# Lessons from CanPath on demystifying data harmonization

Isabel Fortier, Anouar Nechba Webinar, June 10<sup>th</sup>, 2025





## Initiatives harmonizing data Different research areas, focuses, sizes, ...



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

Large initiatives generating and offering access to harmonized data

# Covid-19 IMMUNITY

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.



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



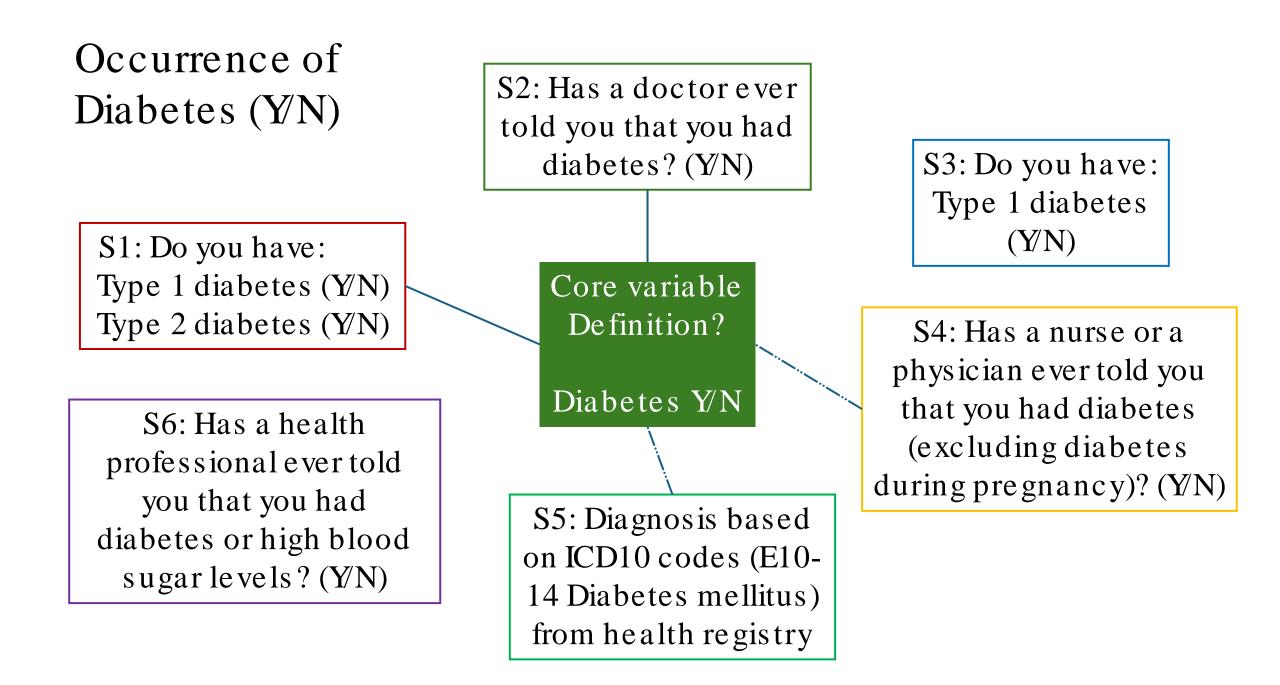
# What is data harmonization?

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









# 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.

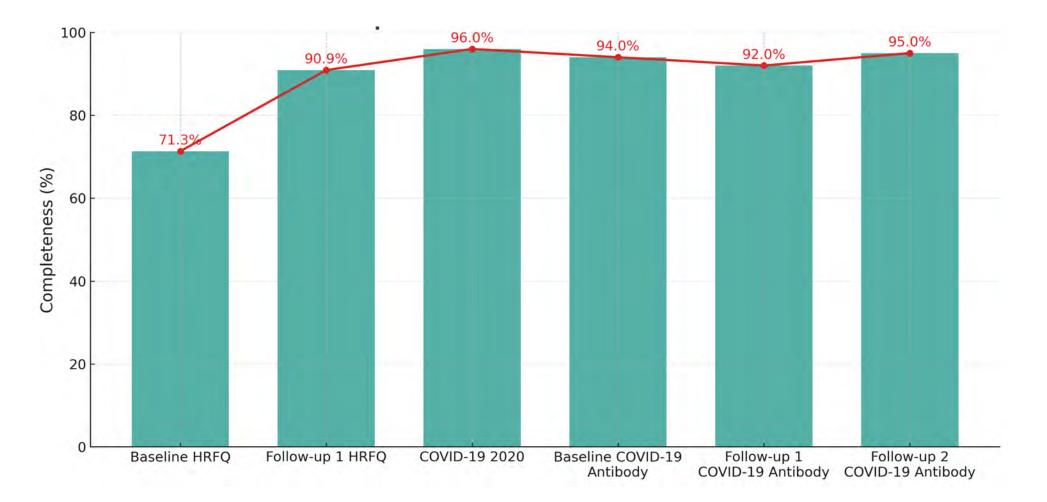






# Harmonization potential





#### www.maelstrom-research.org



- International research platform created in 2012, based on activities initiated in 2004 (P<sup>3</sup>G)
- 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

						psaMark					
	Developmental lealth and Disease	Life course of retrospective harmonization initiatives: key elements to consider									
www.cambrid	ige.org/doh										
Original An	ticle	RESEARCH ARTICLE									
Pinot de Maira A, N	fortier I, Wey TIK, Bergeron J, Rybo-Andersen A-M, Bishop T,	Fost	ering popul	ation-base	ed cohort o	lata					
Enclevort E, Marci Ornelas JP, Sebert Wilson RC, Griffith course of retrospe initiatives: key eler	evic M, Swertz MA, van on Y, Mayhofer MT, 1 S, Santos AC, Rocha A, LE, and Burton P. (2023) Life tothe harmonication ments to consider. Journal of gins of Health and Disease 14:		overy: The N loguing tool		Research						
			International Journal of Epid	Aarcon	<sup>1</sup> , Vincent Ferretti <sup>4</sup> , I	sabel Fortier <sup>1</sup> *					
ISE/	A			Dignal Article Ungelal Article uebec, C	fontreal, Quebec, Canac f Basel, Basel, Switzerla Canada						
riginal Artic	cle			<b>WFA</b>		And a state of the second					
Aaelstro	m Research	n auide	lines for rigoro			Advance Access Publication Data 29 September 201 Dispose antice					
	ctive data h			Constant of							
abel Fortie	1			Data Matters	Sec. 2.						
Stolk, <sup>4</sup> Ba	<b>E</b> A			nel of Epsternoorgy, 2017, 1372–1378 ooi; 10.1085/je/dys140 Pythication: Date: 7 September 2017		lysis to the data,					
eter Grand	LEL/ I		Control rection	Soltware Application Profile	mick Marcon, <sup>2</sup> Julia Is	aeva, <sup>3</sup> Philippe LaFlamme, <sup>2</sup>					
					vy. <sup>5</sup> Marja-Liisa Nuotic						
	Software Applicat				naby Murtagh, <sup>8</sup> Ipek Demir, <sup>9</sup> Dany Doiron, <sup>2</sup> <sup>1</sup> Susan E Wallace, <sup>4</sup> Isabelle Budin-Ljonse, <sup>2</sup> nidt. <sup>11</sup> Paolo Boffetta, <sup>12</sup> Mathies Boniol, <sup>12</sup> V Carter, <sup>13</sup> Nick deklerk, <sup>13</sup> Chris Dibben, <sup>14</sup> <sup>14</sup> Tero Tiekkalinna, <sup>14</sup> Kristian Hveern, <sup>15</sup> <sup>1</sup> Millar, <sup>16</sup> Ivan J Perry, <sup>16</sup> Annette Peters, <sup>17</sup> <sup>1</sup> Millar, <sup>18</sup> Ivan J Perry, <sup>16</sup> Annette Peters, <sup>17</sup>						
			on Profile: Opal								
			are solutions for ta management								
			dissemination		elanie Waldenberger, <sup>1</sup> vel, <sup>20</sup> John Macleod, <sup>1</sup>	Markus Perola, 6.7,18					
	Dany Doit REDO	ap .	C	ollection		r R Harri Augtagh					
	Vincent Ferretti <sup>®</sup>	јх 📃	u	mection	_	· 🕶					
		-		1							
	op	al 📗	Storage,	manager	nent,						
	0	D		nonizatio		Access					
		n –	1	nomzacio	1						
			-	10000							
		100	Analysis	+ Disser	mination	↔					
		R	PataSHIELD	m	ìca	Agate					

# P<sup>3</sup>G-OBiBa-Maelstrom and CanPath Partnership (2008-ongoing)

## Methods, standards...

	Metadata	Requisite				Defin	itions	
RESEARCH IN								
Background		Mandatory			research intiative			
[	Acronym	Optional			research initiative			
	Start year	Mandatory			research initiative			
1	End year	Optional	lf i	annlicable en	vear of the resea	arch initiative		
	Objectives				pecific		Generic standard proposed	
		,		information	peeme			
	Investigators	Optional	Lis			Requisite	Definitions	
	Contact	Optional	Na	Index		Optional	Position of the DS variable in the variable list.	
			ha	Harmonization Table	project	Optional Optional	Acronym of the harmonization project. Name of the table (or dataset) including the DS variables.	
ŀ	Publication(s)	Mandatory if	14	Name		Mandatory	Name of the Tosice (of Dataset) including the US variables.	
	Publication(s)	applicable	12	Label		Mandatory	Short description of the DS variable specifying its content.	
			ua	Classification		Optional	Classification of the DS variable in specific domains.	
	Harmonization	Mandatory	IT		Value type	Mandatory	Value type of DS variable (e.g., boolean, binary, date, datetime, time, decimal, integer, text)	
	protocols		pr		Units	Mandatory (if	For relevant decimal, integer, date, datetime, time DS variables (e.g., height, date of birth), units of the variable	
	Additional	Optional	Su	Format		applicable) Mandatory (if	(e.g., cm, γγγγ-mm-dd) For relevant categorical, integer, boolean and binary DS variables, coding and description of the response options	
	TION PROTOCOL			Format	Categories	applicable)	For relevant categorical, integer, boolean and binary US variables, cooling and description of the response options le e (lacever) Taxometimes 2-antifering Sametre not to asswer()	
Background	Title	Optional	Tit			Mandatory (if		
	Acronym	Optional	Ac		Other	applicable)	for any other type of format, (e.g., postal codes digits, ID), relevent information about the variable format	
	Objectives	Optional	De	Target	Individual/entity	Mandatory	Individual (or entity) the variable is about (e.g., participant, sample, postal code).	
	Information cont		0		Time period	Mandatory	Time period the variable refers to (or is about) (e.g., first trimester of pregnancy, last 30 days, at birth).	
		1	pr	4	Source	Mandatory	Acceptable source of information (e.g., questionnaire, governmental database/registrie, etc.) from which the variable data could be extracted	
ł	Exploratory/actu	al Optional	0	1		Mandatory (if		
	harmonization		die	1	Informant	applicable)	For information that can be provided by different persons, acceptable informant (e.g. participant, partner, proxy)	
	indimization.		da.		Time of collection	Mandatory	Acceptable time of collection of the information (e.g., smoking status during first trimester of pregnancy can be	
ŀ	e	Optional	u a	Requirements	Time of collection		collected at birth(or not)).	
	Ex-post/ ex-ante	Optional	100	Kequirements	Wording	Mandatory (if	For data extracted from questionnaires, acceptable question wording (e.g., binge drinking defined as 5 drinks or	
	approach		ae	-		applicable)	more).	
			co	1	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	Optional	nal De			Mandatory (if	For data that can be generated using different procedures, acceptable procedures (e.g., weight needs to be	
			re		Procedures	applicable)	measured, not reported by the participant).	
	Infrastructure	Optional	De			Mandatory (if	For variables to be generated using specific standard or instructions description of the conditions to be respected	
			ho	Instructions		applicable)	(e.g. ICD-10 must be used, prematurity defined as birth before 37 weeks of pregnancy according to a given	
ſ	Additional	Optional	Su	-			reference, specific information related to management of unique or multiple choices option, etc).	
Participatin	Study/database	Mandatory	Cri	Comments		Optional	Any additional information not covered	
g	selection		up	s/collection ti	me points, countr	y/territory, de	esign, year of recruitment/collection)	
studies/dat	Study/database li	ist Mandatory						
			LIS		eristics (e.g. study	design, partic	pant sampling frame, nb of participants, etc)	
abases		ist iniaridatory						
abases DataSchem	Identification		of	the individual			the harmonization project.	
		Inform	of	the individual	studies/database		the harmonization project. Generic standard proposed	
DataSchem a	Identification Number of var	Inform	of	the individual	studies/database	is included in	the harmonization project. Generic standard proposed Definitions	
DataSchem a	Identification Number of var Variable definit	Inform	of atic	the individual	studies/database Requisite Optional	es included in	the harmonization project. Generic standard proposed Definitions to Svariable lint.	
DataSchem a	Identification Number of var Variable definit	Inform dex tudy-specific DataSch	of atic	the individual	studies/database Requisite Optional Optional	Position of the	the harmonization project.  Generic standard proposed  Definitions  he Sivariable in the variable list.  harmonized dataset including the D5 variables.	
DataSchem a	Identification Number of var Variable definit Variables inclus p	Inform Idex Iudy-specific DataSch ataSchema variable	of atic	the individual	studies/database Requisite Optional Optional Mandatory	Position of the Name of the	the harmonization project.  Generic standard proposed Definitions ac D5 variable in the variable list.  Tarmonized dataset including the D5 variables. DataSchema variable.	
DataSchem a	Identification Number of var Variable definit Variables inclu Variable list	Inform idex iudy-specific DataSch ataSchema variable cronym of the study	of atic	the individual on a table	studies/database  Requisite  Optional  Optional  Mandatory  Mandatory	Position of the Name of the Name of the Acronym of t	the harmonization project.  Generic standard proposed  Definitions  be Svariable in the variable list.  harmonized dataset including the DS variables.  DataSchema variable.  be study.	
DataSchem a Study/data	Identification Number of var Variable definit Variables inclus Variable list Datasets to be N	Inform idex tudy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe	of atio	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory	Position of th Name of the Name of the Acronym of th Name of the	the harmonization project.  Generic standard proposed Definitions harmonized dataset including the DS variables. Desixbothen available list. harmonized dataset after preprocess of cleaning and ready to be used to generate the DS variables.	
DataSchem a Study/data base	Identification Number of var Variable definit Variables inclus Variable list Datasets to be Participants inc	Inform udy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe ame of the study-spe	of atio	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory	Position of th Name of the Name of the Acronym of t Name of the Name of the	the harmonization project.  Generic standard proposed Definitions before the variable list. before a variable. be study. Study specific variable. be study. Study specific variable job used to generate the DS variables. Study specific variable job used to generate the DS variable.	
DataSchem a Study/data base datasets	Identification Number of var Variable definit Variables inclu Variable list Datasets to be Naticipants inc Fi	Inform idex tudy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe	of atio	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory	Position of th Name of the Name of the Name of the Name of the Any informal	the harmonization project.  Generic standard proposed Definitions te D5 variable int. barmonized dataset including the D5 variables. Destaching variables. be study-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. study-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. study-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. study-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. study-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. study-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables.	
DataSchem a Study/data base datasets	Identification Number of var Variable definit Uariables inclus Variable list Datasets to be Participants inc Variables	Inform udy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe ame of the study-spe	of atio	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory	Position of th Name of the Name of the Acronym of th Name of the Any informat Value type :t	the harmonization project.  Generic standard proposed  Definitions  to Svariable in the variable list.  harmonized dataset induding the DS variables. DataSchema variable.  he study. study-apedic variable (a): study-apedic variable (b): s	
DataSchem a Study/data base datasets	Identification Number of var Variable definit Variables inclus Variables inclus Datasets to be Participants inc Variables acceptability	Inform udy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe ame of the study-spe	of atio	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory	Position of th Name of the Name of the Acronym of th Name of the Any informat Value type :t Units : for rel	the harmonization project.  Generic standard proposed Definitions te D5 variable in the variable list. DataSchema variable. Stands-generic dataset including the D5 variables. Stands-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. study-specific variable(s) to be used to generate the D5 variable. Stands-generic dataset after preprocess of cleaning and ready to be used to generate the D5 variables. Study-specific variable(s) (e.g., boolean, binary, date, datetime, time, decimal, integer, text) send decimal, integer and ter variable (e.g., boolean, binary, date, datetime, time, decimal, integer, text)	
DataSchem a Study/data base datasets	Identification Number of var Variable definit Variables inclut Variables inclut Datasets to be Participants inc Fr Variables acceptability Descriptive sta	Inform dex tudy-specific DataSch ataSchema variable cronym of the study ame of the study-spe ame of the study-spe ormat	of atio	the individual on a table : dataset	studies/database Requisite Optional Mandatory Mandatory Mandatory Optional	Position of th Name of the Name of the Acronym of th Name of the Name of the Any informat Value type :t Units : for rel Categories : of	the harmonization project.  Generic standard proposed Definitions be 5% variable in the variable list. harmonized dataset including the DS variables. DataSchema variable. he study. study-specific variable list be used to generate the DS variable. Study-specific variable[10 be used to generate the DS variable. Study-s	1-17
DataSchem a Study/data base datasets Processing	Identification Number of var Variable definit Datasets to be Participants ind Variable list Fi Variables acceptability Descriptive sta Harronization	Inform. idex tudy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe ame of the study-spe armat ime period	of atio	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional	Position of th Name of the Acronym of th Name of the Name of the Name of the Categories : Time period	the harmonization project.  Generic standard proposed Definitions bit Standard proposed Definitions bit Standard proposed Definitions bit Standard proposed Definitions bit Standard proposed bit Standard proposed bit Standard proposed Definitions bit Standard proposed bit Standard proposed bit Standard proposed Definitions bit Standard proposed bit Standard p	
DataSchem a Study/data base datasets Processing	Identification Number of var Variable defini Variables inclu Variable list Datasets to be Narticipants inc Fi Variables acceptability Descriptive sta Harmonization Harmonization	Inform dex udy-specific DataSch ataSchema variable conym of the study-spe arme of the study-spe arme of the study-spe armat ime period me of collection	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional	Position of th Name of the Name of the Acronym of t Name of the Name of the Name of the Categories : C Time period Time period	the harmonization project.  Generic standard proposed Definitions  be 5% variable in the variable list. harmonized dataset including the DS variables. DataSchema variable. he study. study-specific variable (j) be used to generate the DS variables. University of the study-specific variable (j) be used to generate the DS variable. to related to the format of the study-specific variable(j) be used to generate the DS variable. to related to the format of the study-specific variable(j) be used to generate the DS variable. to related to the format of the study-specific variable(j) be used to generate the DS variable. to related to the format of the study-specific variable(j) be used to generate the DS variable. to related to the format of the study-specific variable(j) be used to generate the DS variable. targeted by the study-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable(j) fee used to generate the DS variable. targeted by the study-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable(j) fee used to generate the DS variable. targeted by the study-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable( <i>s</i> = heiht date of hirth) units of the ctude-specific variable( <i>s</i> = heither date of hirth) units of the ctude-specific date of hirth) units of the ctude-specific date of hirth) units of the ctude-specific variable( <i>s</i> = heith date of hirth) units of the ctude-specific date of hirth) units of the ctude-specific date of hirth) units of the ctude-specific date of hirth) units of heither date	075
DataSchem a Study/data base datasets Processing	Identification Number of var Variable definit Variable inclu Variable int Participants inc Variables acceptability Descriptive sta Harmonization T Processing rule S	Inform dex udy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe ame of the study-spe prmat me period me of collection purce of information	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional Optional	Position of th Name of the Name of the Acronym of t Name of the Any informat Value type : t Units : for rel Categories : c Time period Acceptable s	the harmonization project.  Generic standard proposed Definitions  b Svariable in the variable list.  Thermonized dataset including the DS variables. DataSchmen variable. DataSc	075
DataSchem a Study/data base datasets Processing elements	Identification Number of var Variable defini in Variables inclus Variables inclus Datasets to be Participants in N Participants in N Parti	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	Position of th Name of the Acronym of the Acronym of the Ary informat Value type: t Units : for rel Categories : c Time period Time of colle Acceptable s For informat	the harmonization project.  Generic standard proposed Definitions to Svariable in the variable list. harmonized dataset including the DS variables. DataSchema variable. he study. study-specific variable[] to be used to generate the DS variables. Used variable list is the study-specific variable[] to be used to generate the DS variable. to related to the format of the study-specific variable[] to be used to generate the DS variable. to related to the format of the study-specific variable[] to be used to generate the DS variable. to related to the format of the study-specific variable[] to be used to generate the DS variable. to related to the format of the study-specific variable[] to be used to generate the DS variable. to related to the format of the study-specific variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. to methate the period variable[] to be used to generate the DS variable. targeted by the study-specific variable[] to be used to generate the DS variable. target by the study-specific variable[] to be used to generate the DS variable. target by the study-specific variable[] to be used to generate the DS variable. target by the study-specific variable[] to be used to generate the DS variable. target by the study-specific variable[] to be used to generate the DS variable. target by the study-specific variable[] to be used to generate the DS variable. target by the study-specific variable[] to be used to generate the DS variable. Target by the study-specific variable[] to be used to generate the DS variable. Target by t	075
DataSchem a Study/data base datasets Processing elements	Identification Number of var Variable defini in Variables inclus Variables inclus Datasets to be Participants in N Participants in N Parti	Inform dex udy-specific DataSch ataSchema variable cronym of the study-spe ame of the study-spe ame of the study-spe prmat me period me of collection purce of information	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional Optional	Position of th Name of the Name of the Name of the Acronym of t Name of the Any informat Time periodi Time pe	the harmonization project.  Generic standard proposed Definitions  Definitions  Definitions  DataSchman Variable  Buddy, Standard, Strandard, S	075
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	Identification Number of var Variable definin Variable list Ariable list Participants ind Participants ind P	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	s included in Position of the Name of the Name of the Acronym of the Ary informat Value type : Units : for rel Categories : Time period Time of colle Acceptable s for informat Any informat Any informat	the harmonization project.  Generic standard proposed Definitions te 05 variables in the variable list. barmonized dataset including the DS variables. Distribution and the sub-raped control of the	4075 5-34
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	Identification Number of var Variable definit Variables inclus Datasets to be N Participants in N Variables acceptability Descriptive sta Harmonization T Processing rule S Datasets (s) Sample size	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	Position of th Name of the Name of the Acroym of th Aronym of the Any informat Value type : t Units : for rel Categories : c Time period Time period Time of colle Acceptable s For informat Any informat Mording, acc Measures: ac	the harmonization project.  Generic standard proposed Definitions  e D5 variable in the variable list. Inamonized dataset including the D5 variables. DataSchema variable. Be study. Study-specific variable(s) to be used to generate the D5 variables. Study-specific variable(s) to be used to generate the D5 variables. Study-specific variable(s) to be used to generate the D5 variables. Study-specific variable(s) to be used to generate the D5 variables. Study-specific variable(s) to be used to generate the D5 variables. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable. Study-specific variable(s) to be used to generate the D5 variable (s) to be used to generate the D5 variable (s) to be used to generate the D5 variable (s) to be used to generate the D5 variable (s) to be used to generate the D5 variable (s) to be used (s)	071
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	Identification Number of var Variable definit Variables inclus Datasets to be N Participants in N Variables acceptability Descriptive sta Harmonization T Processing rule S Datasets (s) Sample size	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	s included in Position of th Name of the Acronym of th Name of the Aronym of that Arony nof that Any informat Value type :t Units : for rel Categoris : 2 Time period Time of colle Acceptable s For informat Any informat Any informat Any informat Any informat Any informat Categoris : 2 For informat Any informat Any informat Any informat Categoris : 2 For informat Any i	the harmonization project.  Generic standard proposed Definitions te D5 variable in the variable list. Tharmonized dataset including the D5 variables. Subdy-specific dataset after preprocess of cleaning and ready to be used to generate the D5 variables. Subdy-specific dataset after preprocess of cleaning and ready to be used to generate the D5 variables. Subdy-specific variable(s) to be used to generate the D5 variable. Subdy-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables (e.g. built date of birth) units of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables (e.g. built date of birth) units of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables (e.g. built date of birth) units of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables. Subdy-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, i	4075 5Ge
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	Identification Number of var Variable definit Variables inclus Datasets to be N Participants in N Variables acceptability Descriptive sta Harmonization T Processing rule S Datasets (s) Sample size	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	s included in Position of th Name of the Acronym of th Name of the Aronym of that Arony nof that Any informat Value type :t Units : for rel Categoris : 2 Time period Time of colle Acceptable s For informat Any informat Any informat Any informat Any informat Any informat Categoris : 2 For informat Any informat Any informat Any informat Categoris : 2 For informat Any i	the harmonization project.  Generic standard proposed Definitions te D5 variable in the variable list. Tharmonized dataset including the D5 variables. Subdy-specific dataset after preprocess of cleaning and ready to be used to generate the D5 variables. Subdy-specific dataset after preprocess of cleaning and ready to be used to generate the D5 variables. Subdy-specific variable(s) to be used to generate the D5 variable. Subdy-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables (e.g. built date of birth) units of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables (e.g. built date of birth) units of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables (e.g. built date of birth) units of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer and ter variables. Subdy-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, integer, text) send decimal, integer of the study-specific variable(s) (e.g. boolean, binany, date, datetime, time, decimal, i	e075 Side
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	Identification Number of var Variable definit Variables inclus Datasets to be N Participants in N Variables acceptability Descriptive sta Harmonization T Processing rule S Datasets (s) Sample size	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	Position of th Name of the Name of the Name of the Acronym of the Name of the Name of the Name of the Name of the Name of the Categories : Time period Time of colle Acceptable s For informat Wording: acc Measures: at 7, PHC-9, me Procedues :	the harmonization project.	4075 5Ge
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	Identification Number of var Variable definit Variables inclus Datasets to be N Participants in N Variables acceptability Descriptive sta Harmonization T Processing rule S Datasets (s) Sample size	Inform. dex udy-specific DataSch atsSchema variable crowym of the study-spe ame of the study-spe armat ime period ime of collection purce of information formant	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional	Position of th Name of the Name of the Name of the Acronym of the Name of the Name of the Name of the Name of the Name of the Categories : Time period Time of colle Acceptable s For informat Wording: acc Measures: at 7, PHC-9, me Procedues :	the harmonization project.	4075 5Ge
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable definit it Variable ist definit it Variable ist definit Datasets to be n Participants inde Participants inde Participa	Inform dex adv_specific Oatsch advocution of the study spe ame of collection aurce of information formant spwords	of atic	the individual on a table : dataset	studies/database Requisite Optional Optional Optional Mandatory Mandatory Mandatory Mandatory Optional	Position of th Name of the Name of the Categories : for rel Categories : tor Time period: Time period: Time of colle Any informat Wording: acc Measures: ac 7. JPHC 9, me Proceeding: acc Instructions: The of the the the the the the the Proceeding of the the the the the Proceeding of the the the the the the Proceeding of the the the the the the the the Proceeding of the	the harmonization project.  Generic standard proposed Definitions  backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions backstandard proposed Definitions Definitions backstandard proposed Definitions Def	4075 5Ge
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable indenting Variable indenting Variable indention Participants mit Participants mit Sample size Descriptive sta	Inform dex udy-specific DataSch ataSchema variable crown of the study-spe ame of the study-spe armat me of collection surce of information formant epwords ternal Comment	of atic	the individual on a table : dataset : variable(s)	studies/database Requisite Optional Opt	s included in Position of th Name of the Name of the Acronym of In Name of the Acronym of In Name of the Acronym of In Name of the Acronym of In Categories : c Time of collex Acceptable s For informat Wording: acc Procedures : Procedures : prematurity internation protection in the Internation Procedures : prematurity internation	the harmonization project.  Generic standard proposed Definitions  be 55 variable in the variable list.  brannoized dataset inter preprocess of cleaning and ready to be used to generate the DS variables. DataSchema variable.  be datay.  study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  to study-specific variable() to be used to generate the DS variable.  To study specific variable() to be used to generate the DS variable.  To study-specific variable() to be used to study-specific variable() to generate variable() to	e075 Scile
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable indefining Variables indefining Variables indefining Participants into R Participants in	Inform. dex udy specific DataSch atsSchema variable commo of the study. ame of the study spe ame of the study spe ame of the study spe ame of collection user collection user collection formant sywords ternal Comment armonization comm	of atio	the individual on a table : dataset : variable(s)	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Mandatory Optional Optional Optional Optional	Position of the Name of the Acronym of I Name of the Acronym of I Name of the Acronym of I Categories: I Time of colle Acceptables 5 For informat Wording: acc Measures: ac Procedures: prematurity prematurity instructions: prematurity instructions:	the harmonization project.  Generic standard proposed Definitions  te D5 variable in the variable list. harmonized dataset including the D5 variables. Suddy-specific dataset after proposeds of cleaning and ready to be used to generate the D5 variables. Study-specific variable(s) to used to generate the D5 variable. Some of the study-specific variable(s) (e., bolden, blanny, date, dateline, linteger, text) send decimal, integer and ter variables. Some of the study-specific variable(s) (e., bolden, blanny, date, dateline, linteger, text) send decimal, integer and ter variables (e. > built date of blinh) units of the study-specific variable(s) send decimal, integer, text) send decimal, integer and ter variables. Some of the study-specific variable(s) (e., bolden, blanny, date, dateline, linteger, text) send decimal, integer and ter variables (e. > built date of blinh) units of the study-specific variable(s) (e., bolden, blanny, date, dateline, linteger, text) send decimal, integer and ter variables (e. > built date of blinh) units of the study-specific variable(s) (e., bolden, blanny, date, dateline, linteger, text) send decimal, integer, text) send decimal, integer, text) send decimal, integer and ter variables. Some of the study-specific variable(s) (e., and the study-specific variable(s) (e., and the study-specific variable(s) (e., and the study-specific variables) (for the study-specific vari	e075 Scile
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable finden in Variable finden in Variable finden of Pariables to be Pariticipants in the Pariticipants inter Pariticipants inter Pariticipants inter Processing rule in Descriptive sta Based of the Sample size bescriptive sta	Inform dex udy-specific DataSch ataSchema variable crown of the study-spe ame of the study-spe ame of the study-spe ormat me period me of collection succe of information formant eyeords ternal Comment armonization comment armonization status	of atic	the individual on stable dataset variable(s)	studies/database Requisite Optional Opt	Position of the Name of the Acronym of I Name of the Acronym of I Name of the Acronym of I Categories: I Time of colle Acceptables 5 For informat Wording: acc Measures: ac Procedures: prematurity prematurity instructions: prematurity instructions:	the harmonization project.  Generic standard proposed Definitions  te D5 variable in the variable list. harmonized dataset including the D5 variables. Study-specific dataset after proposes of cleaning and ready to be used to generate the D5 variables. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, datelin	e075 Scile
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable identification Variables included Variables include Participants in N Participants	Inform dex udy specific DataSch ataSchema variable commo of the study ame of the study spe ame of the study spe ame of the study spe ame of collection user collection user collection formant sywords ternal Comment armonization comma armonization status	of atic	the individual on stable dataset variable(s)	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional Optional Optional Mandatory Mandatory Mandatory	estino of the Position of the Name of the Nam of the Name of the Name of the Name of the Name of the	the harmonization project.  Generic standard proposed Definitions  Barbonized dataset including the 53 variables. DataSchmen variable. Be dataset including the 53 variables. DataSchmen variable. Be dataset including the 53 variables. DataSchmen variable. Be dataset for proposed of the study-specific variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed of the study-specific variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed of the study-specific variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. DataSchmen variable. DataSchmen variable(i) be used to generate the DS variables. DataSchmen variable. DataSchmen v	e075 Scle
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable identification Variables included Variables include Participants in N Participants	Inform dex udy-specific DataSch ataSchema variable crown of the study-spe ame of the study-spe ame of the study-spe ormat me period me of collection succe of information formant eyeords ternal Comment armonization comment armonization status	of atic	the individual on stable dataset variable(s)	studics/database  Requisite  Optional  Annotatory  Mandatory  Mandatory  Optional  Opt	estino of the Position of the Name of the Nam of the Name of the Name of the Name of the Name of the	the harmonization project.  Generic standard proposed Definitions  te D5 variable in the variable list. harmonized dataset including the D5 variables. Study-specific dataset after proposes of cleaning and ready to be used to generate the D5 variables. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) to used to generate the D5 variable. Study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, decimal, integer, text) study-specific variable(s) (e., bolieta, blanny, date, dateline, time, datelin	e075 Scle
DataSchem a Study/data base datasets Processing elements Harmonized dataset(s)	identification Number of var Variable indentification Variables indentification Variables indentification Variables indentification Participants in M Participants in M Partic	Inform dex udy specific DataSch ataSchema variable commo of the study ame of the study spe ame of the study spe ame of the study spe ame of collection user collection user collection formant sywords ternal Comment armonization comma armonization status	of atic	the individual on stable dataset variable(s)	studies/database Requisite Optional Optional Mandatory Mandatory Mandatory Optional Optional Optional Optional Optional Optional Mandatory Mandatory Mandatory	s included in Position of th Name of the Name of the Acronym of th Anne of the Anny informat Yulae type : Units for rei Categories : Time periodil Acceptable s For informat Any informat Workings, and Any informat Workings, and Proceedures : reported by in Instructions; International Any informat Markings, and Proceedures : Instructions; Instructions; International	the harmonization project.  Generic standard proposed Definitions  Barbonized dataset including the 53 variables. DataSchmen variable. Be dataset including the 53 variables. DataSchmen variable. Be dataset including the 53 variables. DataSchmen variable. Be dataset for proposed of the study-specific variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed of the study-specific variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed of the study-specific variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. Be dataset for proposed variable(i) be used to generate the DS variables. DataSchmen variable. DataSchmen variable. DataSchmen variable(i) be used to generate the DS variables. DataSchmen variable. DataSchmen v	e075 Scile

Software

vare **Spal** 

To facilitate data storage and management

mælstr



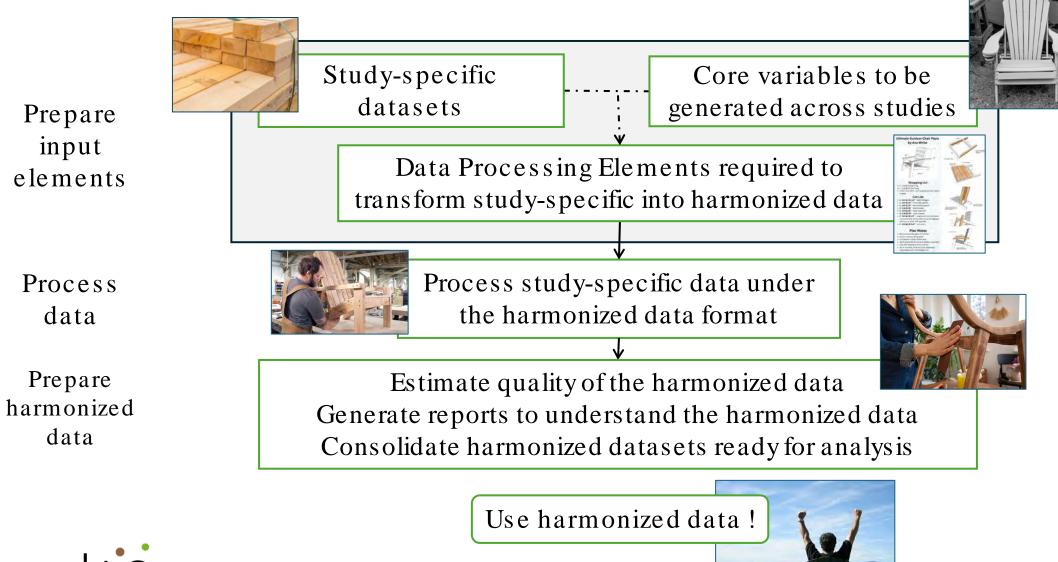
To build online catalogues describing study-specific and harmonized data content, and managing demands for access to data

Rpackages



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

# Harmonization process





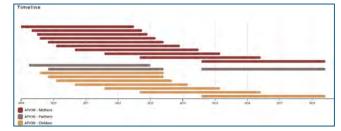


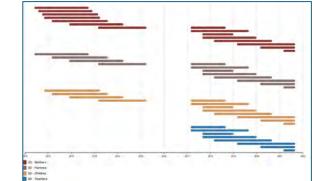
# Study-specific datasets

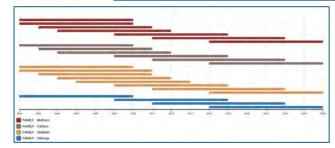
Participants selection criteria? Number of participants? Number/timing of follow ups? Specific information collected? Quality of the data? ....

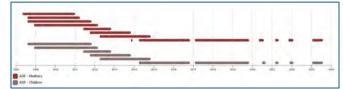
Required to :

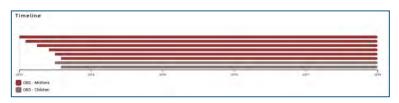
- Understand studies characteristics (e.g., exclusions criteria)
- Transform data and data dictionaries into appropriate formats
- Explore data content and quality (e.g., identify outliers, duplicates)







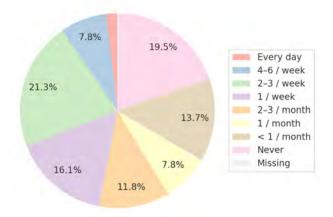




#### Study 1

Alcohol intake during year before pregnancy

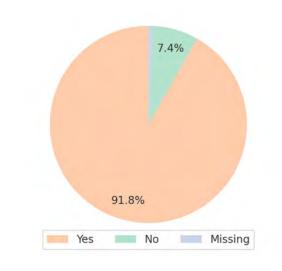
Category	Count
Every day	43
4–6 / week	185
2–3 / week	503
1/week	380
2–3 / month	278
1 / month	185
<1 / month	325
Never	461
Missing	5
Total	2 365



#### Study 2

#### Alcohol use 12 months before pregnancy

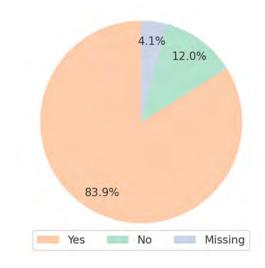
Category	Count
Yes	2 728
No	221
Missing	23
Total	2 972



#### Study 3

#### Ever consumed alcohol

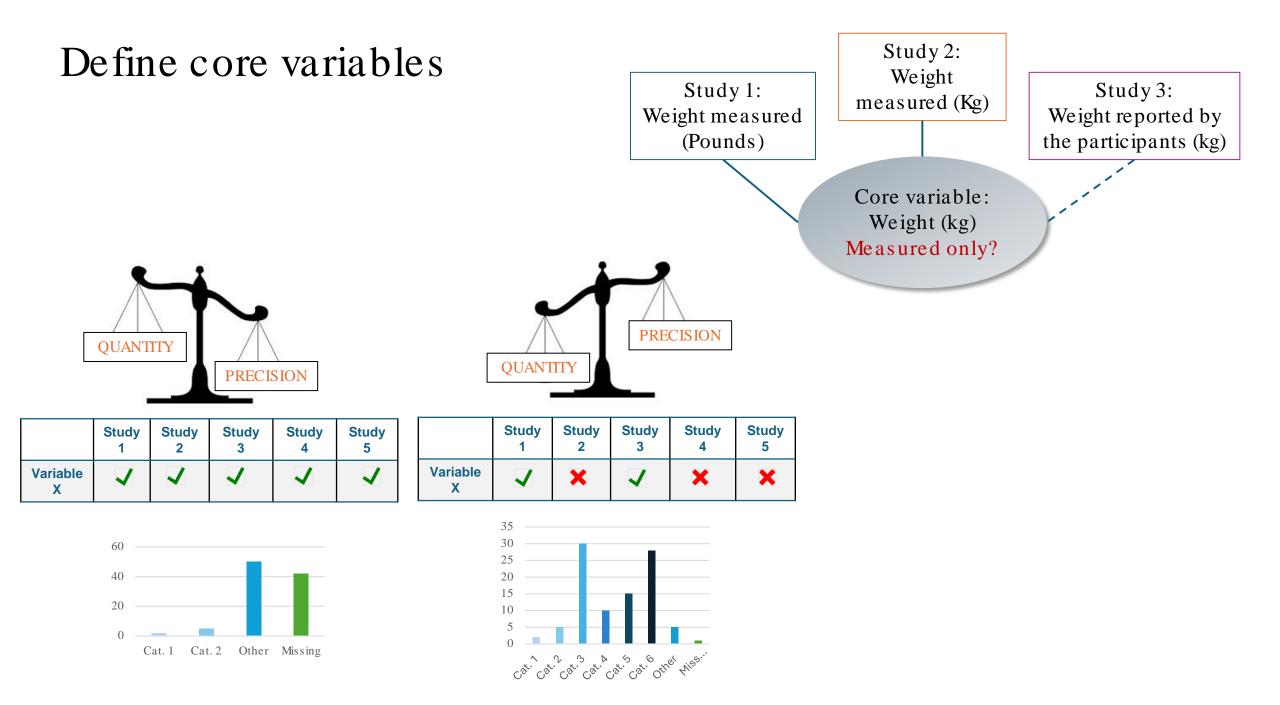
Category	Count
Yes	1 835
No	262
Missing	90
Total	2 187





# Core Variables - exemples

Alcohol consumption of the mother 1 year prior to pregnancy	Categories: <b>0= No ; 1= Yes</b>
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	???



# 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

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

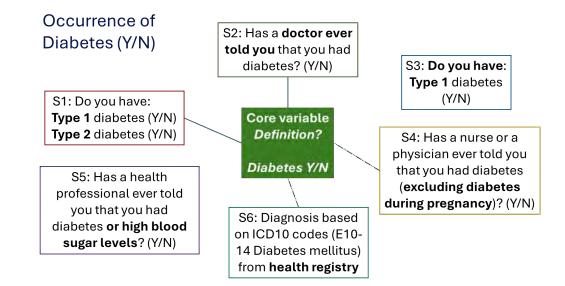
Bran	ALL OUR FAMILIES	APrÖN	FAMILL	ONTARIO
# of mothers: 2,456	# of mothers: 3,387	# of mothers: 2,189	# of mothers: 857	# of mothers: 1,374
Province: QC	Province: AB	Province: AB	Province: ON	Province: ON
Start year: 2010	Start year: 2008	Start year: 2009	Start year: 2004	Start year: 2013

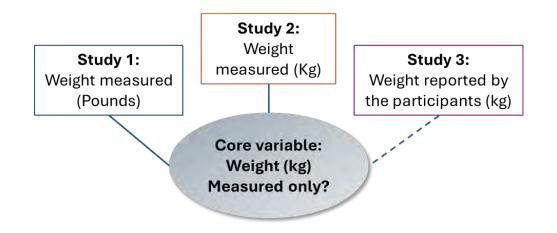
FINAL DISTINCT VARIABLES?
FIRST NATIONS
BLACK
LATIN AMERICAN/HISPANIC
SOUTH ASIAN
WHITE/CAUCASIAN
OTHER



# 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 2

Study 3

#### Study 1

Alcohol intake d	uring year befor	e pregnancy	Alcoh	oluse 12 mon	ths before preg	gnancy	Everconsum	ed alcohol	
Categor	y Count			Category	Count		Category	Count	
Every da	y 43			Yes	2 728		Yes	1 835	
4-6 / wee	ek 185			No	221		No	262	
2–3 / we	ek 503			Missing	23		Missing	90	
1 / week	380			Total	2 972		Total	2 187	
2–3 / mo	nth 278			10101					
1 / mont	h 185								
< 1 / mor	nth 325								
Never	461								
Missing	5								
Total	2 365	]							
•	ecode("Every day" = 1; "4-6 / week" $1 \cdot 2 - 3 / week" = 1 \cdot 1 / week" = 1$		Direct mapping from study variable				Impossible		
"2–3 / month" =	=1; "2-3 / week"=1; "1 / week"=1; "2-3 / month"=1; "1 / month"=1; " <1 / month "=1 " Never"=0; ELSE~NA)						consumption	n on alcohol not collected regnancy	

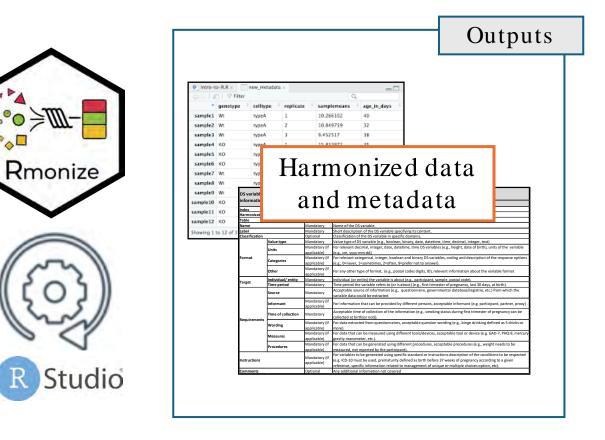
Process study-specific into harmonized data

# Validate and consolidate harmonized data



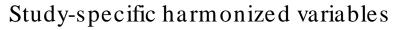
Inputs										
F	mp web									
S	tudy-spec	ific	Category	Count						
	Jan		Everyday	43						
	Category	Count			4–6 / week	185				
	Yes	2 7 2 8			2–3 / week	503				
	No		22	1	1/week	380				
	Missing	23			2–3 / month	278				
	Total	2 972			1 / month	185				
	Category		Count		<1/month	325				
	Yes	1 835		5						
	No	262		2	Never	461				
	Missing	90					Missing	5		
	Total	2 187			Total	2 365				

DS var inform Index Harmor Table Name Label Classific	••••		variables nitions	decimal integer, ted)	equisite nal	Data processing
Format	Categories	Mandatory (if applicable)	For relevant categorical, integer, boolean and binary DS variables, i (e.g., 0=never, 1=sometimes, 2=often, 9=prefer not to answer).	coding and description of the response opti	datory	elements 📃
	Other	Mandatory (if applicable)	For any other type of format, (e.g., postal codes digits, ID), relevent	information about the variable format	iatory	Nam
Target	Individual/ entity	Mandatory	Individual (or entity) the variable is about (e.g., participant, sample		inal	Any Valu
ranget	Time period	Mandatory	Time period the variable refers to (or is about) (e.g., first trimester of	of pregnancy, last 30 days, at birth).		Value Units : for relevant decimal, integer and text variables (e.g., height, date of birth), units of the study-specific variable(s) (e.g., cm, vvvv-mm-dd).
	Source	Mandatory	Acceptable source of information (e.g., questionnaire, government variable data could be extracted	al database/registrie, etc.) from which the		Categories : categories of the study-specific variable(s) as defined in the study-specific questionnaires/data dictionnary.
	Informant	Mandatory (if applicable)	For information that can be provided by different persons, accepta	ble informant (e.g. participant, partner, pro	(v) (nal	Time period targeted by the study-specific variable(s) (e.g., first trimester of pregnancy, last 30 days, at birth). Time of collection of the study-specific variable(s) (e.g., smoking status during first trimester of pregnancy can be collected at birth(or not)).
	Time of collection	Mandatory	Acceptable time of collection of the information (e.g., smoking stat collected at birth(or not)).	us during first trimester of pregnancy can be	inal	Acceptable source of information (e.g., questionnaire, governmental databases/registries, biosample, etc.). For information that can be provided by different persons, acceptable informant (e.g. participant, partner, proxy).
Requirements	Wording	Mandatory (if	For data extracted from questionnaires, acceptable question wordi	ng (e.g., binge drinking defined as 5 drinks o	r	Any information related to the wording, measures or procedures that needs to be specified to estimate the harmonization status : Wording: acceptable or required information from questionnaires to support understanding of the study-specific variable(s).
		applicable) Mandatory (if	more). For data that can be measured using different tools/devices, accept	while teach and endowing (a p. CAD 7, DIO 0, and	_	Measures: acceptable or required information for study-specific variable(s) measured using different tools/devices, acceptable tool or device (e.g. GAD-
	Measures	applicable)	For data that can be measured using different toois/devices, accept gravity manometer, etc.).	able tool of device (e.g. GAD-7, PHQ-9, Mer	Lury	7,PHQ:9, mercury-gravity manometer (Exhibit 2), etc.).
		Applicable) Mandatory (if	gravity manometer, etc.). For data that can be generated using different procedures, accepta	ble annual constant annual be		Procedures: acceptable or required information on different procedures used to collect the study-specific variable(s) (e.g., weight needs to be measured, not
	Procedures	applicable)	For data that can be generated using different procedures, accepta measured, not reported by the participant).	procedures (e.g., weight needs to be		reported by the participant).
		appricable)	For variables to be generated using specific standard or instruction:			instructions: acceptable or required infomation on specific standard or instructions used to collect the study-specific variable(s) (e.g. ICD-10, ISCED,
Instructions		Mandatory (if	(e.g. ICD-10 must be used, prematurity defined as birth before 37 w		inal	prematurity defined before 37 weeks of pregnancy according to a given reference). Internal comments or nuestions to the team members to belo or document the barmonization process and decisions
		applicable)	reference, specific information related to management of unique o		inal	Internal comments or questions to the team members to nelp or document the harmonization process and decisions. Any additional relevant information on the study-specific variables used to generate the DS variable and to be documented online.
Comments		Optional	Any additional information not covered	marapie choices option, etc).	datory	any additional relevant information on the study-spectric variable.
		Tables of the		Harmonization status details	Mandatory	namionization status of the study-specific harmonized variable.
				Rule category	Mandatory (if	namonration status octan or ne stopyspectric namonrato variable.
					applicable)	
				Harmonization rule	Mandatory (if	Simplified algorithms to be used to generate the study-specific harmonized variable.
					applicable)	
						· · · · · · · · · · · · · · · · · · ·



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

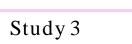
Stu	dy 1
Category	Count
Yes	1 899
No	461
Missing	5
Total	2 365



Stu	dy2
Category	Count
Yes	2 728
No	221
Missing	23
Total	2 972

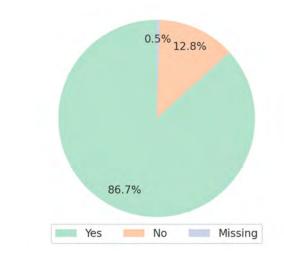
19.5% 80.3%





# Combined harmonized data

Category	Count
Yes	4 627
No	682
Missing	28
Total	5 3 3 7



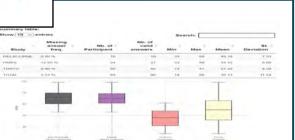


											_											_
Steps	MVPA (se	Bouts	TPA (seco	MPA (sec		M Thresh	V Thresh	Filter (ser	010-1	Q10-2	6	210-3	Q10-4	Q10	-5	Q10-6	Q10-7	Q10-8	Q10-9	Q10-10	Q10-11	1
5262						100	130		410 1	0	•		4104	0		4100	0		0	0	0	•
										-	0			0			-	•	•	-	0	- 0
7628			4092			100	130			0	0		U	0	0		0	-	0	0	0	0
12100	4904	3	6303	4056	848	100	130	4		0	0		D	0	0		0	0	0	0	0	0
10282	3899	3	6099	3875	24	100	130	4		0	0		D	0	0		0	0	0	0	0	0
1104	288	0	693	288	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
0	0	C	0 0	0	0	0	0	0	1	0	0		D	0	0		0	0	0	0	0	0
66	25	C	46	25	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
46	14	C	29	14	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
4191	1465	C	2416	1384	81	100	130	4		0	0		D	0	0		0	0	0	0	0	0
3749	1073	C	2233	975	98	100	130	4		0	0		D	0	0		0	0	0	0	0	0
6929	2375	1	4131	2344	31	100	130	4		0	0		D	0	0		0	0	0	0	0	0
2756	595	C	1750	578	17	100	130	4		0	0		D	0	0		0	0	0	0	0	0
3367	801	C	2114	801	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
19	0	C	13	0	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
18	5	C	10	5	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
109	32	C	65	32	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
21	11	C	11	11	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
59	24	C	30	24	0	100	130	4		0	0		D	0	0		0	0	0	0	0	0
22	6		13	6	0	100	130	1		0	0		n	0	0		n	0	0	0	0	0



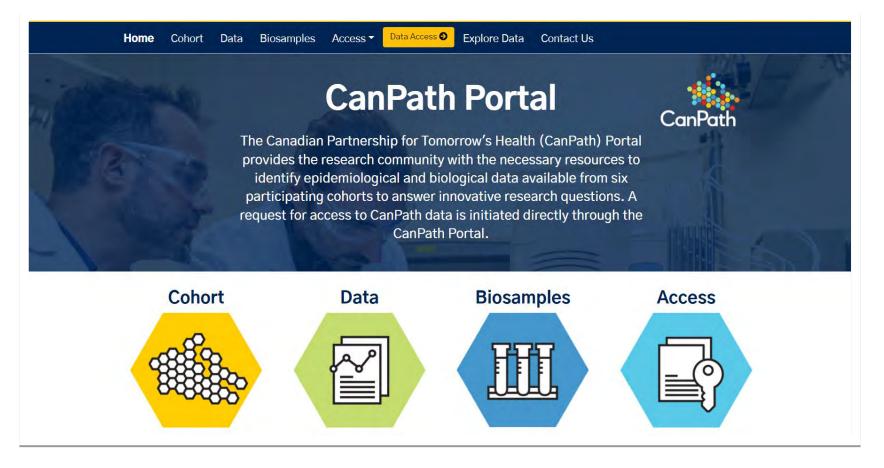
E	F	G	Н	- I	J
valueType Detected Mix cat. and data type other values			Categories in data dictionary	unit	Quality assessment comment
text	character				
integer	numeric		Valid categorical values : 0 = Male 1 = female		[INFO] - Possible duplicated variables: SEX ; Gender
integer	numeric			kg/m^2	[ERR] - ValueTyp 3 status
integer	numeric			years	4 time_completion_day 5 time_completion_minutes
integer	numeric		Valid categorical values : 0 = Non smoker 1 = smoker		6 F AU03_BEER 7 F AU03_UIQUOR 8 F AU03_OTHER ALC 9 F AU03_RED_WINE
integer	numeric	YES	Valid categorical values : 8 = ADMISSION PATTERN	-	[ERR] - Incompa 11 F-AU03_WHITE_WINE [IRR] - Incompa 11 F-AU01_WEIGHT_K6 [INFO] - More ut 12 F-AU01_WEIGHT_L8 13 F-AU02_WAIST_FIRST_CM 14 F-AU02_WAIST_FIRST_IN

	A	В	C	D	E	F	6
1	Variable name	Minimum 1	ist quartile /	Median	3rd quartile 1	Maximum	Mean
2	PROJECT_CODE	47243000003	47243010184	47243020489	47243031877	47243043111	
<b>p</b> 3	status	1	1	1	1	1	
-4	time_completion_day	0	0	0	17	788	
5	time_completion_minutes	-52	31	79	24236	1134596	
6	F_AU03_BEER	0	1	2	5	350	
7	F_AU03_LIQUOR	0	1	1	2	150	
8	F_AU03_OTHER_ALC	0	0	1	2	150	
9	F_AU03_RED_WINE	0	2	4	6	1500	
10	F_AU03_WHITE_WINE	0	1	2	4	150	
11	F_AM01_WEIGHT_KG	27	65	78	90	240	
12	F_AM01_WEIGHT_LB	1	140	164	190	1307	
13	F_AM02_WAIST_FIRST_CM	26	85	93	102	194	
14	F_AM02_WAIST_FIRST_IN	1	34	37	40	89	
15	F_AM03_WAIST_SECOND_CM	30	84	93	102	3740	
16	F_AM03_WAIST_SECOND_IN	1	34	37	40	4145	
17	F_AM04_HIPS_FIRST_CM	30	93	99	105	193	
18	F_AM04_HIPS_FIRST_IN	1	37	-40	42	481	
19	F_AM05_HIPS_SECOND_CM	30	93	99	105	194	
20	F_AM05_HIPS_SECOND_IN	1	37	40	42	371	
21	F_PM03_EYE_CATARACT_ONSET	1	58	63	67	76	
22	F PM03 EYE GLAUCOMA ONSET	1	48	56	64	76	

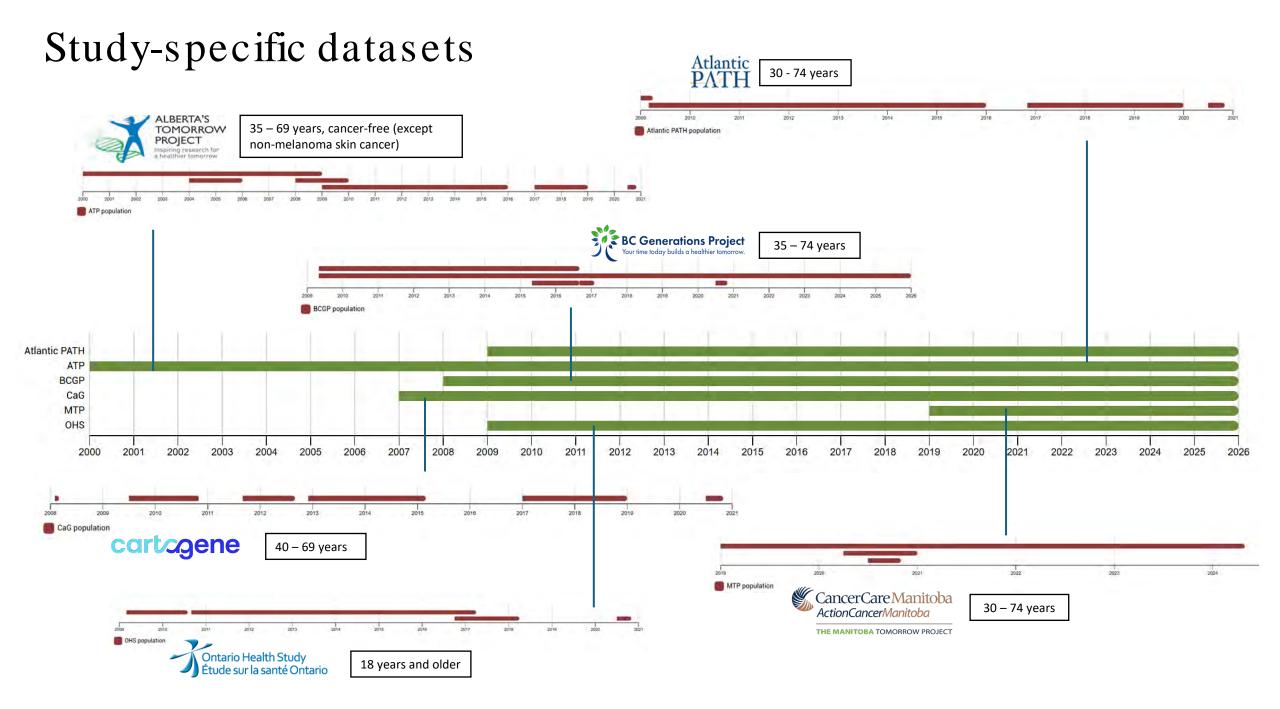




# 4,500 harmonized variables!



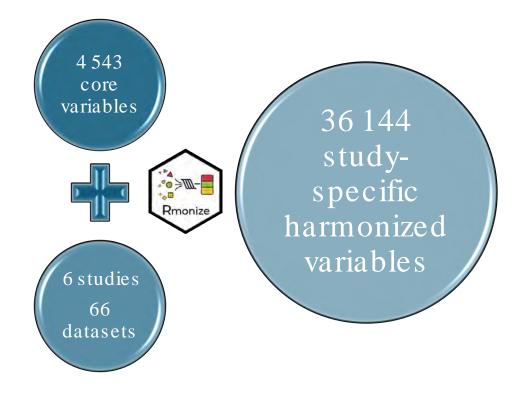
Fortier I, Dragieva N, Saliba M, Craig C, Robson PJ, with the Canadian Partnership for Tomorrow Project's scientific directors and the Harmonization Standing Committee Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis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



# 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 1 1 0

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 4 2 4

# Harmonization potential Health and Risk Factor Questionnaire - Baseline

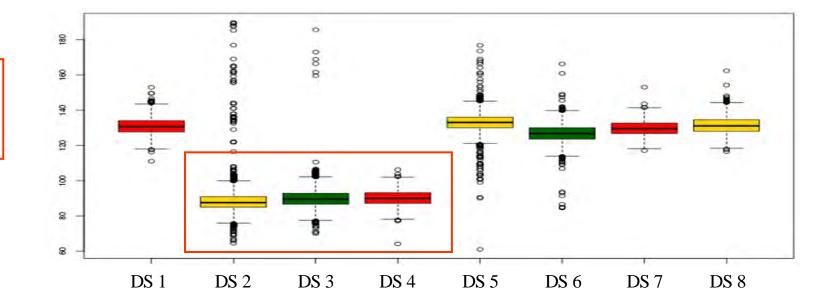
	% Complete 🕕	Atlantic PATH 1 80%	Atlantic PATH 2 49%	ATP 1 61%	ATP 2 79%	BCGP 1 76%	BCGP 2 75%	BCGP 3 65%	CaG 1 62%	CaG 2 77%
Sex	100%	~	~	~	~	~	~	~	~	~
Age	100%	~	~	~	*	~	~	~	*	~
Number of brothers	75%	~	~	×	~	~	~	~	×	×
Number of sisters	75%	~	~	×	*	~	~	~	×	×
Number of siblings	92%	~	~	~	~	~	~	~	~	×
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

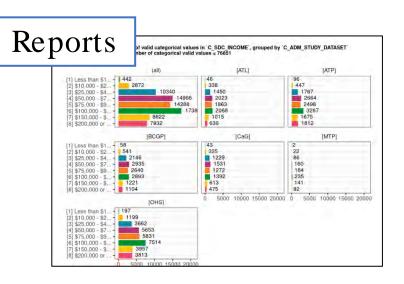
	Ν	Mean		
DS 1	4872	130,88		
DS 2	22703	88,01		
DS 3	29347	89,8		
DS 4	1149	90,2		
DS 5	16363	133,21		
DS 6	19992	126,83		
DS 7	649	129,78		
DS 8	7970	131,3		





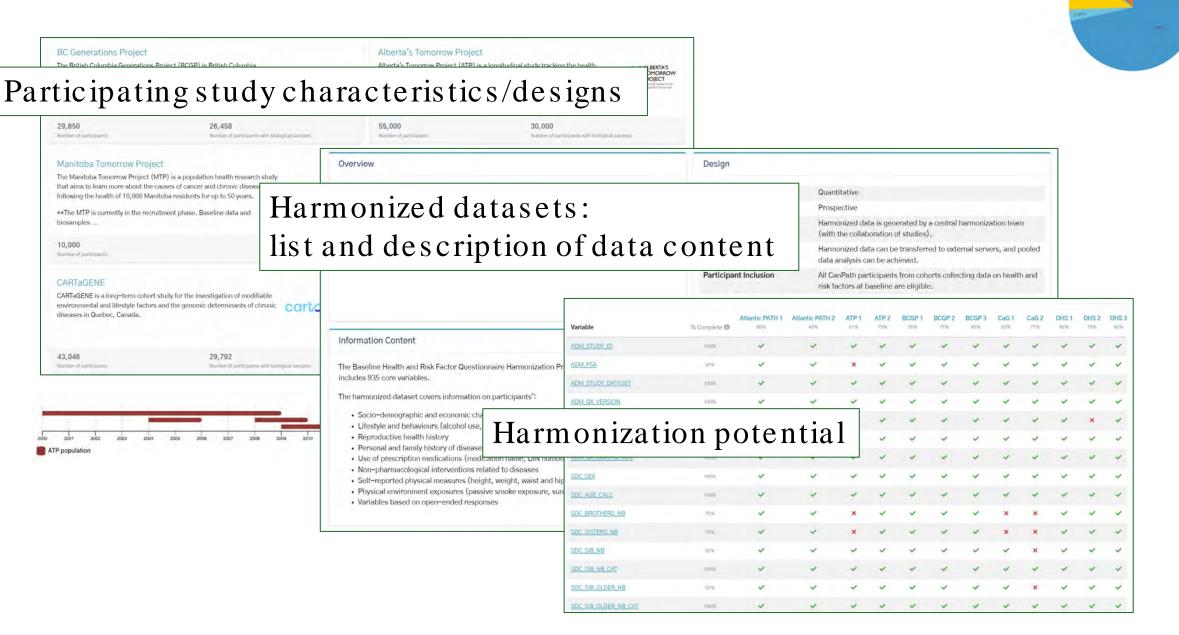


Q10-3 



Repc Grou	D	<b>)</b> a1	tas	set	ts ar	nd o	dat	a	dic	ti	on	ar	ies	5!	shc Fil	ter (sec Q10-1	Q10-2	Q10-3	Q10-4	Q10-5	Q10-6	Q10-7	_
GROU										-					130	4	0	0	•	0	0	0	_
GROUP1			F	-	2018/07/1YES	Y00B25	PiezoRxD	7628	2721	1	4092	2066	655	100	130	4	0	0	-	0	0	0	_
	AMELIE		F	-	2018/07/1YES	Y00B25	PiezoRxD	12100	4904	3	6303	4056	848	100	130	4	0	0	0	0	0	0	-
		ROY	F	-	2018/07/1YES	Y00B25	PiezoRxD	10282	3899	3	6099	3875	24	100	130	4	0	0	0	0	0	0	_
	AMELIE	ROY	F	-	2018/07/1YES	Y00B25	PiezoRxD	1104	288	0	693	288	0	100	130	4	0	0	0	0	0	0	
GROUP1	AMELIE	ROY	F	-	2018/09/1NO	-	PiezoRxD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	_
GROUP1	AMELIE	ROY	F	-	2018/10/2YES	Y00B25	PiezoRxD	66	25	0	46	25	0	100	130	4	0	0	0	0	0	0	
GROUP1	AMELIE	ROY	F	-	2018/10/2YES	Y00B25	PiezoRxD	46	14	0	29	14	0	100	130	4	0	0	0	0	0	0	1
GROUP1	AMELIE	ROY	F	-	2019/01/1YES	Y00B25	PiezoRxD	4191	1465	0	2416	1384	81	100	130	4	0	0	0	0	0	0	1
GROUP1	AMELIE	ROY	F	-	2019/01/1YES	Y00B25	PiezoRxD	3749	1073	0	2233	975	98	100	130	4	0	0	0	0	0	0	1
GROUP1	AMELIE	ROY	F	-	2019/01/1YES	Y00B25	PiezoRxD	6929	2375	1	4131	2344	31	100	130	4	0	0	0	0	0	0	1
GROUP1	AMELIE	ROY	F	-	2019/01/1YES	Y00B25	PiezoRxD	2756	595	0	1750	578	17	100	130	4	0	0	0	0	0	0	1
GROUP1	AMELIE	ROY	F	-	2019/01/1YES	Y00B25	PiezoRxD	3367	801	0	2114	801	0	100	130	4	0	0	0	0	0	0	1
GROUP1	AMELIE	ROY	F	-	2019/01/2YES	Y00B25	PiezoRxD	19	0	0	13	0	0	100	130	4	0	0	0	0	0	0	,
GROUP1	AMELIE	ROY	F	-	2019/01/2YES	Y00B25	PiezoRxD	18	5	0	10	5	0	100	130	4	0	0	0	0	0	0	,
GROUP1	AMELIE	ROY	F	-	2019/02/2YES	Y00B25	PiezoRxD	109	32	0	65	32	0	100	130	4	0	0	0	0	0	0	,
GROUP1	AMELIE	ROY	F	-	2019/02/2YES	Y00B25	PiezoRxD	21	11	0	11	11	0	100	130	4	0	0	0	0	0	0	,
GROUP1	AMELIE	ROY	F	-	2019/02/2YES	Y00B25	PiezoRxD	59	24	0	30	24	0	100	130	4	0	0	0	0	0	0	7
GROUP1	AMELIE	ROY	F	-	2019/02/2YES	Y00B25	PiezoRxD	22	6	0	13	6	0	100	130	4	0	0	0	0	0	0	7

# CanPath Data Portal



# CanPath Data Portal

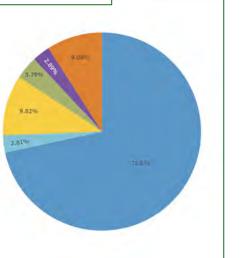
## Variables descriptions

armonize	tion Initiative	CORE	
Harmoniza	tion Protocol	CANPATH-BL-HREQ-HP	
Value type		Integer	
Categorie	s		
Name	Label		Missing
1	Married and/or	r living with a partner	
2	Divorced		
3	Widowed		
4	Separated		
5	Single, never n	narried	
	Single, never n	narried	

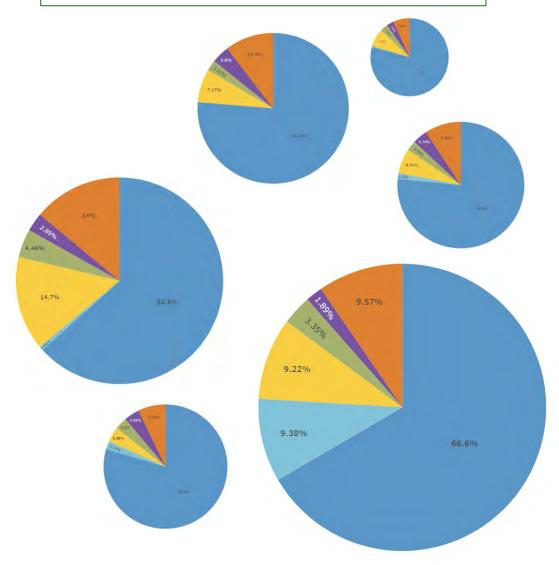
Classifications
Additional information
Source
Questionnaire
Target
Participant
Areas of information
Socio-demographic and economic characteristics
Marital/partner status
Find similar variables Q

# Participants distributions – total

Values	Frequencies	<b>^</b>	
Valid Values			
1 Married and/or living with a partner	234,602 73.68% (71.61%)		2.764
2 Divorced	29,759 9.35% (9.08%)		3.79%
3 Widowed	9,481 2.98% (2.89%)		9,82%
4 Separated	12,402 3.9% <i>(3.79%)</i>		2.81%
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

		• •		riab	~							YR Add to cart ▲ Do	winload			
Variables +	4,557 Protocols 18										20 •	1 2 3				
Name	ie -	Label					Value type	Annotations	3	Filter the torms b	wala					
	GLCUB	Preg	nancy					<ul> <li>Physical and cog</li> <li>Pregnancy, delive</li> </ul>	nitive measure							
	G CUR WK	O Numi	ber of wee	eks pregnant			Integer		A	reas of info	ormation		_			
BL00	OD. TRANS. PREV. 2MTHS	O Occu	rrence of	blood transfusion in pre-	wious 2 months		Integer	Sel	eci	tion	n crit	teria			Marital/partner status	
CHEN	MO_PREV_12WKS	O Occu	rrence of	chemotherapy treatmen	nt in previous 12 weeks		Integer	Radiological inter	ventions	C comony, mos	ano rongion	El randrado	Residen     Labour 1	ice force and retirement	<ul> <li>Birthplace</li> <li>Income, possessions, and b</li> </ul>	
	PREV-12WKS	O Occu	mence of	radiotherapy treatment	in previous 12 weeks		Integer 4	<ul> <li>Physical and cog</li> <li>Radiological inter</li> </ul>	nitive measure							
E FOOD	D. DRINK, TIME	Calc	ulated tim	e since last food or drink	k was consumed, excludin	ng plain water		Physical and cog	nitive measure	Lifestyle and beha	aviours O					
E FOOD	D DRINK PREV 24HRS	O Cons	umption o	f food or drink in previo	us 24 hours, excluding pl	ain water		Physical and cog	nitive measure	Tobacco		Alcohol	Drugs		Nutrition	
										Birth, pregnancy	and reproductive he	alth history <b>O</b>		tion	Fertility and sexual healt	
st Coverage										Birth, pregnancy	and reproductive he	alth history <b>O</b>	X Add to cart Down	- Malicana and Islatic		
				1.1						Birth, pregnancy	and reproductive he	aith history <b>O</b>	₩ Add ta cart 🛓 Down	loar delivery and birth		
	paris	on	ta	able status×	ily and schold structure ×	Education ×	Residence ×	Birthplace ×	Ethnicity, rac and religion ×	ce Language	and reproductive hes	Income, possessions, and benefits ×		delivery and birth	<ul> <li>Fertility and sexual healt</li> </ul>	
omp		011 × 8			sehold		Residence ×	Birthplace × 14	Ethnicity, rac	ce Language	Labour force	Income, possessions, and	◆Fit Other socio~demographic and economic	delivery and birth	Fertility and sexual healt     Other perception of heal     other and functional limit	
omp	paris	<b>O</b> 11 × 8 3		status ×	structure ×	×	×	×	Ethnicity, rac and religion >	ce Language × ×	Labour force and retirement ×	Income, possessions, and benefits ×	Other socio-demographic and economic characteristics x	delivery and birth	Fertility and sexual healt     Other perception of heal     other and functional limit	
D M I	paris *	<b>O</b> 11 × 8 3	×. 1 0	status × 2	structure ×	×	×	×	Ethnicity, rac and religion >	ce Language × ×	Labour force and retirement × 161	Income, possessions, and benefits × 13	Other socio-demographic and economic characteristics ×	delivery and birth	Fertility and sexual healt     Other perception of heal     other and functional limit	
Protocols	paris *	× 8 3	×. 1 0	status × 2 0	27 6	× 3 0	× 5 2	× 14 0	Ethnicity, rac and religion × 58 0	ce Language × × 34 0	Labour force and retirement × 161 9	Income, possessions, and benefits × 13 8	Other socio-demographic and economic characteristics × 1	delivery and birth	Fertility and sexual healt     Other perception of heal     other and functional limit	

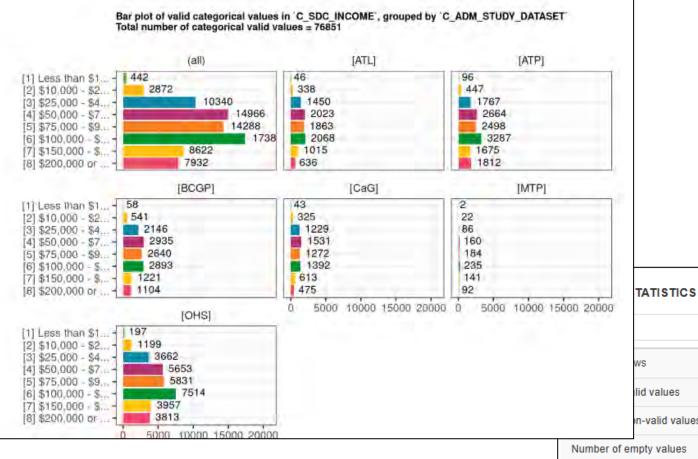
#### CanPath

User name or email	$\sim$
Password	<b>A</b>
	Sign In

# Reports Data Processing Elements

dataschema_variable	ss_variables	ss_format	ss_source_information	ss_informant	harmo_status	armo_status_det	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	<pre>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_) recode("np" = "10010-00"; "smt1" =</pre>
C2_CITF_ASSAY_ID								"10020-00"; "rbd" = "10030-00"; ELSE =

Reports



000 20000							
		[ATL]	[ATP]	[BCGP]	[CaG]	[MTP]	[OH S]
	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 e	empty values	1301	303	901	468	40	1886
% Valid valu	les	73.85	78.4	76.65	84.68	82.76	80.03
% Non-valio	l values	15.97	19.93	18.25	9.56	13.64	15.22
% Empty va	lues	10.18	1.67	5.1	5.76	3.59	4.74
Number of o	listinct values	10	10	10	10	10	10



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

# Thanks!



#### www.maelstrom-research.org



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