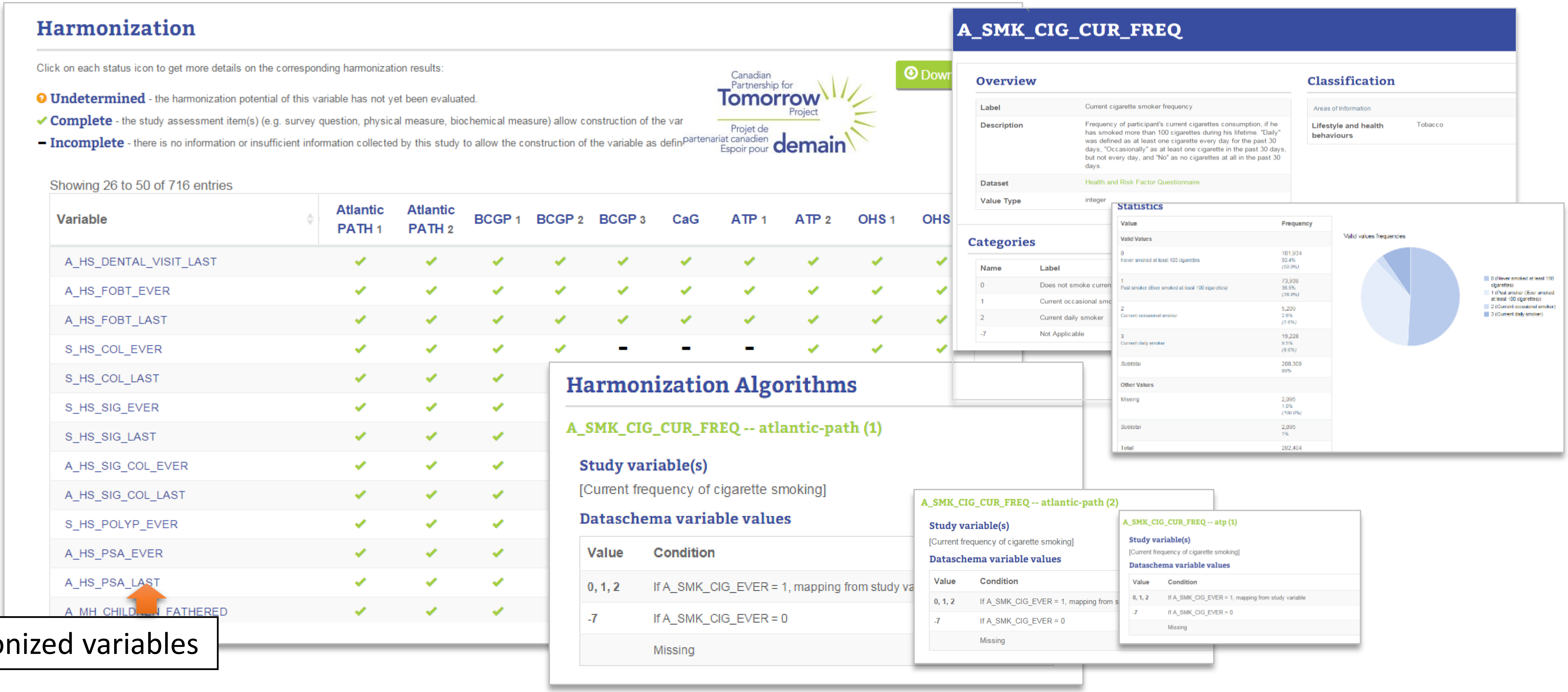


Disseminate and preserve harmonization products



Iterative Harmonization Steps

Ensure transparency and leverage usage of harmonized data



Harmonization

Click on each status icon to get more details on the corresponding harmonization results:

Undetermined - the harmonization potential of this variable has not yet been evaluated.

✓ **Complete** - the study assessment item(s) (e.g. survey question, physical measure, biochemical measure) allow construction of the var

– **Incomplete** - there is no information or insufficient information collected by this study to allow the construction of the variable as defin

Canadian Partnership for Tomorrow Project
Projet de partenariat canadien Espoir pour demain
Down

Showing 26 to 50 of 716 entries

Variable	Atlantic PATH 1	Atlantic PATH 2	BCGP 1	BCGP 2	BCGP 3	CaG	ATP 1	ATP 2	OHS 1	OHS
A_HS_DENTAL_VISIT_LAST	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A_HS_FOBT_EVER	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A_HS_FOBT_LAST	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
S_HS_COL_EVER	✓	✓	✓	✓	–	–	–	✓	✓	✓
S_HS_COL_LAST	✓	✓	✓							
S_HS_SIG_EVER	✓	✓	✓							
S_HS_SIG_LAST	✓	✓	✓							
A_HS_SIG_COL_EVER	✓	✓	✓							
A_HS_SIG_COL_LAST	✓	✓	✓							
S_HS_POLYP_EVER	✓	✓	✓							
A_HS_PSA_EVER	✓	✓	✓							
A_HS_PSA_LAST	✓	✓	✓							
A_MH_CHILDREN_FATHERED	✓	✓	✓							

Harmonized variables

A_SMK_CIG_CUR_FREQ

Overview

Label	Current cigarette smoker frequency
Description	Frequency of participant's current cigarettes consumption, if he has smoked more than 100 cigarettes during his lifetime. "Daily" was defined as at least one cigarette every day for the past 30 days, "Occasionally" as at least one cigarette in the past 30 days, but not every day, and "No" as no cigarettes at all in the past 30 days.
Dataset	Health and Risk Factor Questionnaire
Value Type	integer

Classification

Areas of Information	
Lifestyle and health behaviours	Tobacco

Categories

Name	Label
0	Does not smoke curren
1	Current occasional smo
2	Current daily smoker
-7	Not Applicable

Statistics

Value	Frequency
Valid Values	
0 Never smoked at least 100 cigarettes	101,934 50.4% (50.9%)
1 Past smoker (Ever smoked at least 100 cigarettes)	73,938 36.5% (36.9%)
2 Current occasional smoker	5,209 2.6% (2.6%)
3 Current daily smoker	19,228 9.5% (9.6%)
Subtotal	200,309 99%
Other Values	
Missing	2,095 1.0% (100.0%)
Subtotal	2,095 1%
Total	202,404

Valid values frequencies
0 (Never smoked at least 100 cigarettes)
1 (Past smoker (Ever smoked at least 100 cigarettes))
2 (Current occasional smoker)
3 (Current daily smoker)

Harmonization Algorithms

A_SMK_CIG_CUR_FREQ -- atlantic-path (1)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, mapping from study va
-7	If A_SMK_CIG_EVER = 0
	Missing

A_SMK_CIG_CUR_FREQ -- atlantic-path (2)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, mapping from s
-7	If A_SMK_CIG_EVER = 0
	Missing

A_SMK_CIG_CUR_FREQ -- atp (1)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

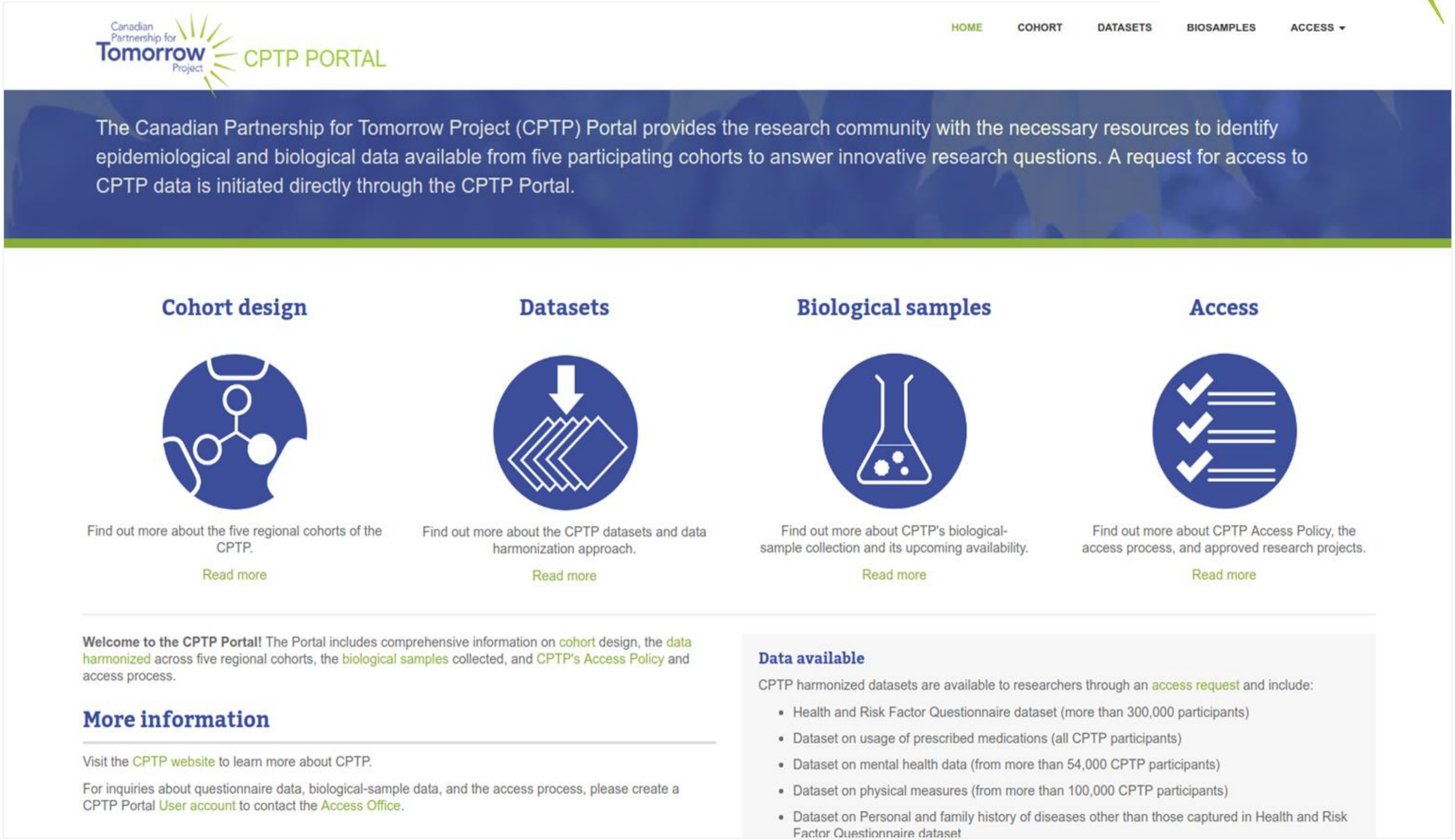
Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, mapping from study variable
-7	If A_SMK_CIG_EVER = 0
	Missing

mica

A data portal application used to describe central data and manage data access requests

Canadian Partnership for Tomorrow Project

Data portal



Canadian Partnership for Tomorrow Project CPTP PORTAL

The Canadian Partnership for Tomorrow Project (CPTP) Portal provides the research community with the necessary resources to identify epidemiological and biological data available from five participating cohorts to answer innovative research questions. A request for access to CPTP data is initiated directly through the CPTP Portal.

Cohort design

Find out more about the five regional cohorts of the CPTP.

Read more

Datasets

Find out more about the CPTP datasets and data harmonization approach.

Read more

Biological samples

Find out more about CPTP's biological-sample collection and its upcoming availability.

Read more

Access

Find out more about CPTP Access Policy, the access process, and approved research projects.

Read more

Welcome to the CPTP Portal! The Portal includes comprehensive information on cohort design, the data harmonized across five regional cohorts, the biological samples collected, and CPTP's Access Policy and access process.

More information

Visit the CPTP website to learn more about CPTP.

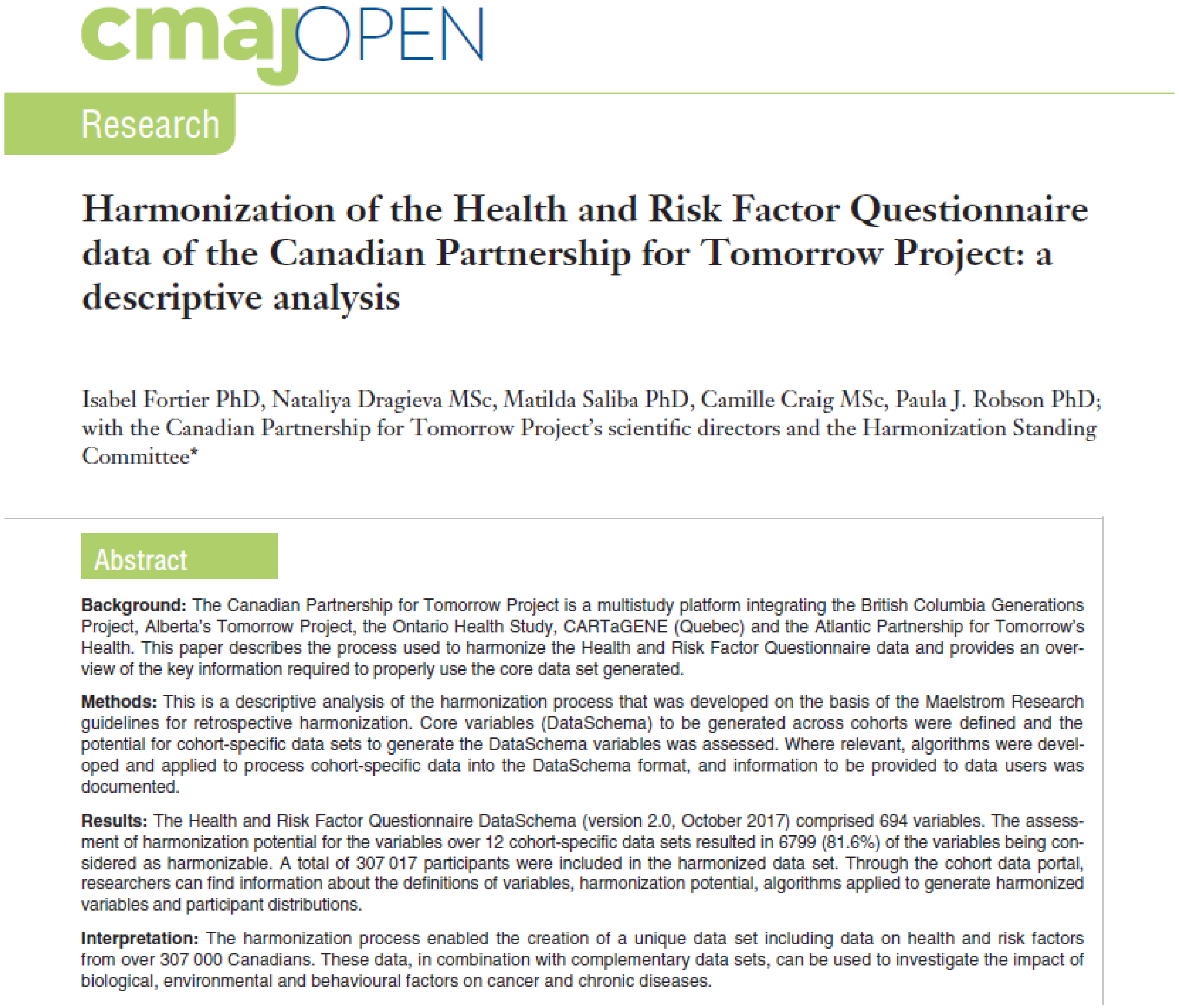
For inquiries about questionnaire data, biological-sample data, and the access process, please create a CPTP Portal User account to contact the Access Office.

Data available

CPTP harmonized datasets are available to researchers through an access request and include:

- Health and Risk Factor Questionnaire dataset (more than 300,000 participants)
- Dataset on usage of prescribed medications (all CPTP participants)
- Dataset on mental health data (from more than 54,000 CPTP participants)
- Dataset on physical measures (from more than 100,000 CPTP participants)
- Dataset on Personal and family history of diseases other than those captured in Health and Risk Factor Questionnaire dataset

Document the overall harmonization process



cmajOPEN

Research

Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis

Isabel Fortier PhD, Nataliya Dragieva MSc, Matilda Saliba PhD, Camille Craig MSc, Paula J. Robson PhD; with the Canadian Partnership for Tomorrow Project's scientific directors and the Harmonization Standing Committee*

Abstract

Background: The Canadian Partnership for Tomorrow Project is a multistudy platform integrating the British Columbia Generations Project, Alberta's Tomorrow Project, the Ontario Health Study, CARTaGENE (Quebec) and the Atlantic Partnership for Tomorrow's Health. This paper describes the process used to harmonize the Health and Risk Factor Questionnaire data and provides an overview of the key information required to properly use the core data set generated.

Methods: This is a descriptive analysis of the harmonization process that was developed on the basis of the Maelstrom Research guidelines for retrospective harmonization. Core variables (DataSchema) to be generated across cohorts were defined and the potential for cohort-specific data sets to generate the DataSchema variables was assessed. Where relevant, algorithms were developed and applied to process cohort-specific data into the DataSchema format, and information to be provided to data users was documented.

Results: The Health and Risk Factor Questionnaire DataSchema (version 2.0, October 2017) comprised 694 variables. The assessment of harmonization potential for the variables over 12 cohort-specific data sets resulted in 6799 (81.6%) of the variables being considered as harmonizable. A total of 307 017 participants were included in the harmonized data set. Through the cohort data portal, researchers can find information about the definitions of variables, harmonization potential, algorithms applied to generate harmonized variables and participant distributions.

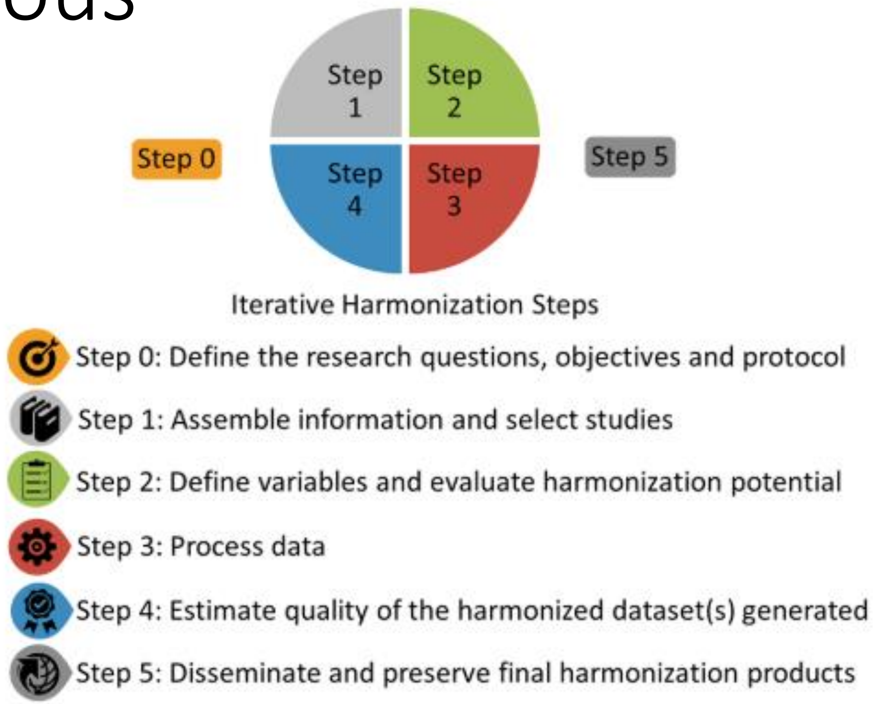
Interpretation: The harmonization process enabled the creation of a unique data set including data on health and risk factors from over 307 000 Canadians. These data, in combination with complementary data sets, can be used to investigate the impact of biological, environmental and behavioural factors on cancer and chronic diseases.

Maelstrom Research process for rigorous data harmonization

Retrospective data harmonization offers many benefits but is necessarily challenging.

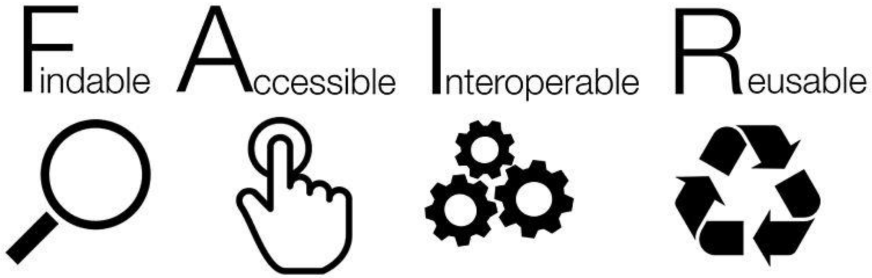
Need a general systematic process that can be adapted to each initiative.

Applying systematic approach to ensure proper quality checks and documentation throughout is critical for assessing and interpreting results.



Iterative Harmonization Steps

- Step 0: Define the research questions, objectives and protocol
- Step 1: Assemble information and select studies
- Step 2: Define variables and evaluate harmonization potential
- Step 3: Process data
- Step 4: Estimate quality of the harmonized dataset(s) generated
- Step 5: Disseminate and preserve final harmonization products



THANK YOU!

Funding and support:

Centre universitaire de santé McGill
Institut de recherche

McGill University Health Centre
Research Institute

MUHC
McGILL UNIVERSITY HEALTH CENTRE
FOUNDATION

Économie, Innovation et Exportations
Québec

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

CANADIAN PARTNERSHIP AGAINST CANCER
PARTENARIAT CANADIEN CONTRE LE CANCER

SEVENTH FRAMEWORK PROGRAMME

EUROPEAN COMMISSION

CIHR IRSC
Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

National Institute on Aging

canarie @25

Canadian Partnership for Tomorrow Project

ReACH
Research Advancement through Cohort Cataloguing and Harmonization

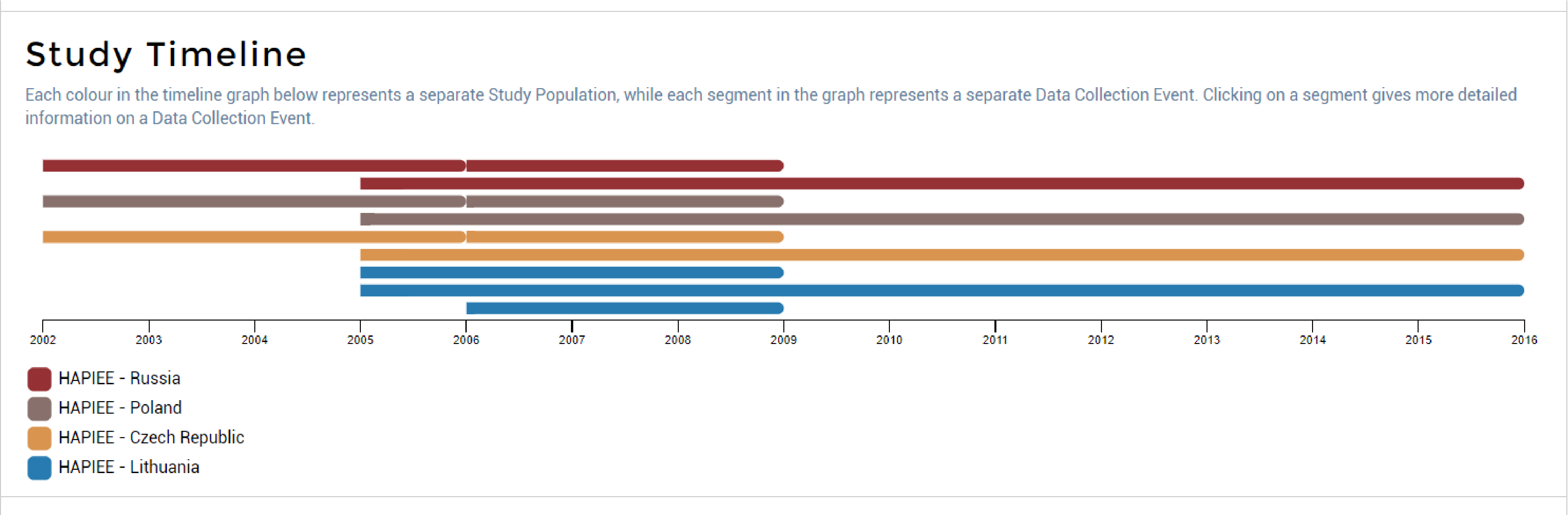
MINDMAP
Promoting mental well-being and healthy ageing in cities

www.maelstrom-research.org

Our numbers continue to grow

Networks	18
Individual Studies	204
Individual Studies with Variables	122
Individual Study Variables	933,144

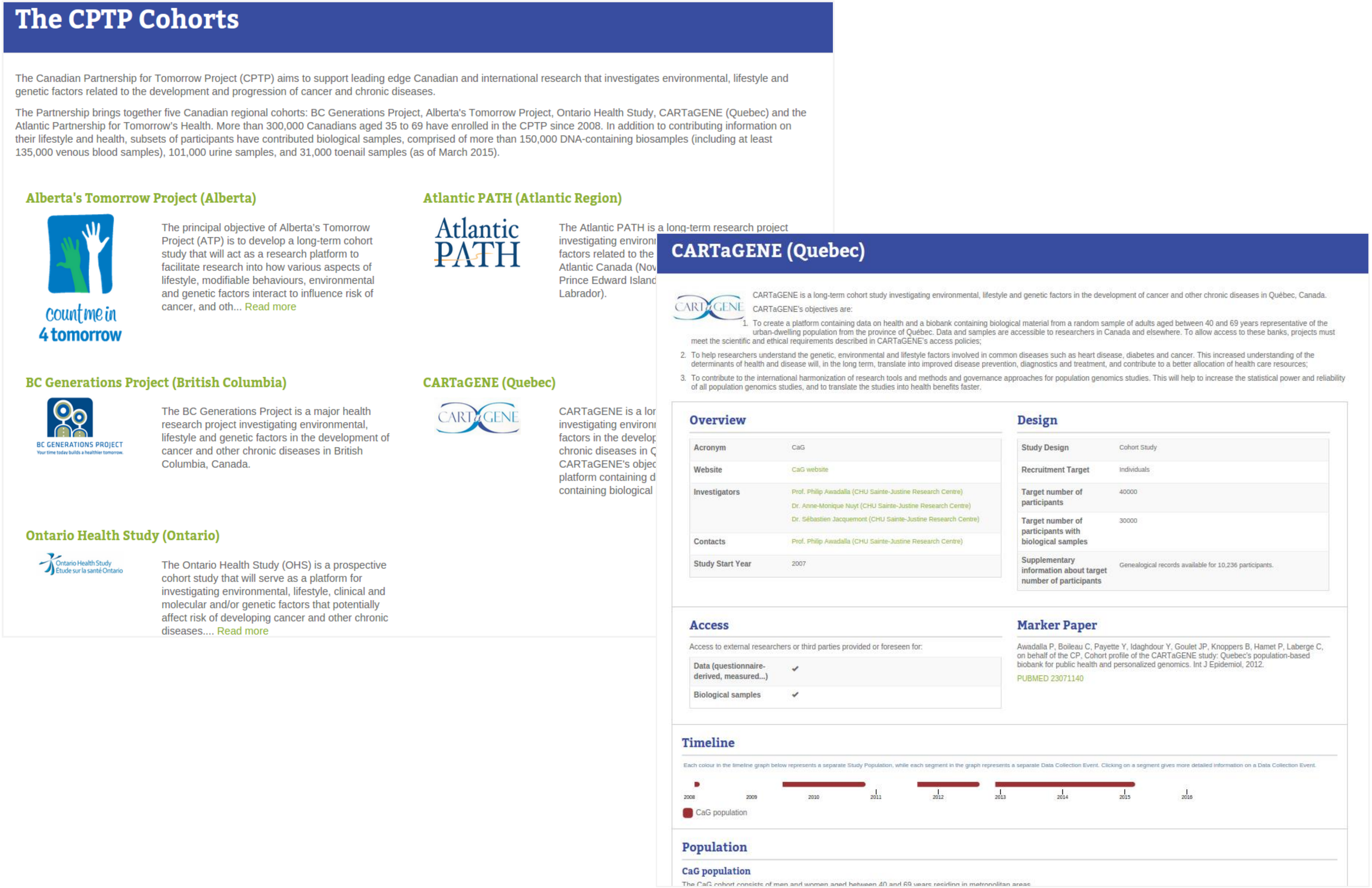
Population-based cohort studies



Study Timeline

Each colour in the timeline graph below represents a separate Study Population, while each segment in the graph represents a separate Data Collection Event. Clicking on a segment gives more detailed information on a Data Collection Event.

Study description



The CPTP Cohorts

The Canadian Partnership for Tomorrow Project (CPTP) aims to support leading edge Canadian and international research that investigates environmental, lifestyle and genetic factors related to the development and progression of cancer and chronic diseases.

The Partnership brings together five Canadian regional cohorts: BC Generations Project, Alberta's Tomorrow Project, Ontario Health Study, CARTaGENE (Quebec) and the Atlantic Partnership for Tomorrow's Health. More than 300,000 Canadians aged 35 to 69 have enrolled in the CPTP since 2008. In addition to contributing information on their lifestyle and health, subsets of participants have contributed biological samples, comprised of more than 150,000 DNA-containing biosamples (including at least 135,000 venous blood samples), 101,000 urine samples, and 31,000 toenail samples (as of March 2015).

Alberta's Tomorrow Project (Alberta)

The principal objective of Alberta's Tomorrow Project (ATP) is to develop a long-term cohort study that will act as a research platform to facilitate research into how various aspects of lifestyle, modifiable behaviours, environmental and genetic factors interact to influence risk of cancer, and oth... Read more

Atlantic PATH (Atlantic Region)

The Atlantic PATH is a long-term research project investigating environ
factors related to the
Atlantic Canada (Nov
Prince Edward Island
Labrador).

BC Generations Project (British Columbia)

The BC Generations Project is a major health research project investigating environmental, lifestyle and genetic factors in the development of cancer and other chronic diseases in British Columbia, Canada.

CARTaGENE (Quebec)

CARTaGENE is a lor
investigating environ
factors in the develop
chronic diseases in Q
CARTaGENE's objec
platform containing d
containing biological

Ontario Health Study (Ontario)

The Ontario Health Study (OHS) is a prospective cohort study that will serve as a platform for investigating environmental, lifestyle, clinical and molecular and/or genetic factors that potentially affect risk of developing cancer and other chronic diseases.... Read more

CARTaGENE (Quebec)

CARTaGENE is a long-term cohort study investigating environmental, lifestyle and genetic factors in the development of cancer and other chronic diseases in Québec, Canada.

CARTaGENE's objectives are:

1. To create a platform containing data on health and a biobank containing biological material from a random sample of adults aged between 40 and 69 years representative of the urban-dwelling population from the province of Québec. Data and samples are accessible to researchers in Canada and elsewhere. To allow access to these banks, projects must meet the scientific and ethical requirements described in CARTaGENE's access policies;
2. To help researchers understand the genetic, environmental and lifestyle factors involved in common diseases such as heart disease, diabetes and cancer. This increased understanding of the determinants of health and disease will, in the long term, translate into improved disease prevention, diagnostics and treatment, and contribute to a better allocation of health care resources;
3. To contribute to the international harmonization of research tools and methods and governance approaches for population genomics studies. This will help to increase the statistical power and reliability of all population genomics studies, and to translate the studies into health benefits faster.

Overview

Acronym	CaG
Website	CaG website
Investigators	Prof. Philip Awadalla (CHU Sainte-Justine Research Centre) Dr. Anne-Monique Nuyt (CHU Sainte-Justine Research Centre) Dr. Sébastien Jacquemont (CHU Sainte-Justine Research Centre)
Contacts	Prof. Philip Awadalla (CHU Sainte-Justine Research Centre)
Study Start Year	2007

Design

Study Design	Cohort Study
Recruitment Target	Individuals
Target number of participants	40000
Target number of participants with biological samples	30000
Supplementary information about target number of participants	Genealogical records available for 10,236 participants.

Access

Access to external researchers or third parties provided or foreseen for:

Data (questionnaire-derived, measured...)	✓
Biological samples	✓

Marker Paper

Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, Hamet P, Laberge C, on behalf of the CP, Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol, 2012.

PUBMED 23071140

Timeline

Each colour in the timeline graph below represents a separate Study Population, while each segment in the graph represents a separate Data Collection Event. Clicking on a segment gives more detailed information on a Data Collection Event.

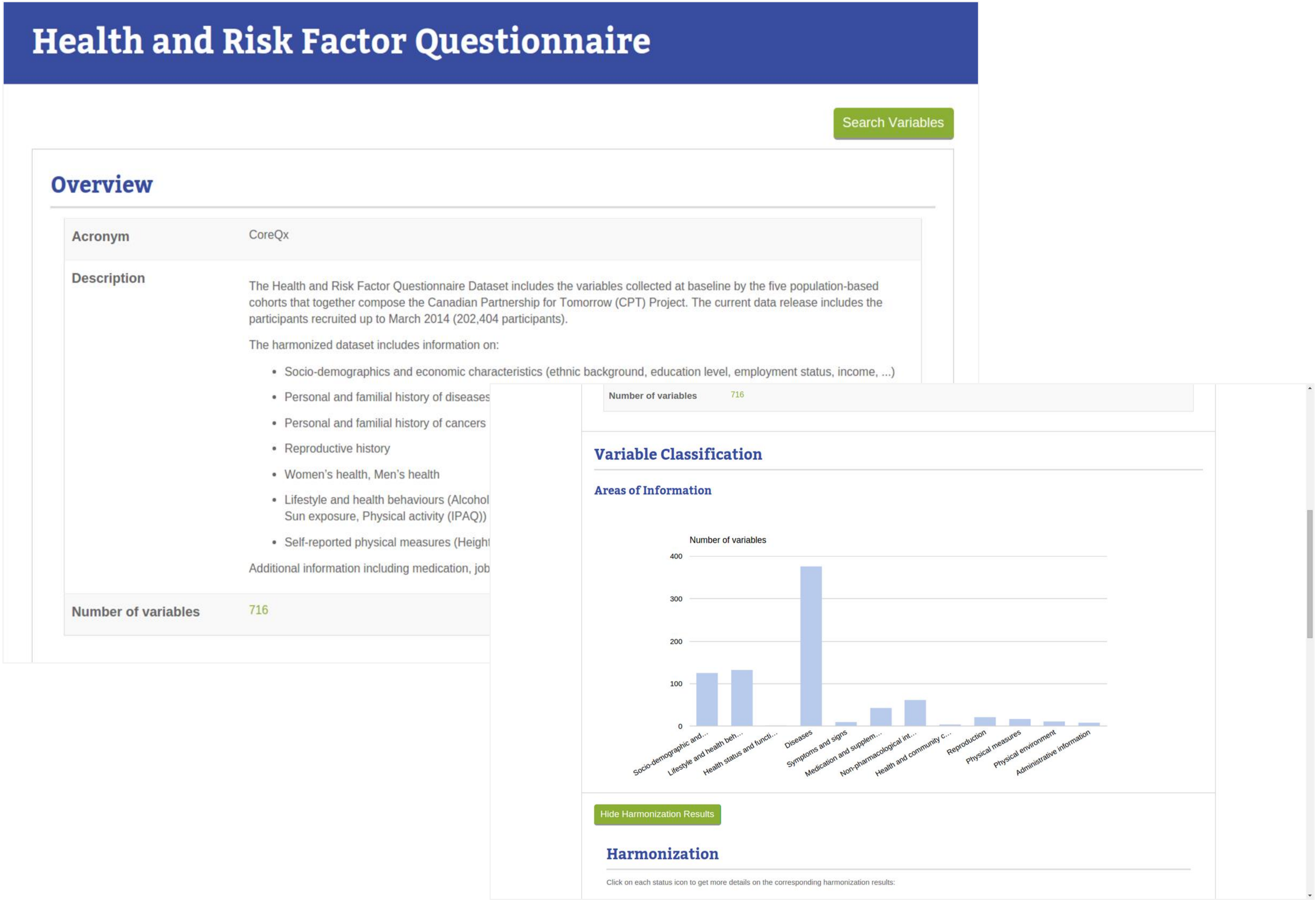
Population

CaG population

The CaG cohort consists of men and women aged between 40 and 69 years residing in metropolitan areas

mica

Dataset description



Health and Risk Factor Questionnaire

Overview

Acronym	CoreQx
Description	The Health and Risk Factor Questionnaire Dataset includes the variables collected at baseline by the five population-based cohorts that together compose the Canadian Partnership for Tomorrow (CPT) Project. The current data release includes the participants recruited up to March 2014 (202,404 participants). The harmonized dataset includes information on: • Socio-demographics and economic characteristics (ethnic background, education level, employment status, income, ...) • Personal and familial history of diseases • Personal and familial history of cancers • Reproductive history • Women's health, Men's health • Lifestyle and health behaviours (Alcohol Sun exposure, Physical activity (IPAQ)) • Self-reported physical measures (Height Additional information including medication, job
Number of variables	716

Number of variables	716

Variable Classification

Areas of Information

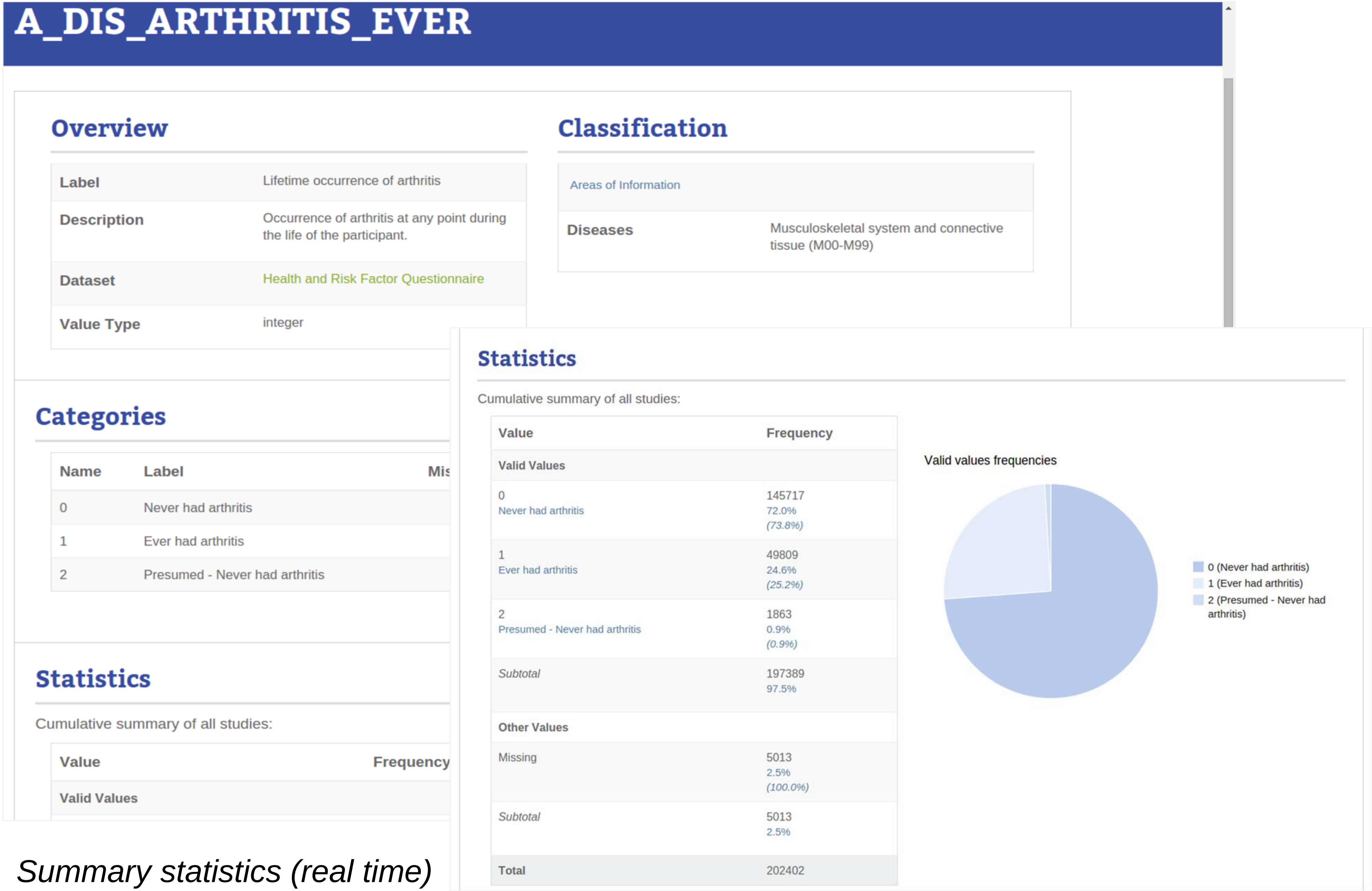
Number of variables
400
300
200
100
0
Socio-demographic and...
Lifestyle and health beh...
Health status and functi...
Diseases
Symptoms and signs
Medication and supplem...
Non-pharmacological int...
Health and community c...
Reproduction
Physical measures
Physical environment
Administrative information
Hide Harmonization Results

Harmonization

Click on each status icon to get more details on the corresponding harmonization results:

mica

Variable description



A_DIS_ARTHRITIS_EVER

Overview

Label	Lifetime occurrence of arthritis
Description	Occurrence of arthritis at any point during the life of the participant.
Dataset	Health and Risk Factor Questionnaire
Value Type	integer

Classification

Areas of Information

Diseases	Musculoskeletal system and connective tissue (M00-M99)

Categories

Name	Label	Mis
0	Never had arthritis	
1	Ever had arthritis	
2	Presumed - Never had arthritis	

Statistics

Cumulative summary of all studies:

Value	Frequency
Valid Values	

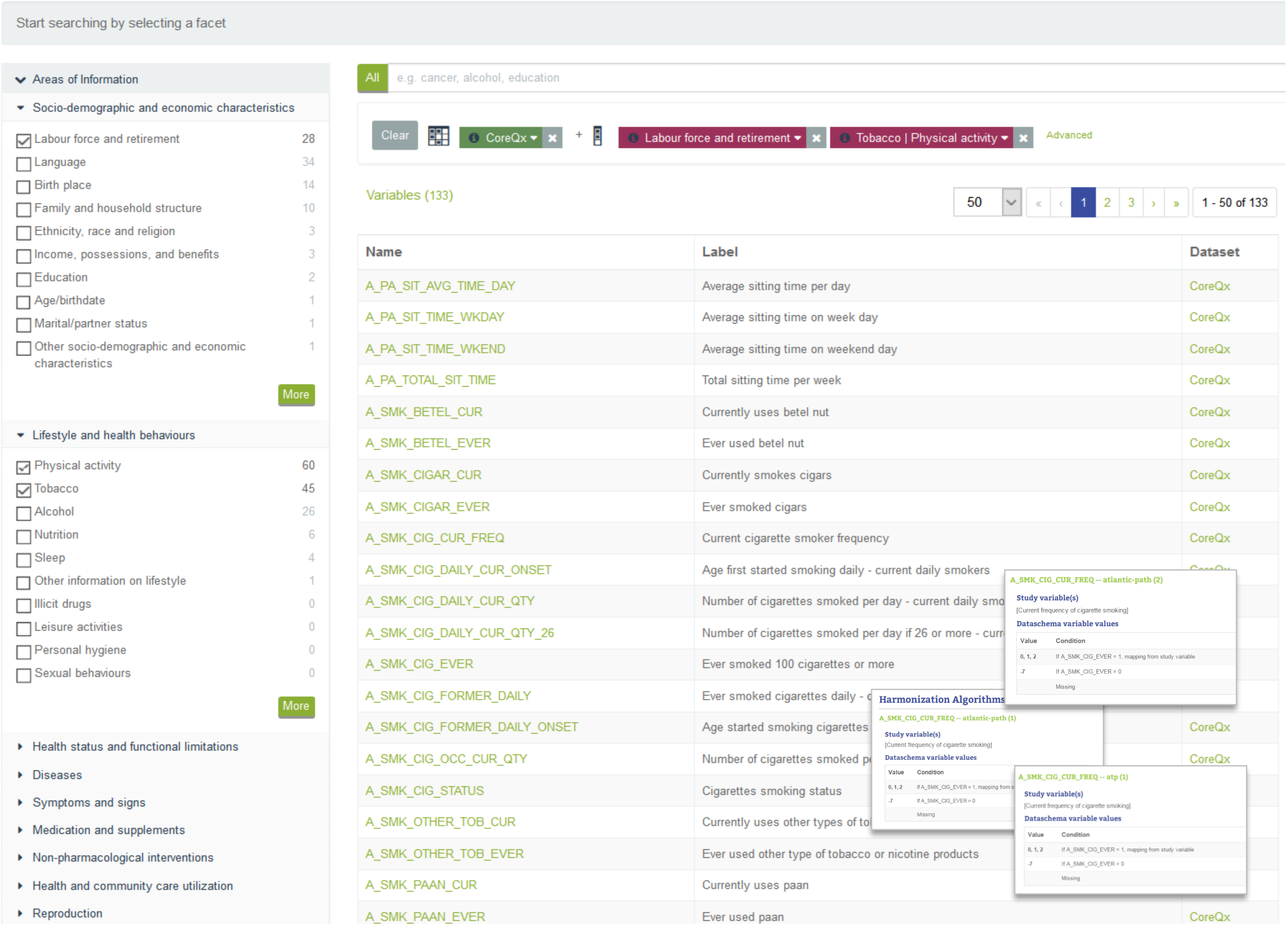
Statistics

Cumulative summary of all studies:

Value	Frequency
Valid Values	
0 Never had arthritis	145717 72.0% *(73.8%)*
1 Ever had arthritis	49809 24.6% *(25.2%)*
2 Presumed - Never had arthritis	1863 0.9% *(0.9%)*
Subtotal	197389 97.5%
Other Values	
Missing	5013 2.5% *(100.0%)*
Subtotal	5013 2.5%
Total	202402

Summary statistics (real time)

Variable search



Start searching by selecting a facet

Areas of Information

Socio-demographic and economic characteristics

☑ Labour force and retirement	28
□ Language	34
□ Birth place	14
□ Family and household structure	10
□ Ethnicity, race and religion	3
□ Income, possessions, and benefits	3
□ Education	2
□ Age/birthdate	1
□ Marital/partner status	1
□ Other socio-demographic and economic characteristics	1

More

Lifestyle and health behaviours

☑ Physical activity	60
☑ Tobacco	45
□ Alcohol	26
□ Nutrition	6
□ Sleep	4
□ Other information on lifestyle	1
□ Illicit drugs	0
□ Leisure activities	0
□ Personal hygiene	0
□ Sexual behaviours	0

More

Health status and functional limitations

Diseases

Symptoms and signs

Medication and supplements

Non-pharmacological interventions

Health and community care utilization

Reproduction

All e.g. cancer, alcohol, education

Clear | CoreQx ✖ | + | Labour force and retirement ✖ | Tobacco | Physical activity ✖ | Advanced

Variables (133)

50 « ‹ 1 2 3 › » 1 - 50 of 133

Name	Label	Dataset
A_PA_SIT_AVG_TIME_DAY	Average sitting time per day	CoreQx
A_PA_SIT_TIME_WKDAY	Average sitting time on week day	CoreQx
A_PA_SIT_TIME_WKEND	Average sitting time on weekend day	CoreQx
A_PA_TOTAL_SIT_TIME	Total sitting time per week	CoreQx
A_SMK_BETEL_CUR	Currently uses betel nut	CoreQx
A_SMK_BETEL_EVER	Ever used betel nut	CoreQx
A_SMK_CIGAR_CUR	Currently smokes cigars	CoreQx
A_SMK_CIGAR_EVER	Ever smoked cigars	CoreQx
A_SMK_CIG_CUR_FREQ	Current cigarette smoker frequency	CoreQx
A_SMK_CIG_DAILY_CUR_ONSET	Age first started smoking daily - current daily smokers	CoreQx
A_SMK_CIG_DAILY_CUR_QTY	Number of cigarettes smoked per day - current daily smo	
A_SMK_CIG_DAILY_CUR_QTY_26	Number of cigarettes smoked per day if 26 or more - curr	
A_SMK_CIG_EVER	Ever smoked 100 cigarettes or more	
A_SMK_CIG_FORMER_DAILY	Ever smoked cigarettes daily - c	
A_SMK_CIG_FORMER_DAILY_ONSET	Age started smoking cigarettes	CoreQx
A_SMK_CIG_OCC_CUR_QTY	Number of cigarettes smoked p	CoreQx
A_SMK_CIG_STATUS	Cigarettes smoking status	
A_SMK_OTHER_TOB_CUR	Currently uses other types of to	
A_SMK_OTHER_TOB_EVER	Ever used other type of tobacco or nicotine products	
A_SMK_PAAN_CUR	Currently uses paan	
A_SMK_PAAN_EVER	Ever used paan	CoreQx

A_SMK_CIG_CUR_FREQ -- atlantic-path (2)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, mapping from study variable
-7	If A_SMK_CIG_EVER = 0
	Missing

Harmonization Algorithms

A_SMK_CIG_CUR_FREQ -- atlantic-path (1)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, mapping from s
-7	If A_SMK_CIG_EVER = 0
	Missing

A_SMK_CIG_CUR_FREQ -- atp (1)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, mapping from study variable
-7	If A_SMK_CIG_EVER = 0
	Missing