Lessons from CanPath on demystifying data harmonization

Isabel Fortier, Anouar Nechba

Webinar, June 10th, 2025

Centre universitaire de santé McGill
Institut de recherche
McGill University Health Centre
Research Institute

mælstrom

Initiatives harmonizing data

Different research areas, focuses, sizes, …

A paradigm shift in the way we conduct research

Influenced by the need to:
obtain **larger sample sizes** and statistical power;
conduct **comparative research** across studies/jurisdictions;
optimize the **impact/usage** of individual studies/data sources.

ReACH
Research Advancement through Cohort Cataloguing and Harmonization

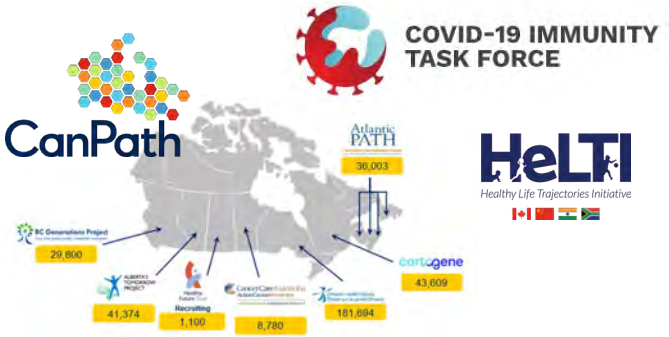
SPHN

Sepsis Canada

EUCAN connect

Research networks aiming to facilitate data usage, harmonization, sharing, …

Large initiatives generating and offering access to harmonized data

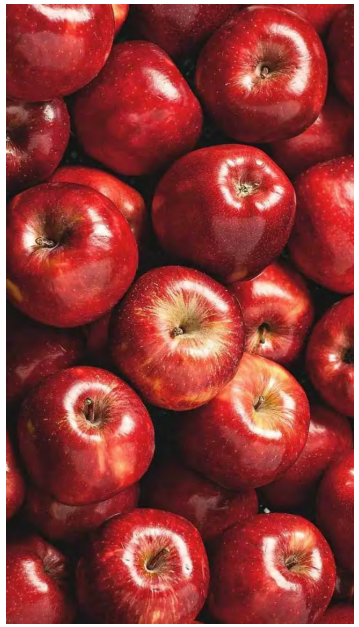


Focussed research projects

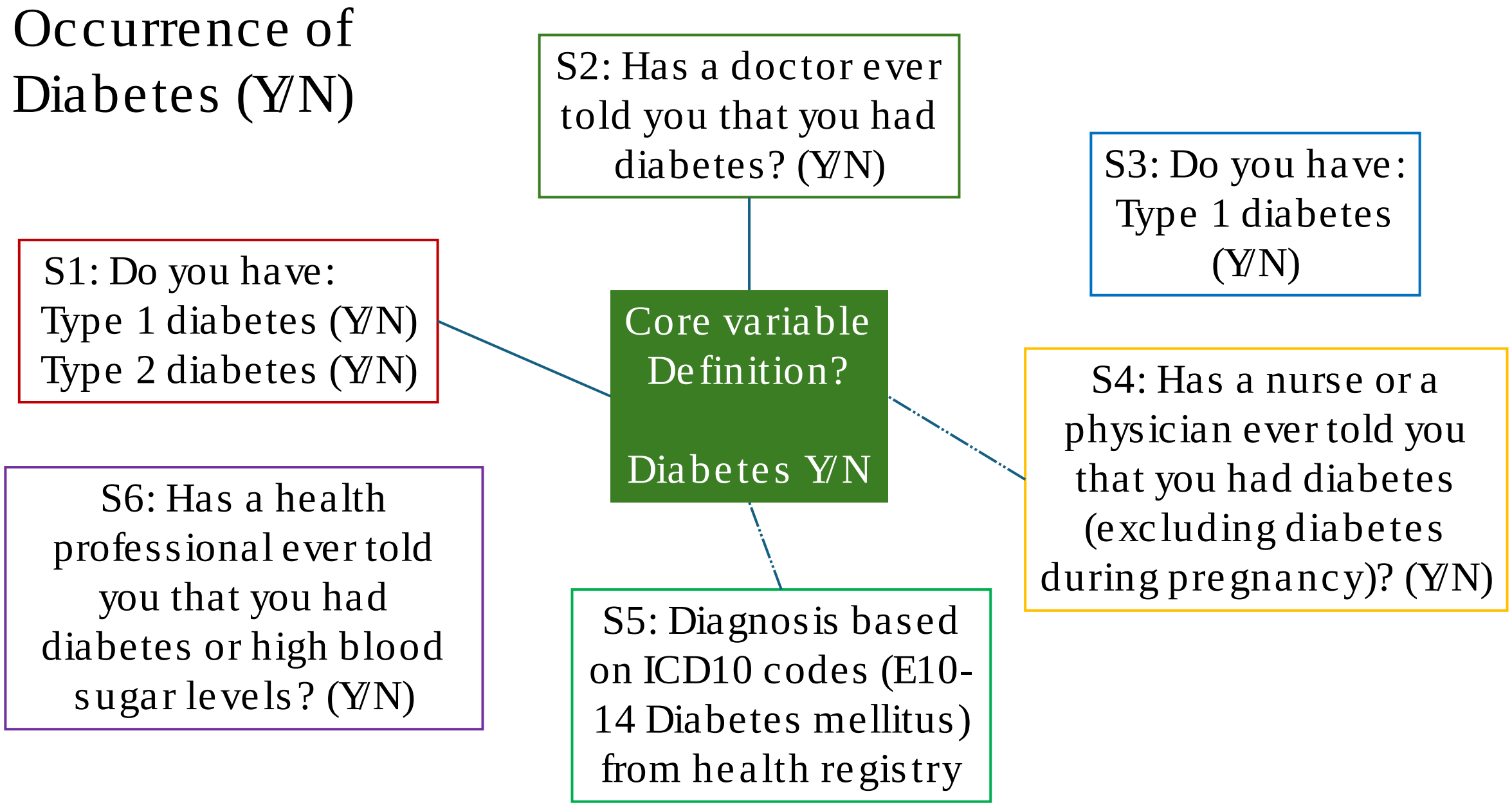
Paper on factors associated with alcohol consumption during pregnancy (ABocking)

What is data harmonization?

Process achieved to ‘’transform’’ study-specific individual participant data at an acceptable level of compatibility to support co-analysis across studies.



Occurrence of Diabetes (Y/N)



Harmonize

Data from existing databases?

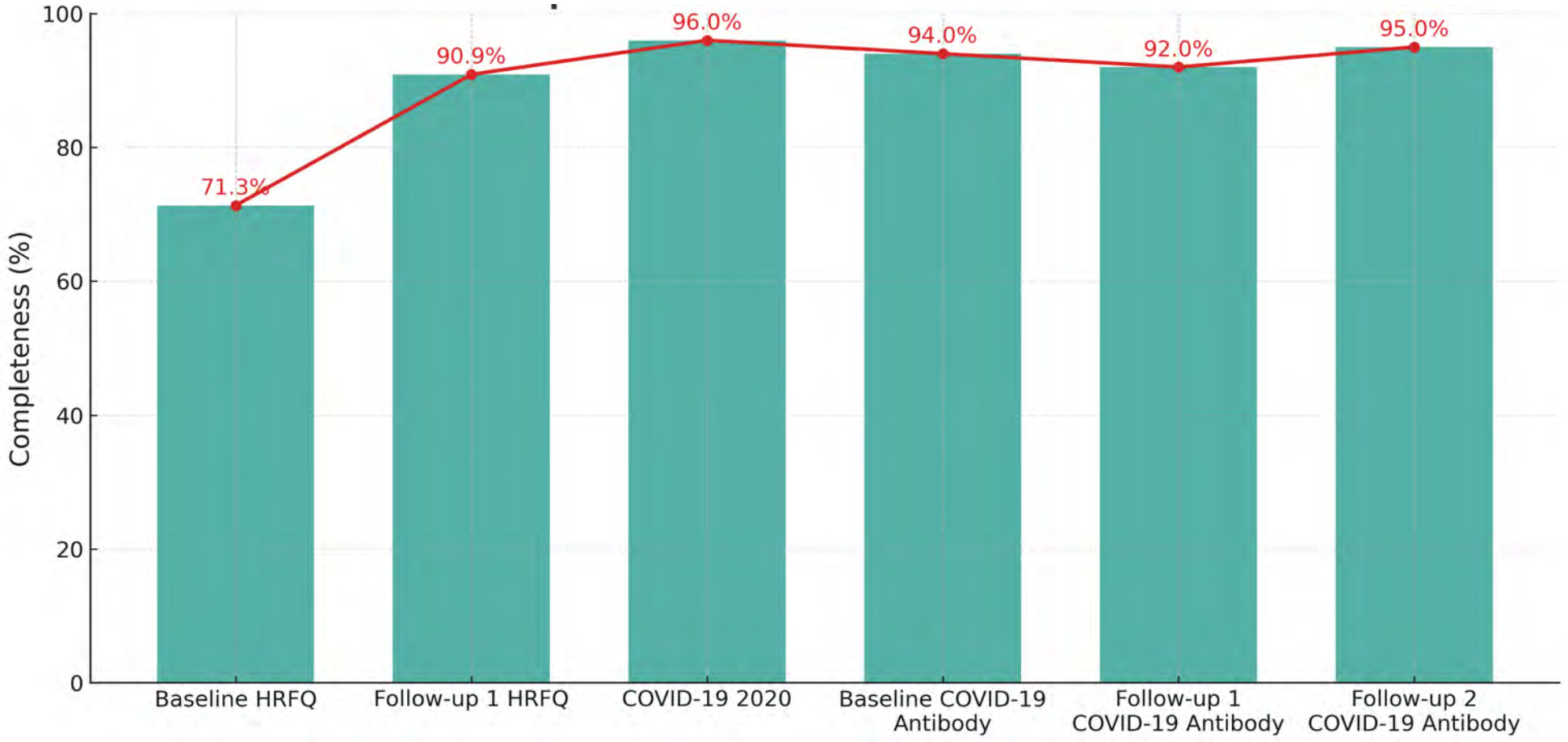
Quantity and quality of information that can be shared limited by heterogeneity.

Data to be collected in the future?

Common standards and methods to collect information can improve harmonization potential.

CanPath

Harmonization potential



www.maelstrom-research.org

- International research platform created in 2012, based on activities initiated in 2004 (P^3G)
- Collaborations with over 60 national and international partners

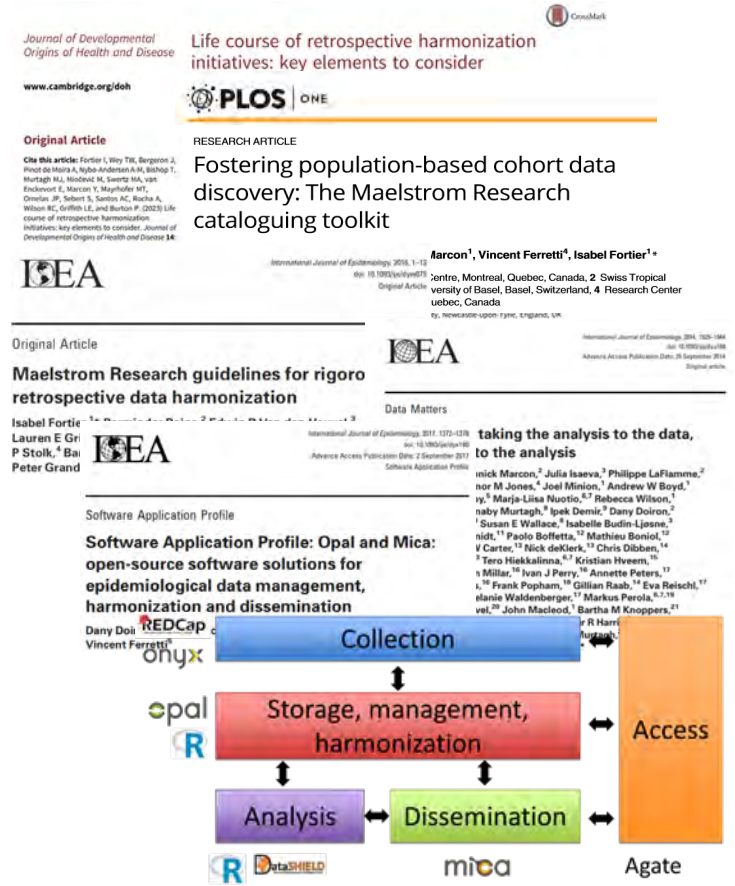
In collaboration with our partners we:

Develop tools and methodological guidelines to support data cataloguing, management, harmonization and co-analysis

Maintain a central study catalogue to foster discovery and usage of collected data

31 networks, 446 studies, > 1,840,000 variables

Support research initiatives to implement and maintain data harmonization platforms



P³G-OBiBa-Maelstrom and CanPath Partnership (2008-ongoing)

Methods, standards…

	Metadata	Requisite	Definitions
RESEARCH INITIATIVE			
Background	Name	Mandatory	Full name of the research intiative
	Acronym	Optional	Acronym of the research intiative
	Start year	Mandatory	Start year of the research intiative
	End year	Optional	If applicable, end year of the research initiative
	Objectives	Mandatory	Ob
	Investigators	Optional	Lis
	Contact	Optional	Na ha
	Publication(s)	Mandatory if applicable	If a da
	Harmonization protocols	Mandatory	If pr
	Additional	Optional	Su
HARMONIZATION PROTOCOL			
Background	Title	Optional	Tit
	Acronym	Optional	Ac
	Objectives	Optional	De
	Information content	Mandatory	Ou pr
	Exploratory/actual harmonization	Optional	Ou die da
	Ex-post/ ex-ante approach	Optional	Ou de co
	Procedures	Optional	De re
	Infrastructure	Optional	De ho
	Additional	Optional	Su
Participating studies/databases	Study/database selection	Mandatory	Cr
	Study/database list	Mandatory	List and characteristics (e.g. study design, participant sampling frame, nb of participants, etc) of the individual studies/databases included in the harmonization project.
DataSchema	Identification		
	Number of var		
	Variable defini		
	Variables inclu		
	Variable list		
Study/database datasets	Datasets to be		
	Participants inc		
	Variables acceptability		
	Descriptive sta		
Processing elements	Harmonization		
	Processing rule		
Harmonized dataset(s)	Dataset(s)		
	Sample size		
	Descriptive sta		

DS variable-specific information		Generic standard proposed	
		Requisite	Definitions
Index		Optional	Position of the DS variable in the variable list.
Harmonization project		Optional	Acronym of the harmonization project.
Table		Optional	Name of the table (or dataset) including the DS variables.
Name		Mandatory	Name of the DS variable.
Label		Mandatory	Short description of the DS variable specifying its content.
Classification		Optional	Classification of the DS variable in specific domains.
Format	**Value type**	Mandatory	Value type of DS variable (e.g., boolean, binary, date, datetime, time, decimal, integer, text)
	Units	Mandatory (if applicable)	For relevant decimal, integer, date, datetime, time DS variables (e.g., height, date of birth), units of the variable (e.g., cm, yyyy-mm-dd)
	Categories	Mandatory (if applicable)	For relevant categorical, integer, boolean and binary DS variables, coding and description of the response options (e.g., 0=never, 1=sometimes, 2=often, 9=prefer not to answer).
	Other	Mandatory (if applicable)	For any other type of format, (e.g., postal codes digits, ID), relevent information about the variable format
Target	**Individual/ entity**	Mandatory	Individual (or entity) the variable is about (e.g., participant, sample, postal code).
	Time period	Mandatory	Time period the variable refers to (or is about) (e.g., first trimester of pregnancy, last 30 days, at birth).
Requirements	**Source**	Mandatory	Acceptable source of information (e.g., questionnaire, governmental database/registrie, etc.) from which the variable data could be extracted
	Informant	Mandatory (if applicable)	For information that can be provided by different persons, acceptable informant (e.g. participant, partner, proxy)
	Time of collection	Mandatory	Acceptable time of collection of the information (e.g., smoking status during first trimester of pregnancy can be collected at birth(or not)).
	Wording	Mandatory (if applicable)	For data extracted from questionnaires, acceptable question wording (e.g., binge drinking defined as 5 drinks or more).
	Measures	Mandatory (if applicable)	For data that can be measured using different tools/devices, acceptable tool or device (e.g. GAD-7, PHQ-9, mercury gravity manometer, etc.).
	Procedures	Mandatory (if applicable)	For data that can be generated using different procedures, acceptable procedures (e.g., weight needs to be measured, not reported by the participant).
Instructions		Mandatory (if applicable)	For variables to be generated using specific standard or instructions description of the conditions to be respected (e.g. ICD-10 must be used, prematurity defined as birth before 37 weeks of pregnancy according to a given reference, specific information related to management of unique or multiple choices option, etc).
Comments		Optional	Any additional information not covered

Information	Generic standard proposed	
	Requisite	Definitions
Index	Optional	Position of the DS variable in the variable list.
Study-specific DataSchema table	Optional	Name of the harmonized dataset including the DS variables.
DataSchema variable	Mandatory	Name of the DataSchema variable.
Acronym of the study	Mandatory	Acronym of the study.
Name of the study-specific dataset	Mandatory	Name of the study-specific dataset after preprocess of cleaning and ready to be used to generate the DS variables.
Name of the study-specific variable(s)	Mandatory	Name of the study-specific variable(s) to be used to generate the DS variable.
Format	Optional	Any information related to the format of the study-specific variable (s) : Value type :type of study-specific variable(s) (e.g., boolean, binary, date, datetime, time, decimal, integer, text) Units : for relevant decimal, integer and text variables (e.g., height, date of birth), units of the study-specific variable(s) (e.g., cm, yyyy-mm-dd) Categories : categories of the study-specifi
Time period	Optional	Time period targeted by the study-specific
Time of collection	Optional	Time of collection of the study-specific var
Source of information	Optional	Acceptable source of information (e.g., qu
Informant	Optional	For information that can be provided by d
Keywords	Optional	Any information related to the wording, m Wording: acceptable or required informat Measures: acceptable or required informa 7,PHQ-9, mercury-gravity manometer (Exh Procedures: acceptable or required inform reported by the participant). Instructions: acceptable or required infom prematurity defined before 37 weeks of pr
Internal Comment	Optional	Internal comments or questions to the tea
Harmonization comment to be published	Optional	Any additional relevant information on th
Harmonization status	Mandatory	Harmonization status of the study-specific
Harmonization status details	Mandatory	Harmonization status detail of the study-s
Rule category	Mandatory (if applicable)	Rule category used to generate the study-s
Harmonization rule	Mandatory (if applicable)	Simplified algorithms to be used to generate the study-specific harmonized variable.

Software

epal

To facilitate data storage and management

micca

To build online catalogues describing study-specific and harmonized data content, and managing demands for access to data

Rpackages

Rmonize

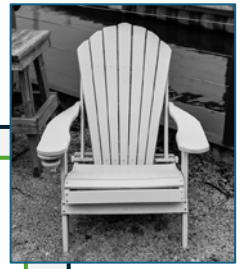
To improve cost-effectiveness, consistency and transparency of data harmonization processes

Harmonization process

Prepare input elements

Study-specific datasets

Core variables to be generated across studies



Data Processing Elements required to transform study-specific into harmonized data

Process data

Process study-specific data under the harmonized data format

Prepare harmonized data

Estimate quality of the harmonized data
Generate reports to understand the harmonized data
Consolidate harmonized datasets ready for analysis

Use harmonized data !

Rmonize

mælstrom

Study-specific datasets

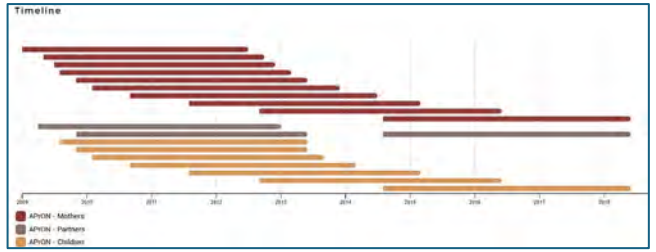
Participants selection criteria?

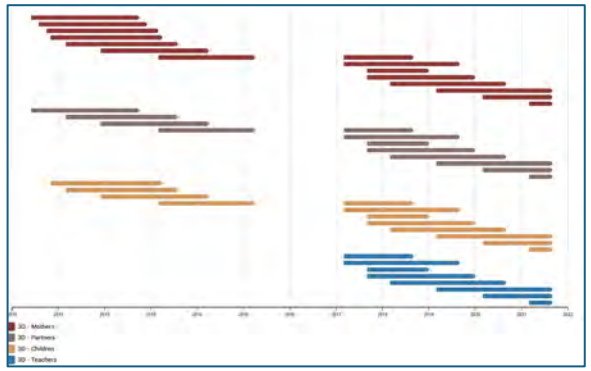
Number of participants?

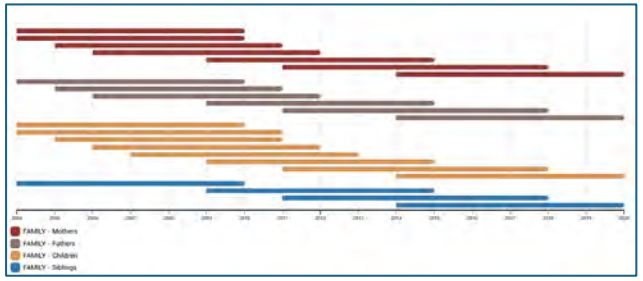
Number/timing of follow ups?

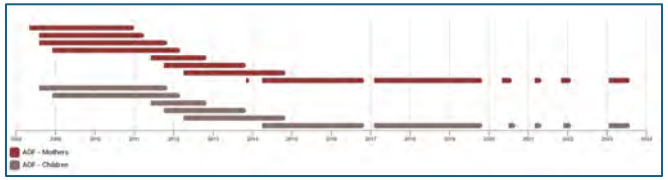
Specific information collected?

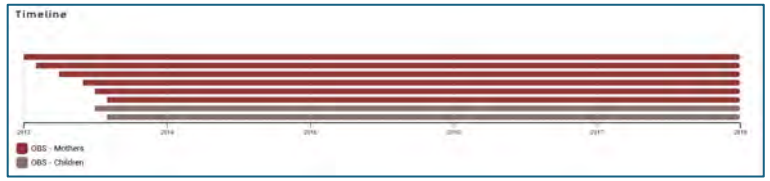
Quality of the data? ….











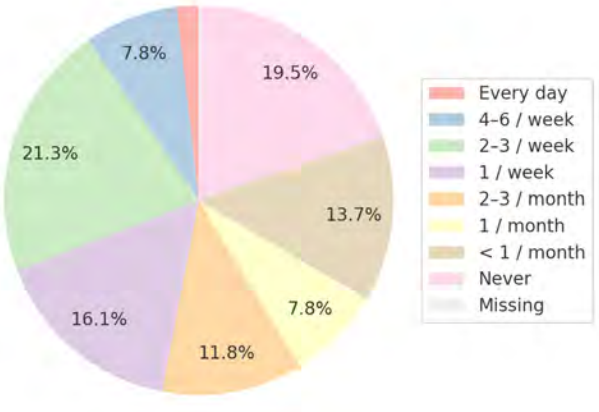
Required to :

- Understand studies characteristics (e.g., exclusions criteria)
- Transform data and data dictionaries into appropriate formats
- Explore data content and quality (e.g., identify outliers, duplicates)

Study 1

Alcohol intake during year before pregnancy

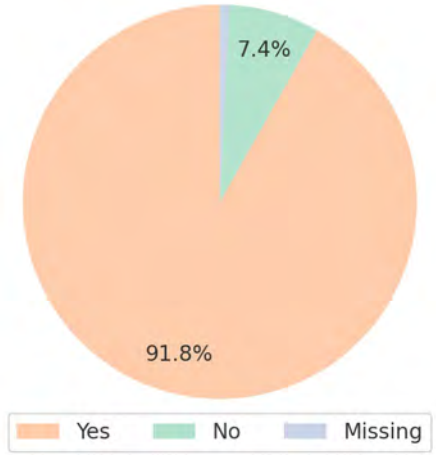
Category	Count
Every day	43
4–6 / week	185
2–3 / week	503
1 / week	380
2–3 / month	278
1 / month	185
< 1 / month	325
Never	461
Missing	5
Total	2 365



Study 2

Alcohol use 12 months before pregnancy

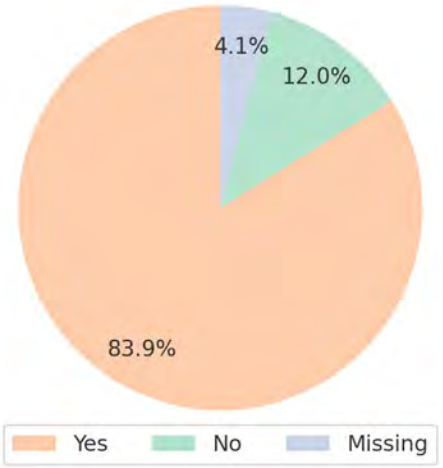
Category	Count
Yes	2 728
No	221
Missing	23
Total	2 972



Study 3

Ever consumed alcohol

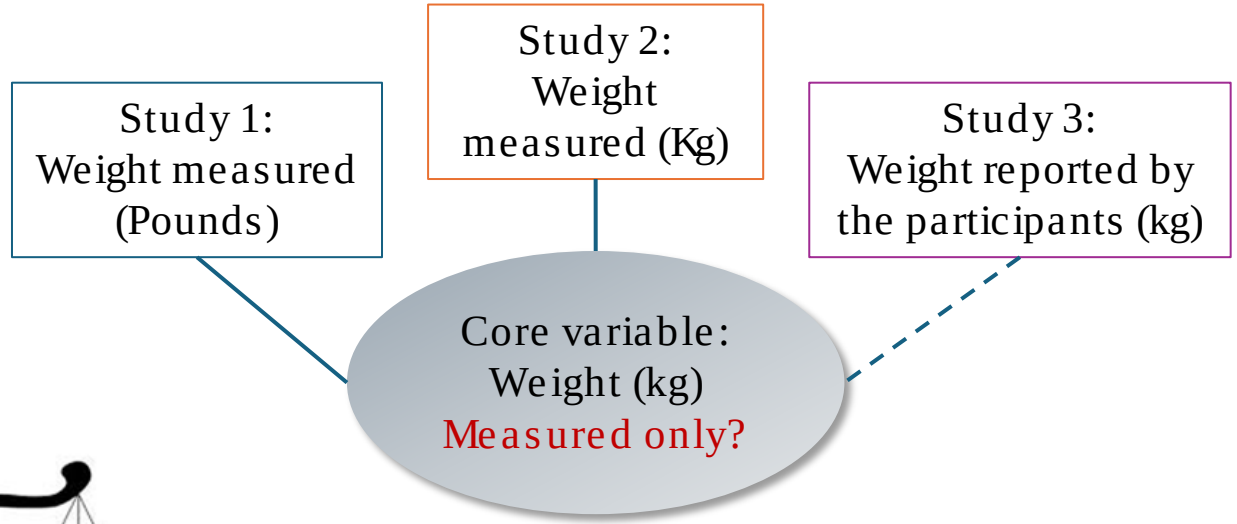
Category	Count
Yes	1 835
No	262
Missing	90
Total	2 187

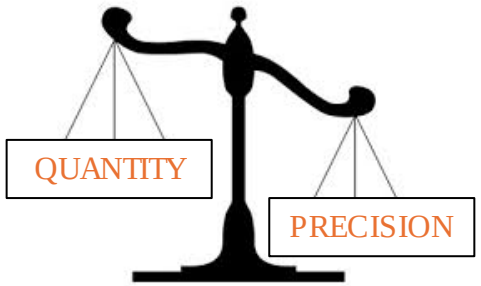


Core Variables - exemples

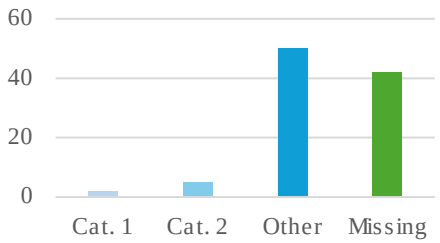
Alcohol consumption of the mother 1 year prior to pregnancy	Categories: **0= No ; 1= Yes**
Binge drinking of the mother during pregnancy	Categories: 0= No ; 1= Yes
Marital status of the mother at first visit	Categories: 1= Single (never married) or not living with partner, 2= Married or living with partner, 3= Divorced or separated, 4=Widowed
Weight of the mother at recruitment	Units=grams
Gestational age of the baby at delivery	Units=weeks
Biological sex of the baby	Categories: 1=Female, 2=Male
Birthweight of the baby	Units=grams
Ethnicity	….???

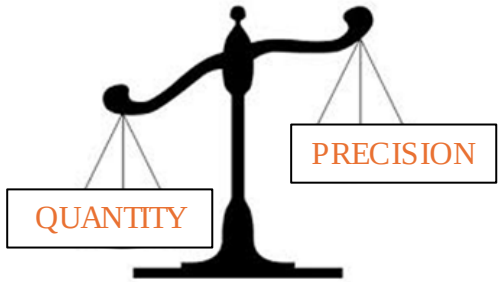
Define core variables



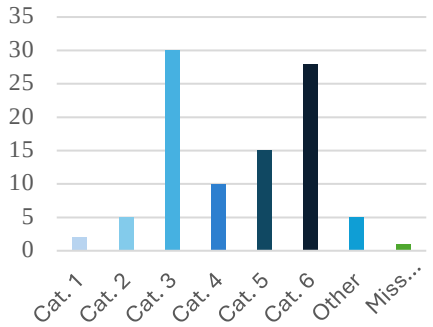


	Study 1	Study 2	Study 3	Study 4	Study 5
Variable X	✓	✓	✓	✓	✓





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



Variable definition

Occurrence of Diabetes

S1: Do you have: Type 1 diabetes (Y/N)
Type 2 diabetes (Y/N)

S2: Has a doctor ever told you that you had diabetes? (Y/N)

S3: Do you have: Type 1 diabetes (Y/N)

S4: Has a nurse or a physician ever told you that you had diabetes (excluding diabetes during pregnancy)? (Y/N)

S5: Has a health professional ever told you that you had diabetes or high blood sugar levels? (Y/N)

S6: Diagnosis based on ICD10 codes (E10-14 Diabetes mellitus) from health registry

Core variable to be generated

Definition:
Type 1, Type 2 or both?

Definition:
Inclusion/exclusion of gestational diabetes?

Definition:
Wording should mention that diagnosed by a physician, a health professional or not?

Definition:
Diagnosis can include high blood sugar?

Source:
Questionnaire or health registry?

Ethnicity

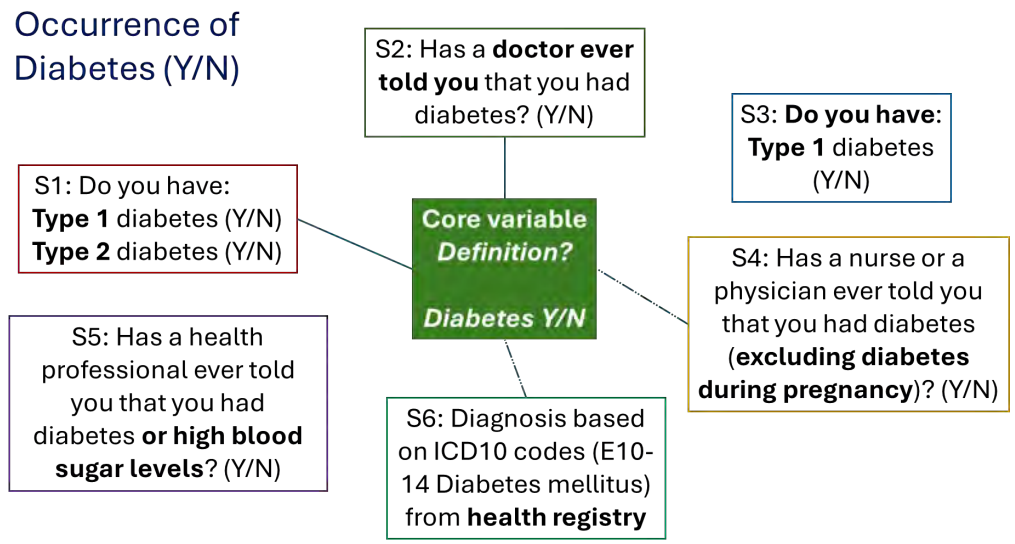
	STUDY				
ETHNICITY	**S 1**	**S 2**	**S 3**	**S 4**	**S 5**
MULTIPLE SELECTION	YES	YES	NO	NO	NO
FIRST NATIONS	x	x	x	x	x
ARAB		x	x	x	x
WEST ASIAN		x	x	x	
ARAB/WEST ASIAN	x				
BLACK	x	x	x	x	x
JEWISH		x			
LATIN AMERICAN/HISPANIC	x	x	x	x	x
EAST ASIAN	x	x			
CHINESE			x	x	
KOREAN				x	
JAPANESE				x	
FILIPINO		x	x	x	
SOUTH ASIAN	x	x	x	x	x
SOUTHEAST ASIAN		x	x	x	
EAST/ SOUTHEAST ASIAN					x
WHITE/CAUCASIAN	x	x	x	x	x
OTHER	x	x		x	x
OTHER SPECIFY	x				x

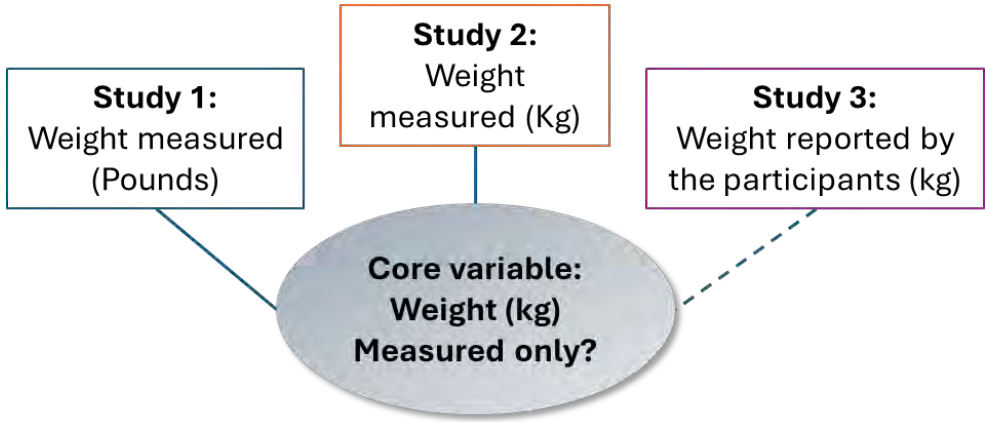
# of mothers: 2,456 Province: QC Start year: 2010	# of mothers: 3,387 Province: AB Start year: 2008	# of mothers: 2,189 Province: AB Start year: 2009	# of mothers: 857 Province: ON Start year: 2004	# of mothers: 1,374 Province: ON Start year: 2013

FINAL DISTINCT VARIABLES?
FIRST NATIONS
BLACK
LATIN AMERICAN/HISPANIC
SOUTH ASIAN
WHITE/CAUCASIAN
OTHER

Data Processing Elements

- Harmonization statuses outlining the possibility (or not), for each study-specific dataset, to generate each core variable
  - If deemed possible, the algorithms to be applied to process input into harmonized data and
    - were relevant, information helping future users to understand the harmonized data content.





Alcohol consumption of the mother 1 year prior to pregnancy	Categories: **0= No ; 1= Yes**

Study 1

Alcohol intake during year before pregnancy

Category	Count
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

recode("Every day" = 1; " 4–6 / week " = 1; " 2–3 / week " = 1; " 1 / week " = 1 ; " 2–3 / month" = 1; " 1 / month " = 1; " < 1 / month "=1 " Never " = 0 ; ELSE~NA)

Study 2

Alcohol use 12 months before pregnancy

Category	Count
Yes	2 728
No	221
Missing	23
Total	2 972

Direct mapping from study variable

Study 3

Ever consumed alcohol

Category	Count
Yes	1 835
No	262
Missing	90
Total	2 187

Impossible

Information on alcohol consumption not collected before pregnancy

Process study-specific into harmonized data

Validate and consolidate harmonized data

Inputs

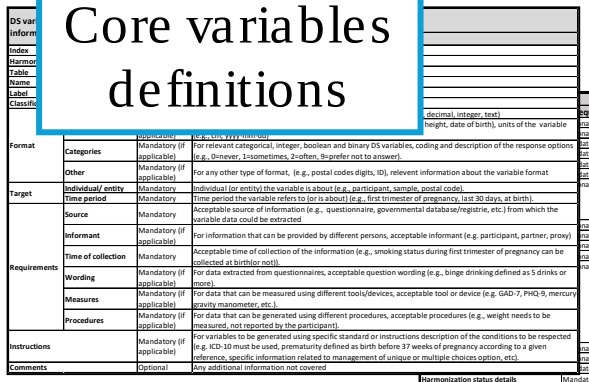
Study-specific data

Category	Count
Yes	2 728
No	221
Missing	23
Total	2 972

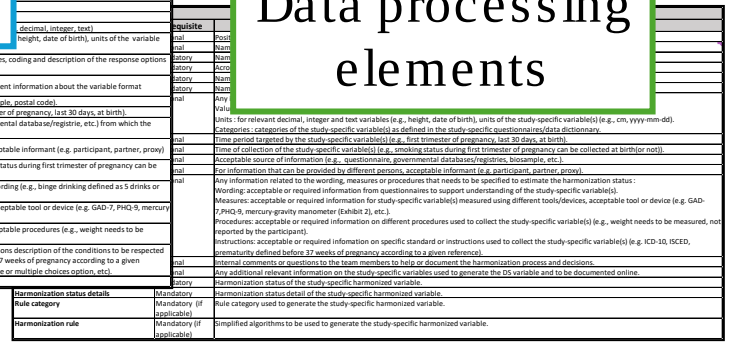
Category	Count
Yes	1 835
No	262
Missing	90
Total	2 187

Category	Count
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

Core variables definitions



Data processing elements

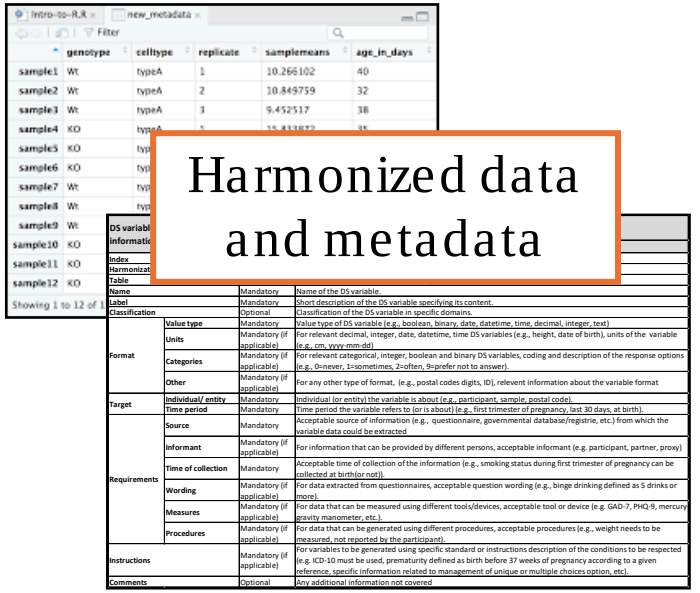


Rmonize

RStudio

Outputs

Harmonized data and metadata

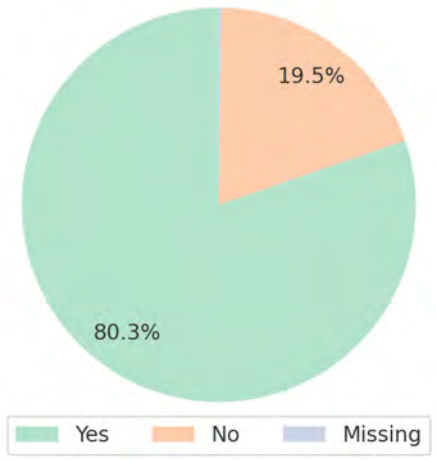


Alcohol consumption of the mother 1 year prior to pregnancy (Yes/No)

Study-specific harmonized variables

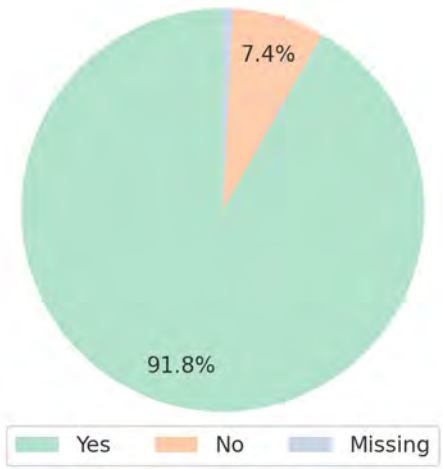
Study 1

Category	Count
Yes	1 899
No	461
Missing	5
Total	2 365



Study 2

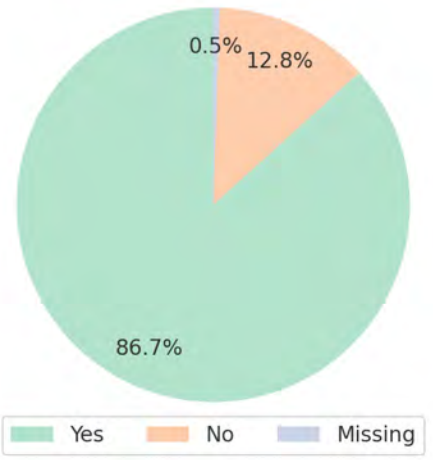
Category	Count
Yes	2 728
No	221
Missing	23
Total	2 972



Study 3

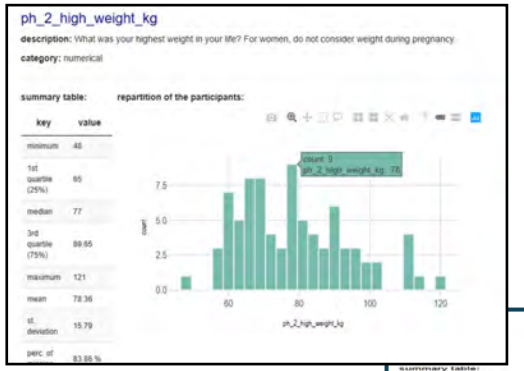
Combined harmonized data

Category	Count
Yes	4 627
No	682
Missing	28
Total	5 337

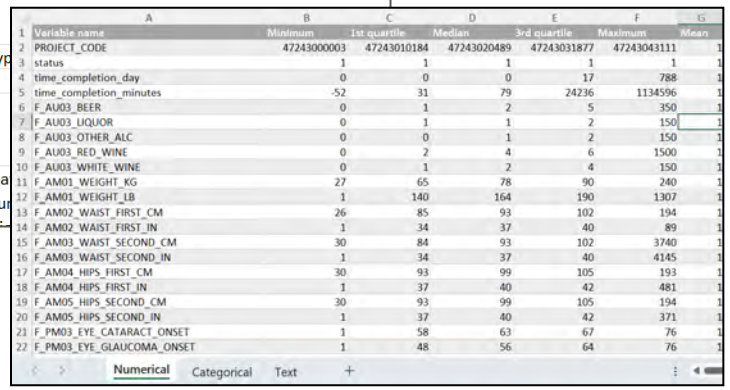


Use data!

Steps	MVPA (se	Bouts	TPA (seco	MPA (secc	VPA (seco	M Thresh	V Thresh	Filter (sec	Q10-1	Q10-2	Q10-3	Q10-4	Q10-5	Q10-6	Q10-7	Q10-8	Q10-9	Q10-10	Q10-11
5262	2074	1	3025	2045	29	100	130	4	0	0	0	0	0	0	0	0	0	0	0
7628	2721	1	4092	2066	655	100	130	4	0	0	0	0	0	0	0	0	0	0	0
12100	4904	3	6303	4056	848	100	130	4	0	0	0	0	0	0	0	0	0	0	0
10282	3899	3	6099	3875	24	100	130	4	0	0	0	0	0	0	0	0	0	0	0
1104	288	0	693	288	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66	25	0	46	25	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
46	14	0	29	14	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
4191	1465	0	2416	1384	81	100	130	4	0	0	0	0	0	0	0	0	0	0	0
3749	1073	0	2233	975	98	100	130	4	0	0	0	0	0	0	0	0	0	0	0
6929	2375	1	4131	2344	31	100	130	4	0	0	0	0	0	0	0	0	0	0	0
2756	595	0	1750	578	17	100	130	4	0	0	0	0	0	0	0	0	0	0	0
3367	801	0	2114	801	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
19	0	0	13	0	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
18	5	0	10	5	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
109	32	0	65	32	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
21	11	0	11	11	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
59	24	0	30	24	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0
22	6	0	13	6	0	100	130	4	0	0	0	0	0	0	0	0	0	0	0

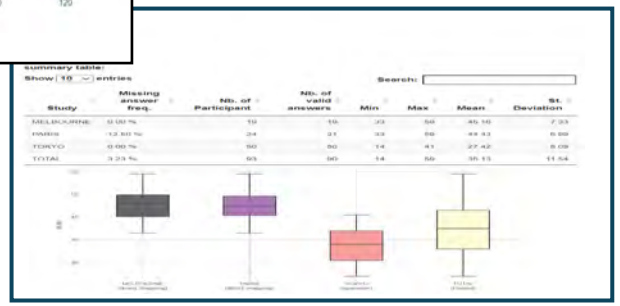


E	F	G	H	I	J
valueType	**Detected data type**	**Mix cat. and other values**	**Categories in data dictionary**	**unit**	**Quality assessment comment**
text	character				
integer	numeric		Valid categorical values : 0 = Male 1 = female		[INFO] - Possible duplicated variables: SEX ; Gender
integer	numeric			kg/m^2	[ERR] - ValueTyp
integer	numeric			years	
integer	numeric		Valid categorical values : 0 = Non smoker 1 = smoker		
integer	numeric	YES	Valid categorical values : 8 = ADMISSION PATTERN	cigarette per week	[ERR] - Incompa [INFO] - More u

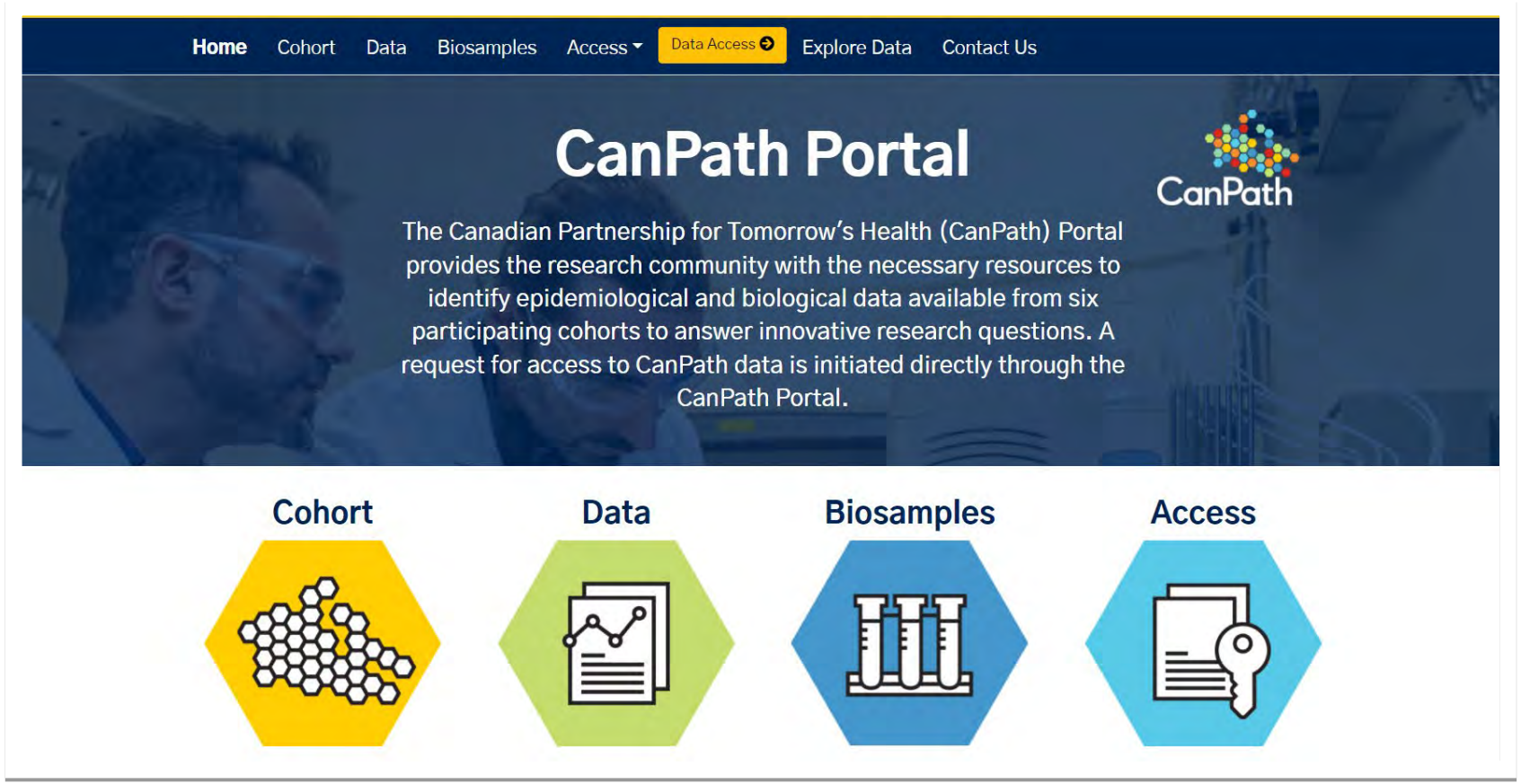


	A	B	C	D	E	F	G
1	Variable name	Minimum	1st quartile	Median	3rd quartile	Maximum	Mean
2	PROJECT_CODE	47243000003	47243010184	47243020489	47243031877	47243043111	1
3	status	1	1	1	1	1	1
4	time_completion_day	0	0	0	17	788	1
5	time_completion_minutes	-52	31	79	24236	1134596	1
6	F_AU03_BEER	0	1	2	5	350	1
7	F_AU03_LIQUOR	0	1	1	2	150	1
8	F_AU03_OTHER_ALC	0	0	1	2	150	1
9	F_AU03_RED_WINE	0	2	4	6	1500	1
10	F_AU03_WHITE_WINE	0	1	2	4	150	1
11	F_AM01_WEIGHT_KG	27	65	78	90	240	1
12	F_AM01_WEIGHT_LB	1	140	164	190	1307	1
13	F_AM02_WAIST_FIRST_CM	26	85	93	102	194	1
14	F_AM02_WAIST_FIRST_IN	1	34	37	40	89	1
15	F_AM03_WAIST_SECOND_CM	30	84	93	102	3740	1
16	F_AM03_WAIST_SECOND_IN	1	34	37	40	4145	1
17	F_AM04_HIPS_FIRST_CM	30	93	99	105	193	1
18	F_AM04_HIPS_FIRST_IN	1	37	40	42	481	1
19	F_AM05_HIPS_SECOND_CM	30	93	99	105	194	1
20	F_AM05_HIPS_SECOND_IN	1	37	40	42	371	1
21	F_PM03_EYE_CATARACT_ONSET	1	58	63	67	76	1
22	F_PM03_EYE_GLAUCOMA_ONSET	1	48	56	64	76	1

Numerical | Categorical | Text



4,500 harmonized variables!

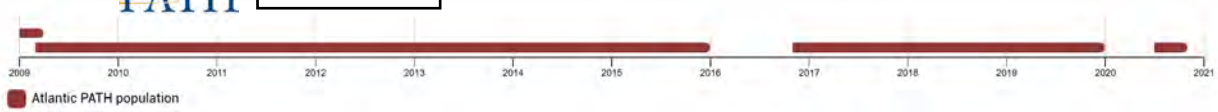


Fortier I, Dragieva N, Saliba M, Craig C, Robson PJ, with the Canadian Partnership for Tomorrow Project's scientific directors and the Harmonization Standing Committee Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis CMAJ Open 2019; 7:E272-E282

Study-specific datasets

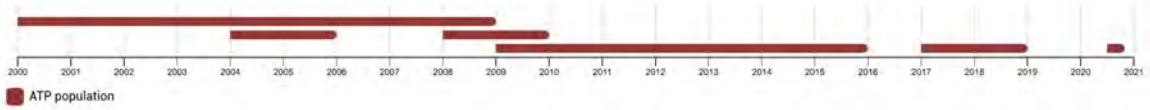
Atlantic PATH

30 - 74 years



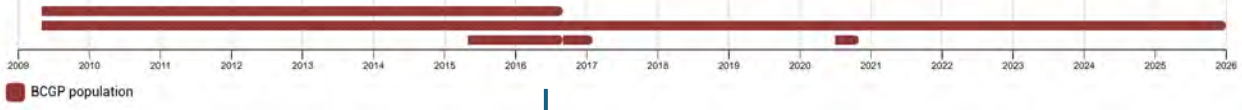
ALBERTA'S TOMORROW PROJECT
Inspiring research for a healthier tomorrow

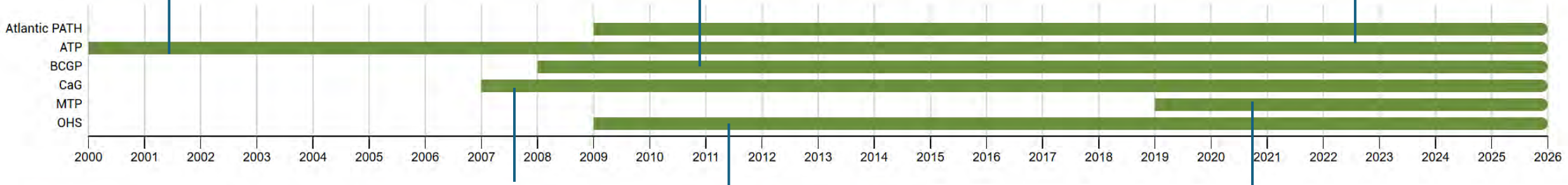
35 – 69 years, cancer-free (except non-melanoma skin cancer)

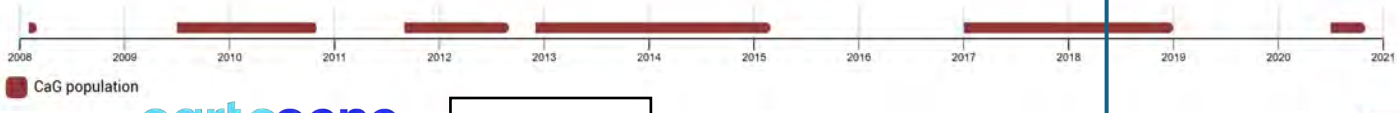


BC Generations Project
Your time today builds a healthier tomorrow.

35 – 74 years





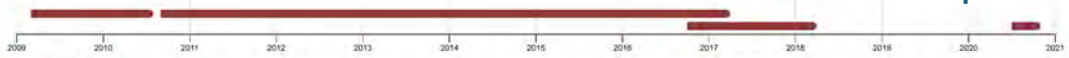


cartagene

40 – 69 years

CancerCare Manitoba
ActionCancerManitoba
THE MANITOBA TOMORROW PROJECT

30 – 74 years



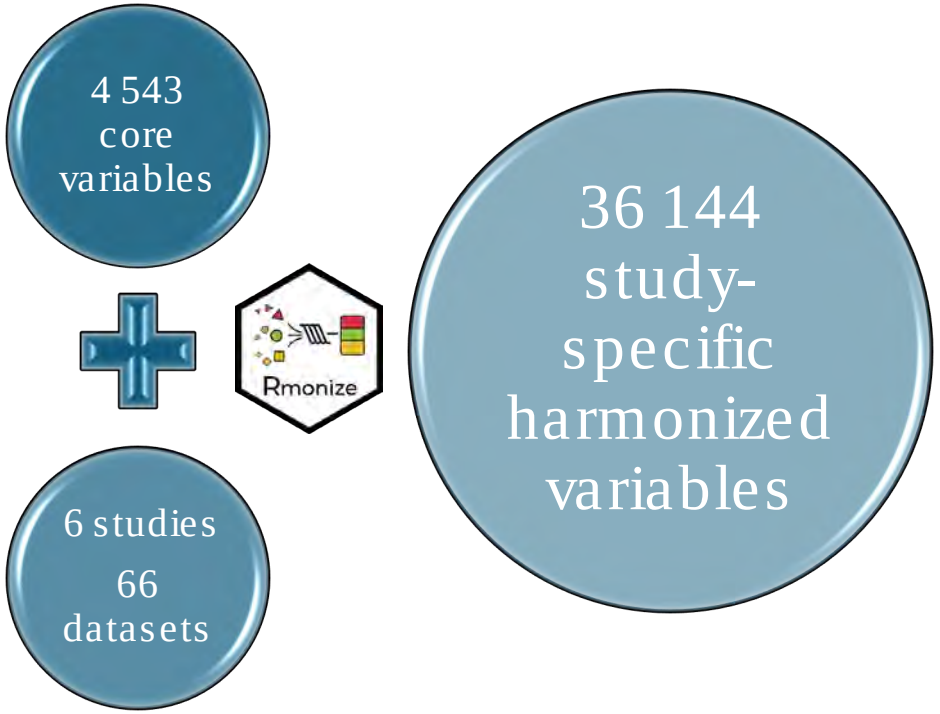
Ontario Health Study
Étude sur la santé Ontario

18 years and older

4,543 core variables

Questionnaires/Data collection events	Core variables
Health and Risk Factor Questionnaire - Baseline	1 481
Health and Risk Factor Questionnaire - Follow-up 1	858
COVID-19 (2020)	487
COVID-19 Antibody – 1st Phase	644
COVID-19 Antibody – 2nd Phase	267
COVID-19 Antibody – 3rd Phase	545

Biological samples collections	Core variables
Biological Sample - Baseline	103
COVID-19 Antibody – 1st Phase	24
COVID-19 Antibody – 2nd Phase	67
COVID-19 Antibody – 3rd Phase	67



Biological samples collections	Participants
Biological Sample - Baseline	152 935
COVID-19 Antibody – 1st Phase	26 236
COVID-19 Antibody – 2nd Phase	21 503
COVID-19 Antibody – 3rd Phase	9 110

Questionnaires/Data collection events	Participants
Health and Risk Factor Questionnaire - Baseline	327 630
Health and Risk Factor Questionnaire - Follow-up 1	134 435
COVID-19 (2020)	97 619
COVID-19 Antibody – 1st Phase	29 104
COVID-19 Antibody – 2nd Phase	22 911
COVID-19 Antibody – 3rd Phase	9 424

Harmonization potential

Health and Risk Factor Questionnaire - Baseline

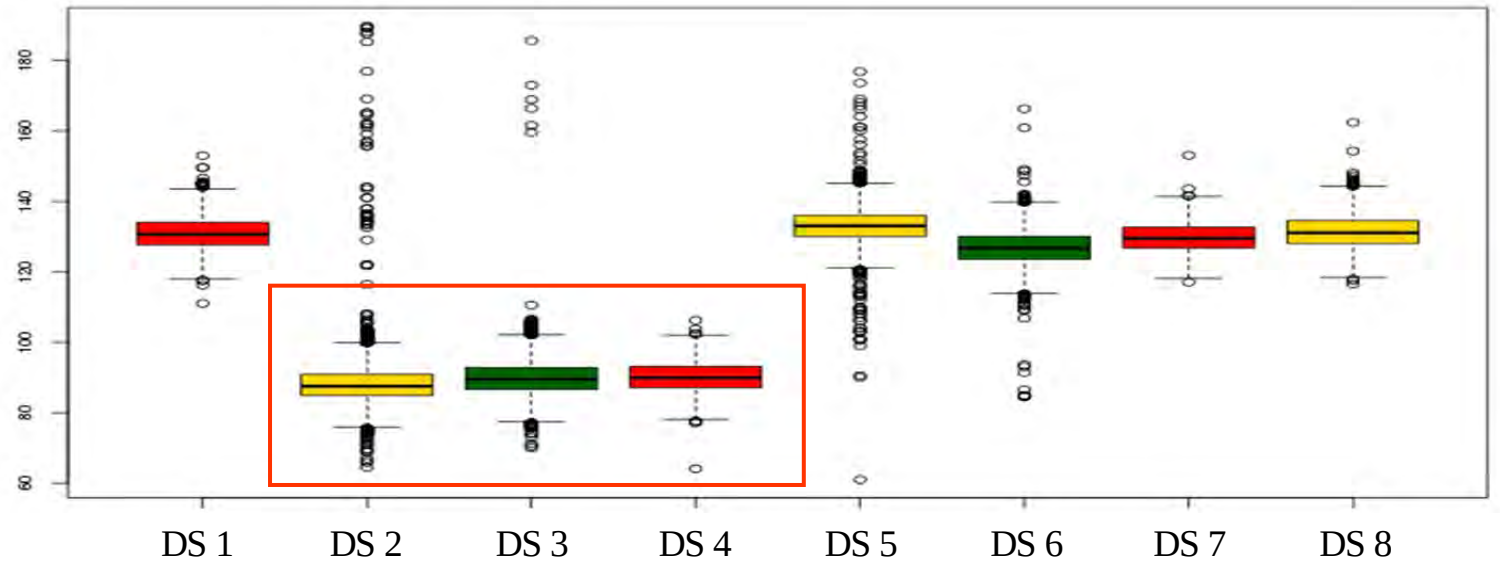
	% Complete	Atlantic PATH 1	Atlantic PATH 2	ATP 1	ATP 2	BCGP 1	BCGP 2	BCGP 3	CaG 1	CaG 2
		80%	49%	61%	79%	76%	75%	65%	62%	77%
Sex	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓
Number of brothers	75%	✓	✓	✗	✓	✓	✓	✓	✗	✗
Number of sisters	75%	✓	✓	✗	✓	✓	✓	✓	✗	✗
Number of siblings	92%	✓	✓	✓	✓	✓	✓	✓	✓	✗
Number of siblings/ Categories	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓

Quality control

Sitting height :

Distance (cm) from the buttocks to the head of the participant when he/she is sitting

	N	Mean
DS 1	4872	130,88
DS 2	22703	88,01
DS 3	29347	89,8
DS 4	1149	90,2
DS 5	16363	133,21
DS 6	19992	126,83
DS 7	649	129,78
DS 8	7970	131,3



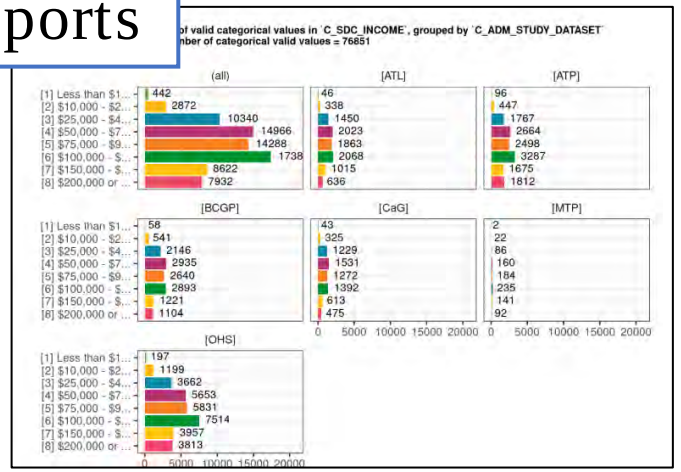
Use CanPath data!

Data portal

Datasets and data dictionaries!

GROUP1	AMELIE	ROY	F	-	2018/07/1	YES	Y00B25	PiezoRxD	7628	2721	1	4092	2066	655	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2018/07/1	YES	Y00B25	PiezoRxD	12100	4904	3	6303	4056	848	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2018/07/1	YES	Y00B25	PiezoRxD	10282	3899	3	6099	3875	24	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2018/07/1	YES	Y00B25	PiezoRxD	1104	288	0	693	288	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2018/09/1	NO	-	PiezoRxD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2018/10/2	YES	Y00B25	PiezoRxD	66	25	0	46	25	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2018/10/2	YES	Y00B25	PiezoRxD	46	14	0	29	14	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/1	YES	Y00B25	PiezoRxD	4191	1465	0	2416	1384	81	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/1	YES	Y00B25	PiezoRxD	3749	1073	0	2233	975	98	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/1	YES	Y00B25	PiezoRxD	6929	2375	1	4131	2344	31	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/1	YES	Y00B25	PiezoRxD	2756	595	0	1750	578	17	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/1	YES	Y00B25	PiezoRxD	3367	801	0	2114	801	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/2	YES	Y00B25	PiezoRxD	19	0	0	13	0	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/01/2	YES	Y00B25	PiezoRxD	18	5	0	10	5	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/02/2	YES	Y00B25	PiezoRxD	109	32	0	65	32	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/02/2	YES	Y00B25	PiezoRxD	21	11	0	11	11	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/02/2	YES	Y00B25	PiezoRxD	59	24	0	30	24	0	100	130	4	0	0	0	0	0	0	0
GROUP1	AMELIE	ROY	F	-	2019/02/2	YES	Y00B25	PiezoRxD	22	6	0	13	6	0	100	130	4	0	0	0	0	0	0	0

Reports



CanPath Data Portal

Participating study characteristics/designs

BC Generations Project

The British Columbia Generations Project (BCGP) is British Columbia

29,850
Number of participants

26,458
Number of participants with biological samples

Alberta's Tomorrow Project

Alberta's Tomorrow Project (ATP) is a longitudinal study tracking the health

55,000
Number of participants

30,000
Number of participants with biological samples

Manitoba Tomorrow Project

The Manitoba Tomorrow Project (MTP) is a population health research study that aims to learn more about the causes of cancer and chronic diseas following the health of 10,000 Manitoba residents for up to 50 years.

**The MTP is currently in the recruitment phase. Baseline data and biosamples ...

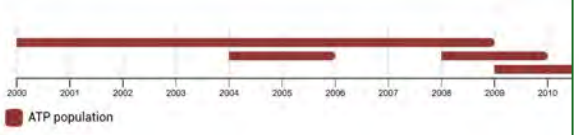
10,000
Number of participants

CARTaGENE

CARTaGENE is a long-term cohort study for the investigation of modifiable environmental and lifestyle factors and the genomic determinants of chronic diseases in Quebec, Canada.

43,046
Number of participants

29,792
Number of participants with biological samples



Harmonized datasets: list and description of data content

Overview

Information Content

The Baseline Health and Risk Factor Questionnaire Harmonization Pr includes 935 core variables.

The harmonized dataset covers information on participants':

- Socio-demographic and economic ch
- Lifestyle and behaviours (alcohol use,
- Reproductive health history
- Personal and family history of disease
- Use of prescription medications (medication name, DIN numbe
- Non-pharmacological interventions related to diseases
- Self-reported physical measures (height, weight, waist and hip
- Physical environment exposures (passive smoke exposure, sun
- Variables based on open-ended responses

Design

	Quantitative
	Prospective
	Harmonized data is generated by a central harmonization team (with the collaboration of studies).
	Harmonized data can be transferred to external servers, and pooled data analysis can be achieved.
Participant Inclusion	All CanPath participants from cohorts collecting data on health and risk factors at baseline are eligible.

Harmonization potential

Variable	% Complete	Atlantic PATH 1 80%	Atlantic PATH 2 49%	ATP 1 61%	ATP 2 79%	BCGP 1 76%	BCGP 2 75%	BCGP 3 65%	CaG 1 62%	CaG 2 77%	OHS 1 86%	OHS 2 79%	OHS 3 66%
ADM_STUDY_ID	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ADM_PSA	92%	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓
ADM_STUDY_DATASET	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ADM_QX_VERSION	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
					✓	✓	✓	✓	✓	✓	✓	×	✓
					✓	✓	✓	✓	✓	✓	✓	✓	✓
					✓	✓	✓	✓	✓	✓	✓	✓	✓
SDC_SEX	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SDC_AGE_CALC	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SDC_BROTHERS_NB	75%	✓	✓	×	✓	✓	✓	✓	×	×	✓	✓	✓
SDC_SISTERS_NB	75%	✓	✓	×	✓	✓	✓	✓	×	×	✓	✓	✓
SDC_SIB_NB	92%	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
SDC_SIB_NB_CAT	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SDC_SIB_OLDER_NB	92%	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
SDC_SIB_OLDER_NB_CAT	100%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

CanPath Data Portal

Variables descriptions

Harmonization Initiative	CORE
Harmonization Protocol	CANPATH-BL-HRFQ-HP
Value type	Integer

Categories

Name	Label	Missing
1	Married and/or living with a partner	
2	Divorced	
3	Widowed	
4	Separated	
5	Single, never married	

Classifications

Additional information

Source
Questionnaire

Target
Participant

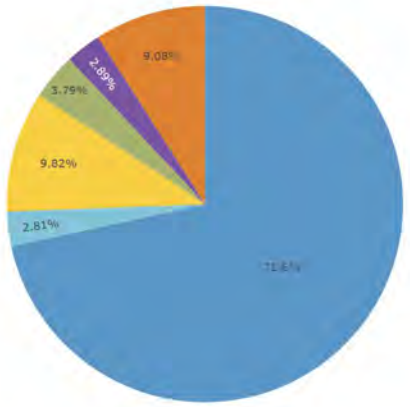
Areas of information

Socio-demographic and economic characteristics
Marital/partner status

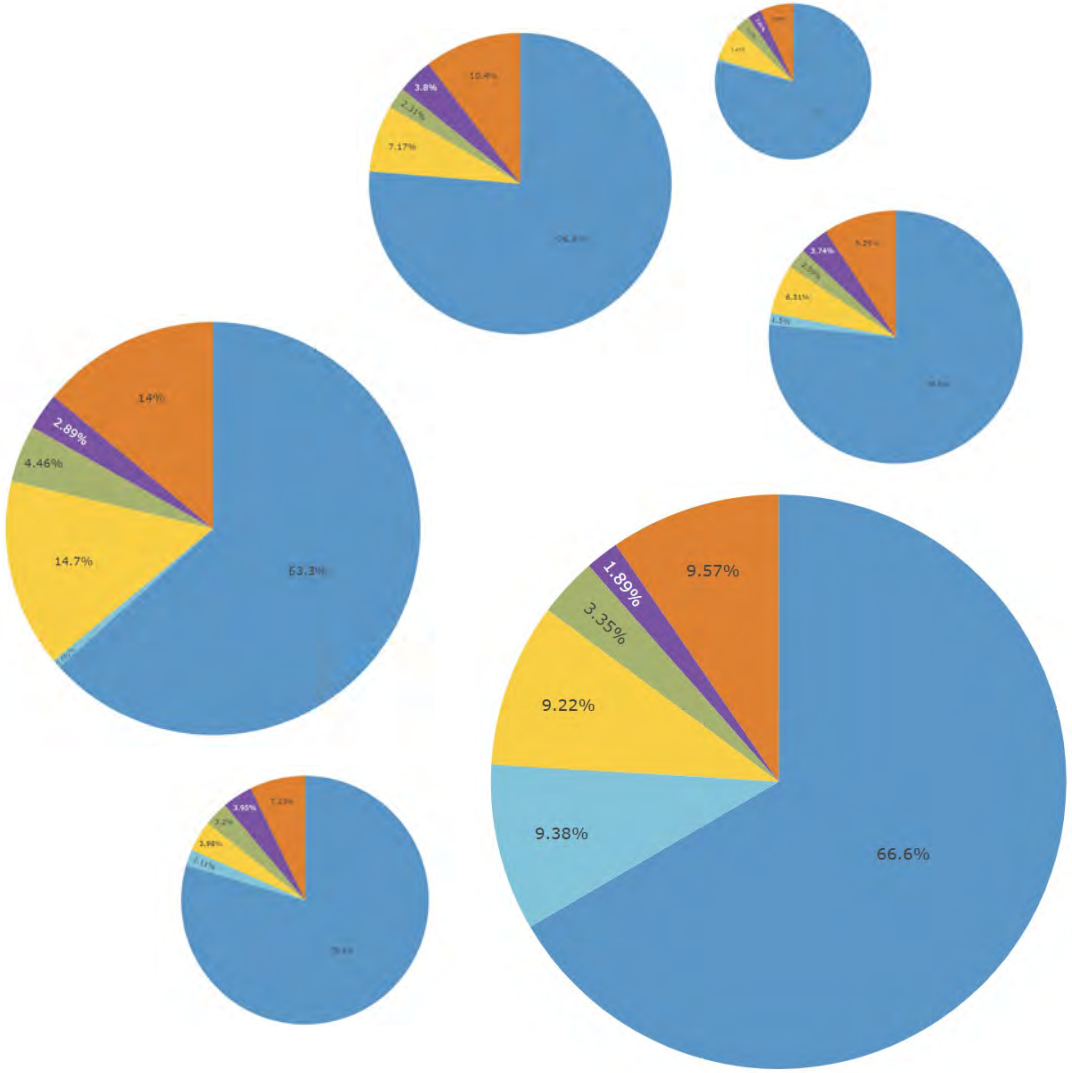
Find similar variables

Participants distributions – total

Values	Frequencies
Valid Values	
1 Married and/or living with a partner	234,602 73.68% *(71.61%)*
2 Divorced	29,759 9.35% *(9.08%)*
3 Widowed	9,481 2.98% *(2.89%)*
4 Separated	12,402 3.9% *(3.79%)*
5 Single, never married	32,164 10.1% *(9.82%)*
Subtotal	318,408 97.19%
Other Values	
Missing	9,222



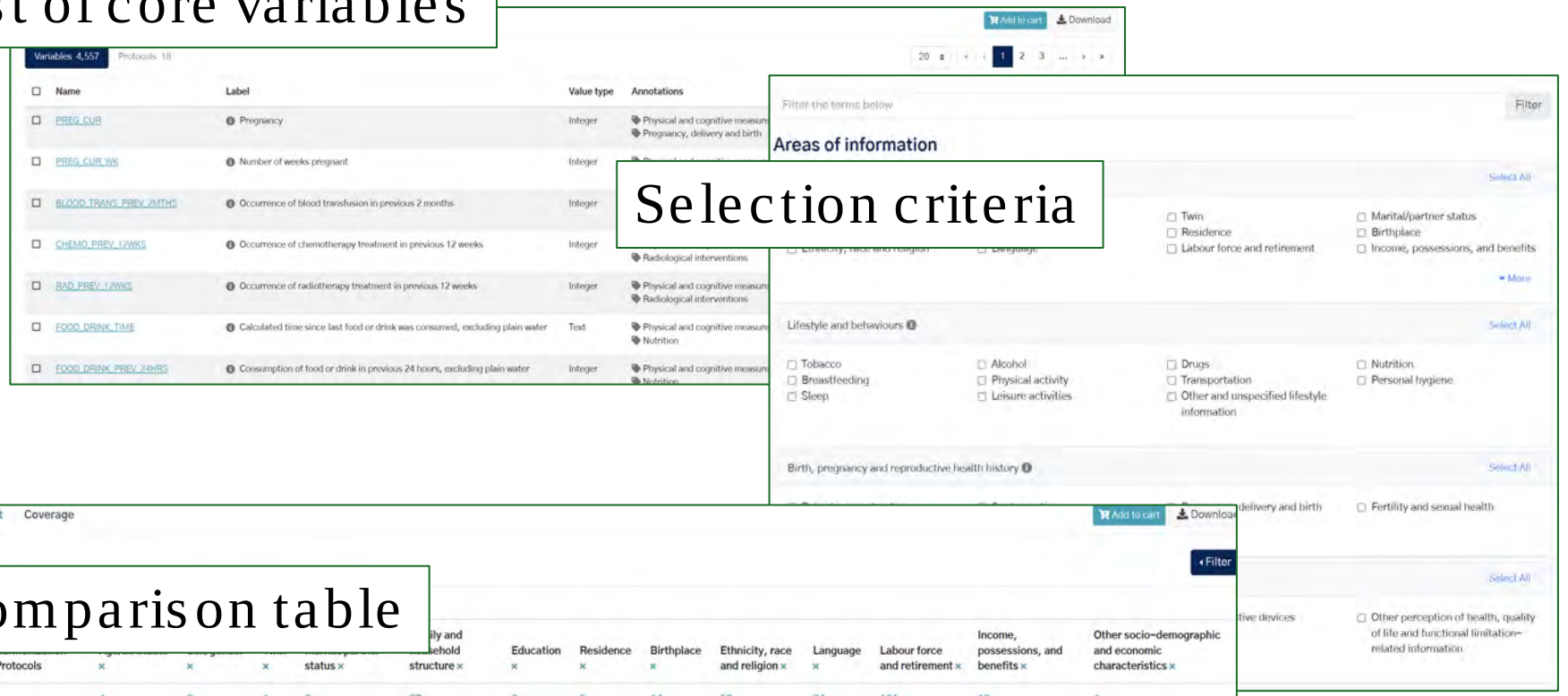
Participants distributions – for each study



CanPath Data Portal

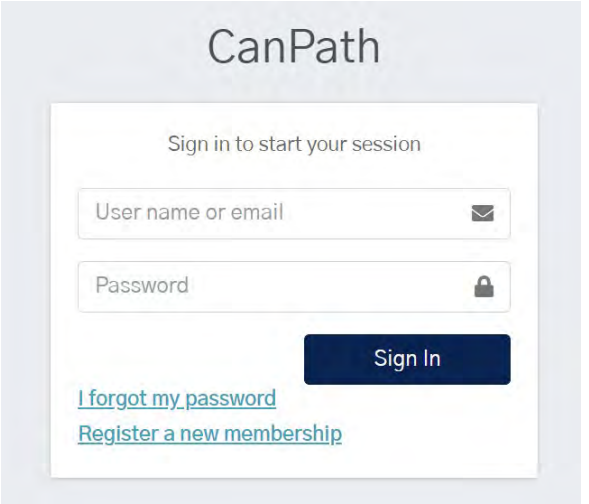
Variables Search

List of core variables



Selection criteria

Comparison table

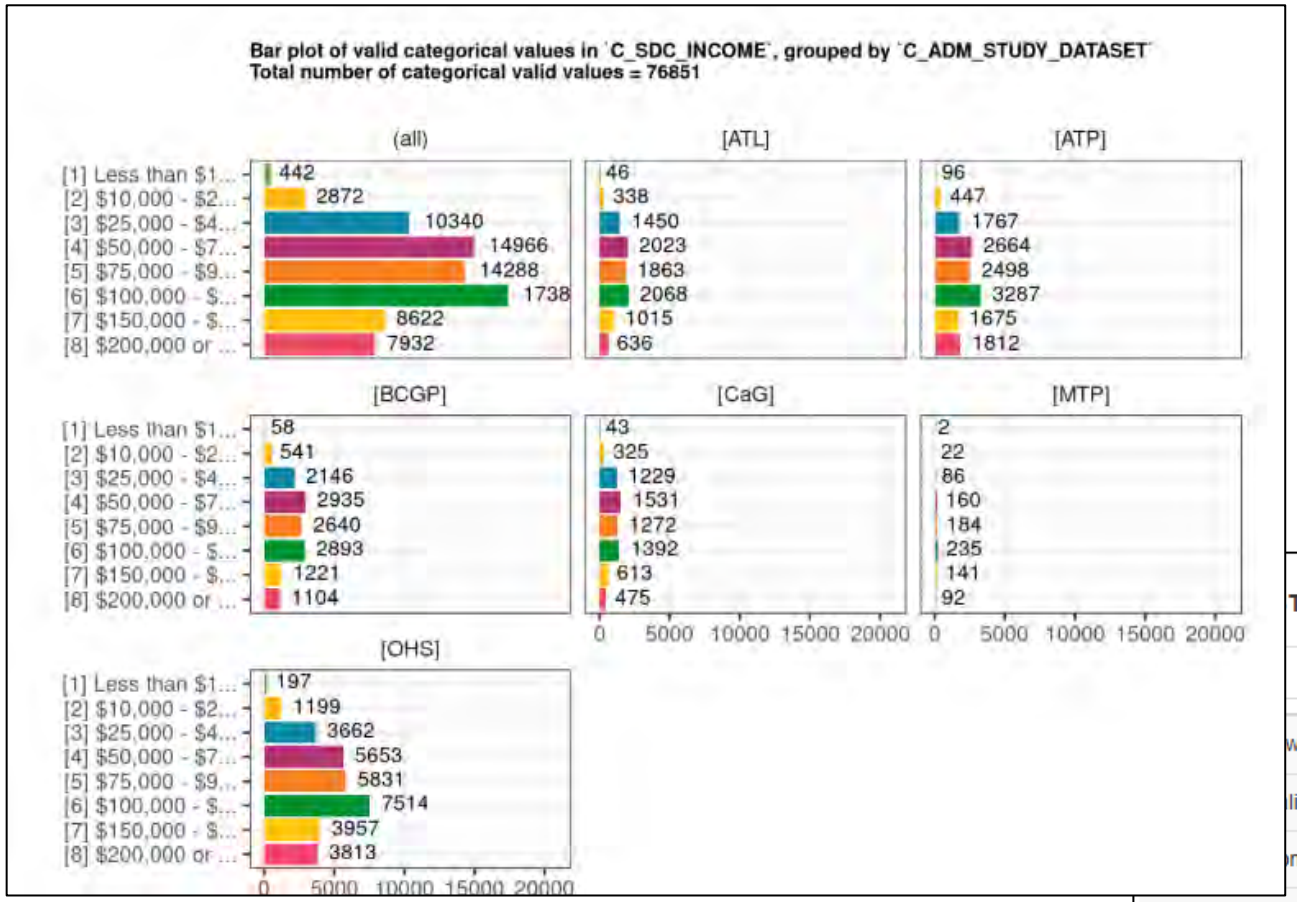


Reports

Data Processing Elements

dataschema_variable	ss_variables	ss_format	ss_source_information	ss_informant	harmo_status	ırmo_status_deta	rule_category	harmo_rule
C2_ADM_PART_ID	index	text	biosamples	participant	complete	identical	id_creation	index
C2_ADM_SPECIMEN_ID	external_id	text	biosamples	participant	complete	identical	direct_mapping	direct_mapping
C2_ADM_SEROLOGY_PART_ID	study_id	text	biosamples	participant	complete	compatible	operation	str_sub(study_id,4)
C2_SAMPLE_SERO_PHASE	__BLANK__	integer	biosamples	participant	complete	identical	paste	
C2_SAMPLE_SPECIMEN_SOURCE	__BLANK__	integer	biosamples	participant	complete	identical	paste	
C2_SAMPLE_LABORATORY_NAME	__BLANK__	text	biosamples	participant	complete	identical	paste	"Sinai Health System-Gingras"
C2_SAMPLE_LABORATORY_ZIPCODE	__BLANK__	text	biosamples	participant	complete	identical	paste	"M5G 1X5"
C2_SAMPLE_OVERALL_DESCRIPTION	voting_result	integer	biosamples	participant	complete	compatible	recode	recode("Negative" = 2L; "Positive" = 1L;ELSE=NA_integer_)
C2_SAMPLE_SUGGESTED_STATUS	suggested_status	integer	biosamples	participant	complete	compatible	case_when	case_when(suggested_status == "Indeterminate" ~ 6L ;suggested_status == "No_antibody response detected" ~ 1L ;suggested_status == "Past_infection" ~ 2L; suggested_status == "Past_infection, may also be vaccinated" ~ 3L ;suggested_status == "Technical Failure (NSQ)" ~ 5L; suggested_status == "Technical failure (NSQ)" ~ 5L;suggested_status == "Technical Failure" ~ 5L;suggested_status == "Vaccinated or past infection" ~ 4L; suggested_status == "Questioned" ~ 6L; ELSE ~ NA_integer_)
C2_CITF_ASSAY_ID								recode("np" = "10010-00"; "smt1" = "10020-00"; "rbd" = "10030-00"; ELSE =

Reports



TATISTICS

	[ATL]	[ATP]	[BCGP]	[CaG]	[MTP]	[OHS]
ws	12781	18171	17662	8125	1114	39766
lid values	9439	14246	13538	6880	922	31826
n-valid values	2041	3622	3223	777	152	6054
Number of empty values	1301	303	901	468	40	1886
% Valid values	73.85	78.4	76.65	84.68	82.76	80.03
% Non-valid values	15.97	19.93	18.25	9.56	13.64	15.22
% Empty values	10.18	1.67	5.1	5.76	3.59	4.74
Number of distinct values	10	10	10	10	10	10

Thanks!

Rita Wissa,
Anouar Nechba,
Guillaume Fabre,
Karla Ordonez,
Tina Wey,
Éva Séa,
Alexandre Trottier,
Jérome Laforme,
Zishu Chen,
Ramin Haeri Azad,
Samuel El-Bouzaidi Tiali,
Chloé Marchand,
Jeanne Sophie Ndiaye

mælstrom

www.maelstrom-research.org

anechba@maelstrom-research.org
isabel.fortier2@mcgill.ca