Harnessing Canadian cohorts potential: Data harmonization and co-analysis

Isabel Fortier Ph.D.

The Canadian Association for Population Therapeutical

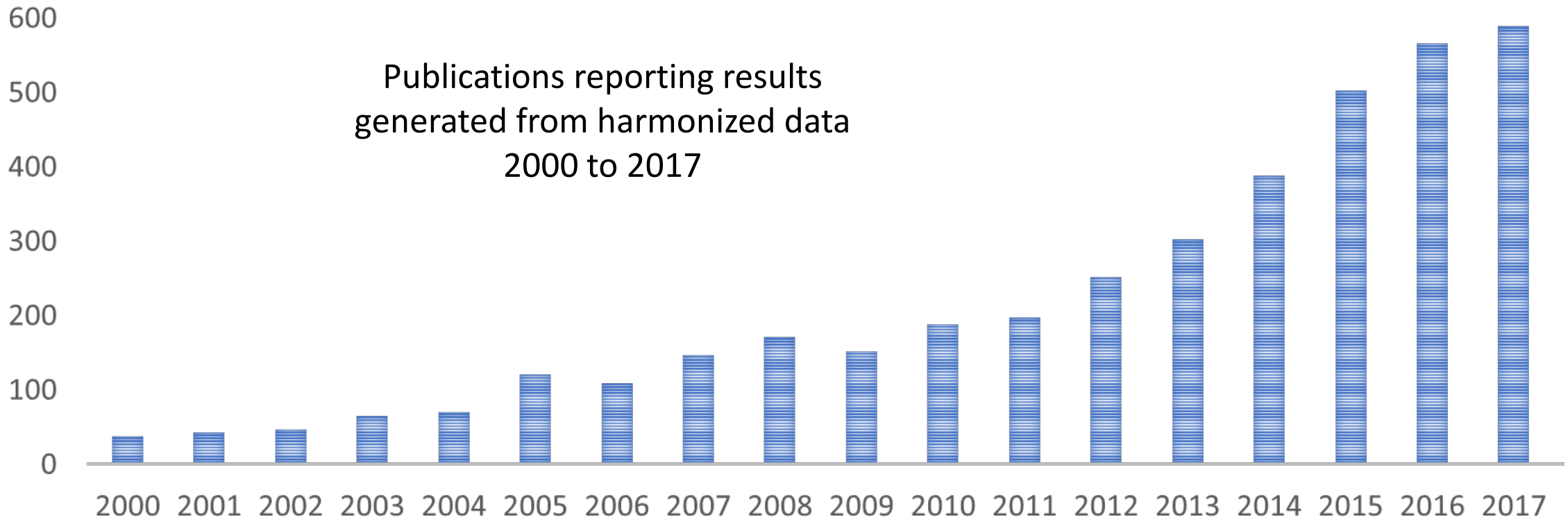
Driving Health Innovation: Harnessing the Power and value of real-world evidence

Toronto , September 21th 2019

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

A paradigm shift in the manner we conduct research

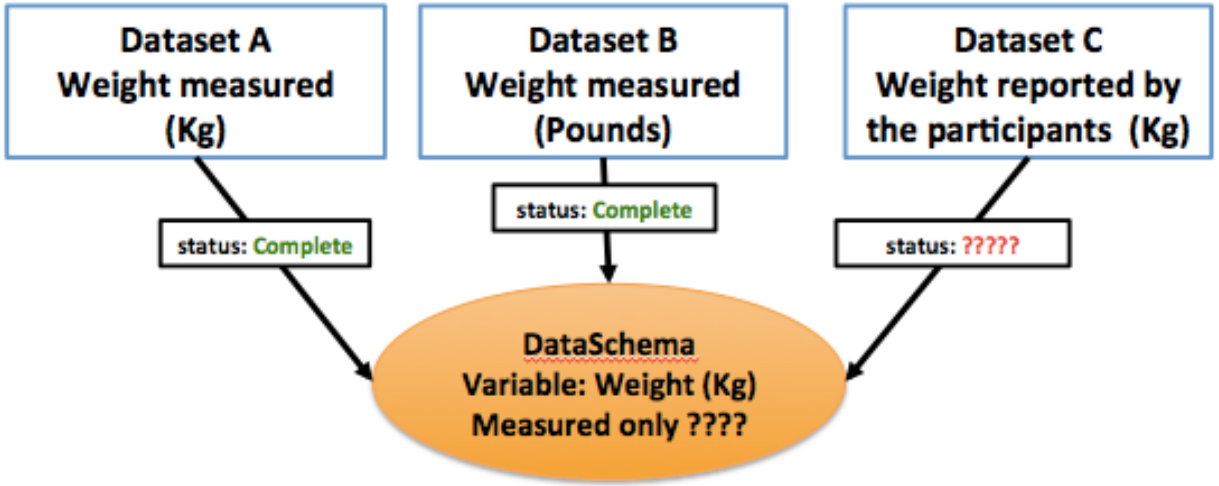
Influenced by the need to: obtain **larger sample sizes** and statistical power; conduct **comparative research** across studies/jurisdictions; **extend** the scientific **impact** of individual studies/data sources.

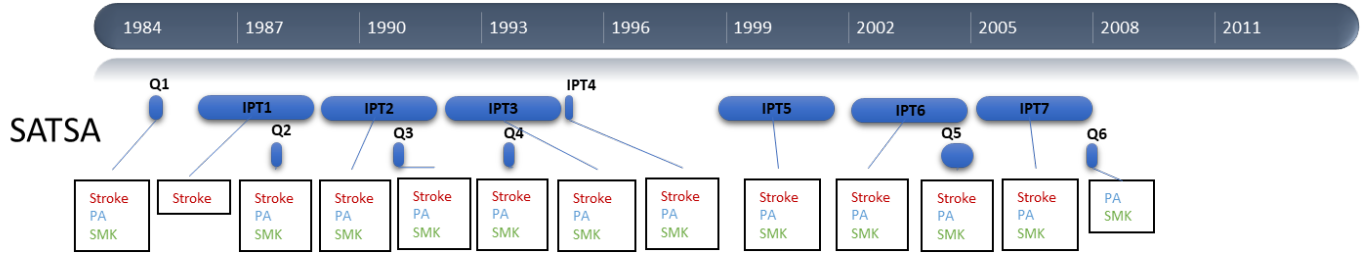



Anxiety Level
None
Mild
Mild-Moderate
Moderate
Moderate-High
Highest
Faces
Worry/Anxiety
5
freaking out panic
4
really worried
3
worried nervous anxious
2
a little worried
1
okay
Chromosome
Excellent
Very Good
Average
provide us feedback on

Retrospective harmonization: Reality check…

- **STUDY 1: Do you currently have a job or do any unpaid work outside your home?**
  - |_| Yes |_| No
- **STUDY 2: What is your job title?**
  - OPEN_____
- **STUDY 3: What is your current employment status?**
  - |_| Working full-time
  - |_| Working part-time
  - |_| Unable to work because of sickness or disability
  - |_| Looking after home and/or family
  - |_| Student
  - |_| Retired
  - |_| Unemployed
  - |_| Doing unpaid or voluntary work

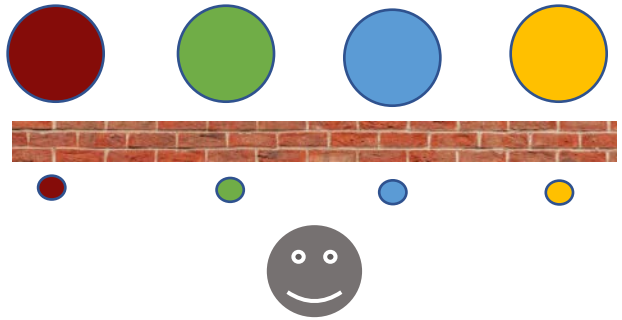




Data infrastructures supporting co-analysis across studies

Summary data meta-analysis

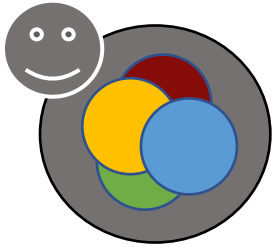
Study-specific data analyses done locally followed by a meta-analysis combining the study-level estimates



No sharing of IPD
Harmonization: local
Analysis: local

Pooled analysis

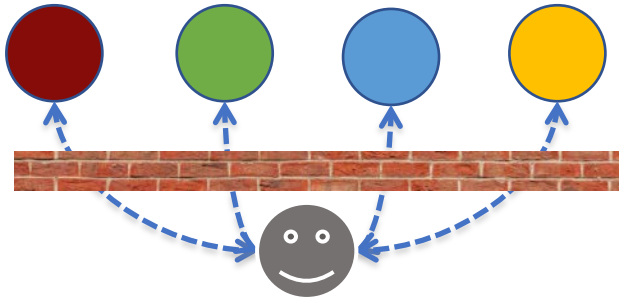
Data pooled and analyzed in a central location



Sharing of IPD
Harmonization: local or central
Analysis: central

Federated analysis

Analyses done centrally, but the individual-level participant data can remain on local servers



Sharing (or not) of IPD
Harmonization: local
Analysis: central

mælstrom

- Tools and methodological guidelines to support data cataloguing, management, harmonization and co-analysis.
- A central study catalogue developed with our partners to foster usage of available data

  200 studies and over 900 000 variables documented
- Expert support to national and international research initiatives to implement catalogues and harmonization platforms


PLOS ONE

RESEARCH ARTICLE

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

Julie Bergeron[1], Dany

1 Research Institute of the and Public Health Institute, of the Sainte-Justine Univer

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

Original Article

Maelstrom Research guidelines for rigorous retrospective data harmonization

Isabel
Laure
P Sto
Peter

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

Software Application Profile

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

Dany Doiron,
Vincent Ferre

Data Matters

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

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

www.maelstrom-research.org

BRAIN - Broad and Deep Analyses in Neurodegeneration

PHQE: Quebec Europe data harmonization platform

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

ReACH

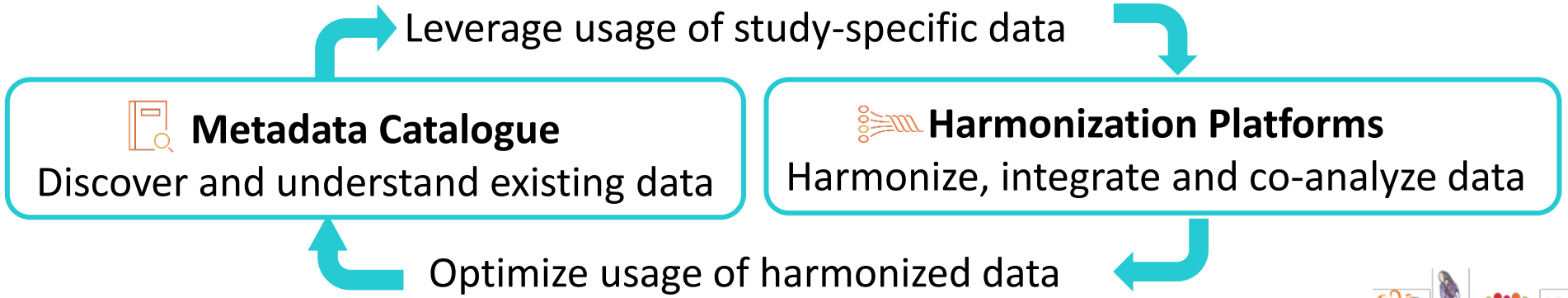
Research Advancement through
Cohort Cataloguing and Harmonization

53 investigators from different fields of research

27 Canadian pregnancy and birth cohorts (DOHaD)

Aims to provide the research community with the means to:

- **Leverage usage** of existing study-specific data and samples.
- Facilitate **research** co-analyzing data across studies
- Support implementation of **new** mother/child **cohorts**.

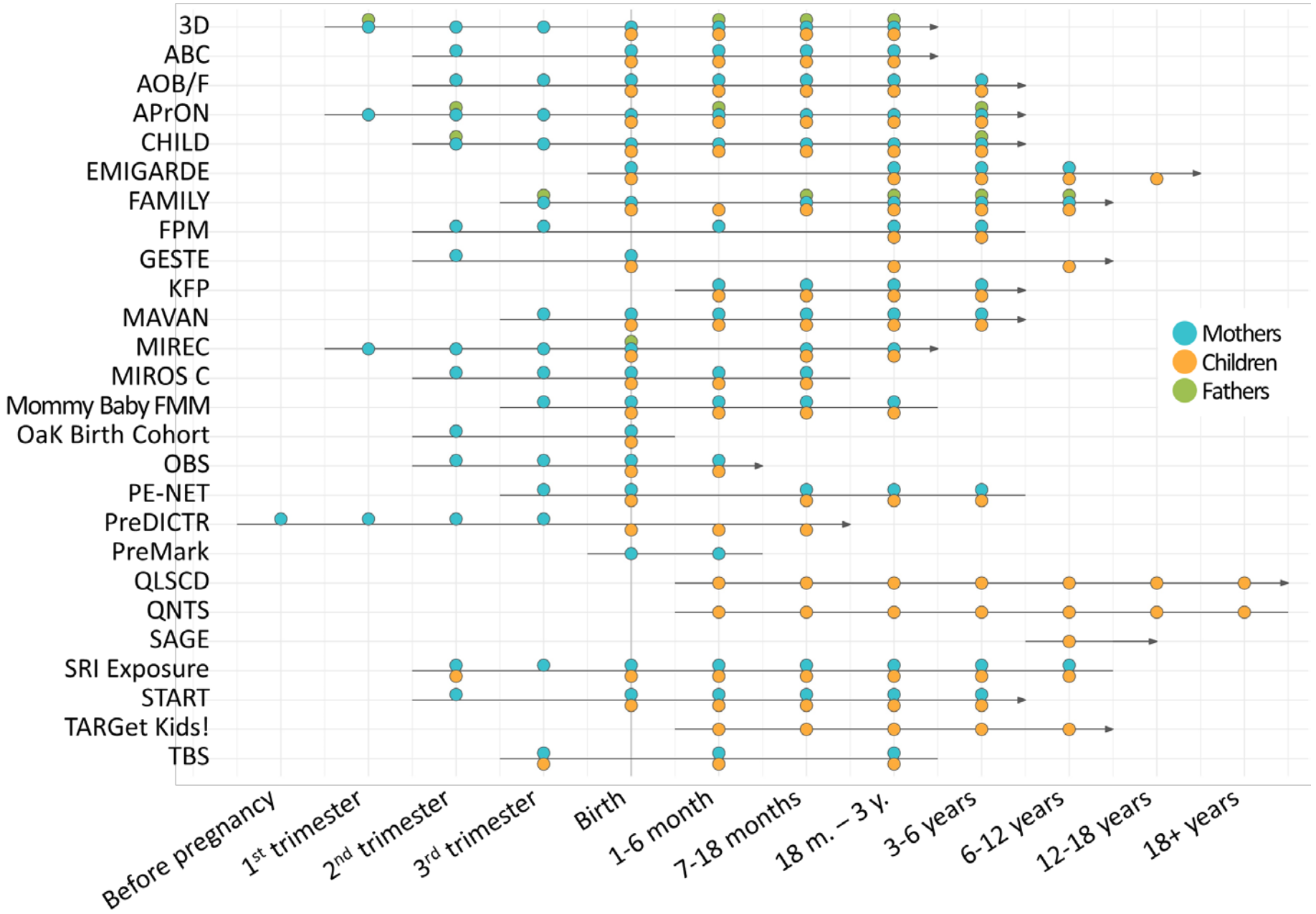


EMIGARDE, GESTE, MAVAN, MIROS, PreMark, QNTS, SRI_Exposure

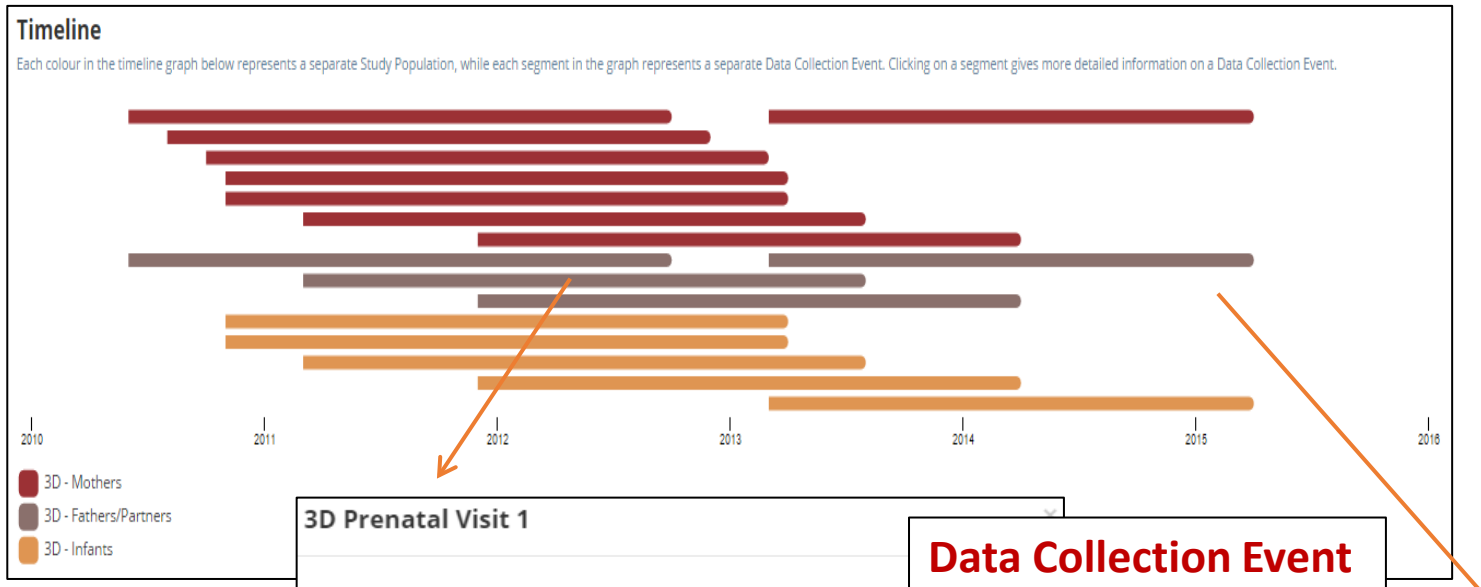
ReACH
Research Advancement through Cohort Cataloguing and Harmonization

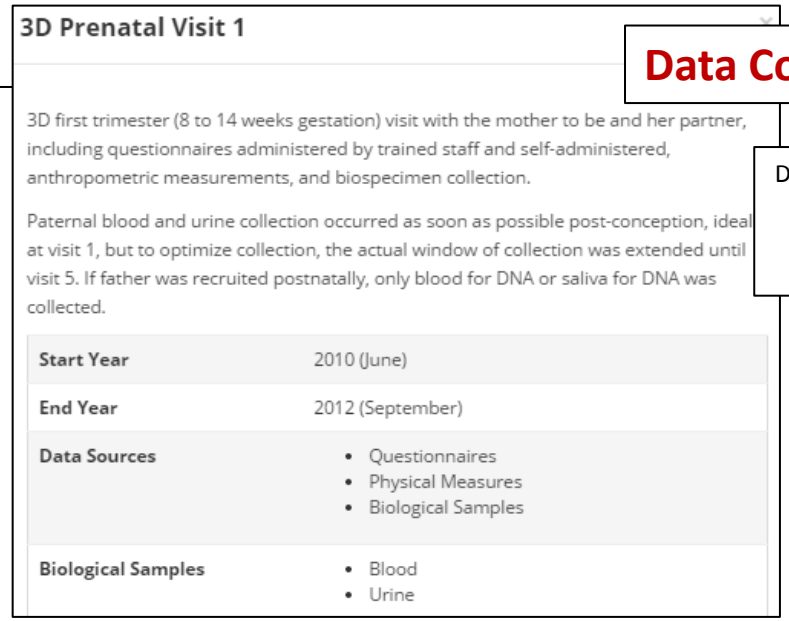
Participants and follow-up

35,070 mothers
39,835 children
7,239 fathers
82,144 participants



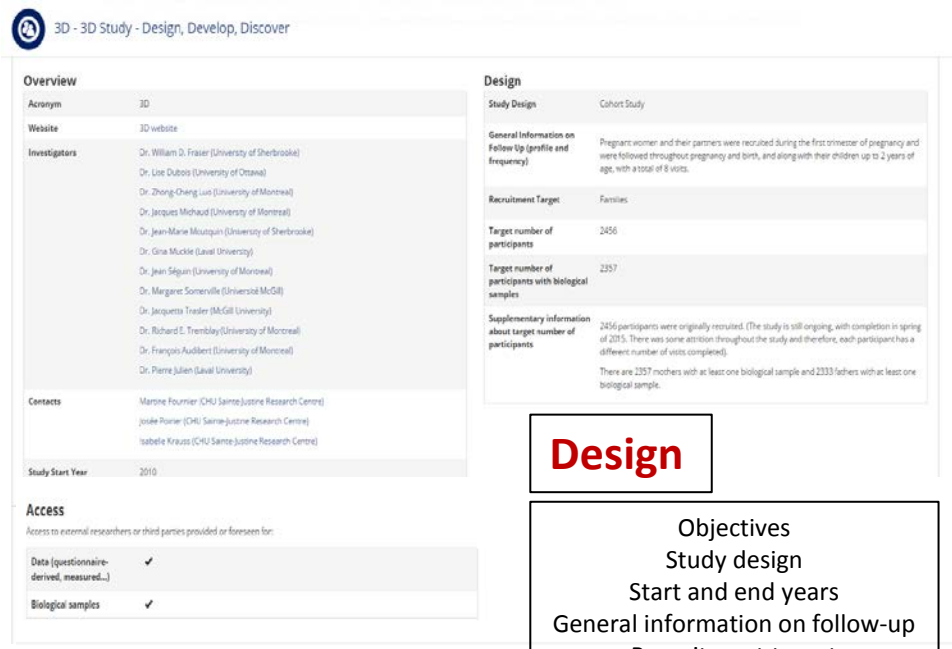
Study Description





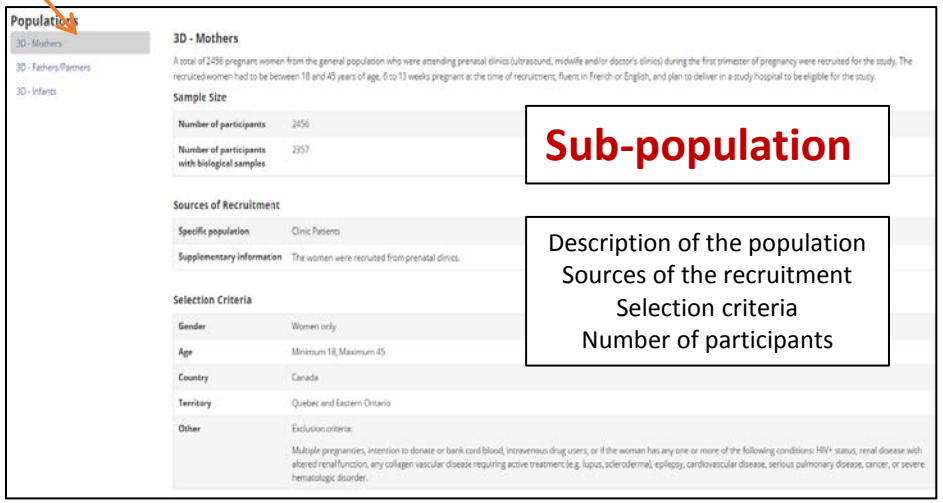
Data Collection Event

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



Design

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



Sub-population

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

Areas of information (and scales)

17 Sections; 132 Categories

- Socio-demographic and economic characteristics
- **Lifestyle and health behaviors**
- Health status and functional limitations
- Diseases (ICD-10)
- Symptoms and signs (ICD-10)
- Medications and supplements
- Non-pharmacological interventions
- Health and community care utilization
- Reproduction
- Birth, infancy and childhood
- End of life
- Physical measures
- Cognition, personality and other psychological measures
- Laboratory measures
- Social environment and life events
- Physical environment
- Administrative information

Tobacco
Alcohol
Illicit drugs
Nutrition
Physical activity
Transportation
Personal hygiene
Sleep
Sexual behaviors
Leisure activities
Other and unspecified lifestyle information

Clear

ReACH × + Selection criteria - Pregnant ... × Start year:[2010,2018] × + Nutrition | Physical activity × Age/birthdate | Education ×

Anthropometry × Advanced

Variables

- Areas of Information
- Scales/Measures
- Source & target
- Properties

Studies

- Properties

Networks

- Properties

List | Comparison Table | Summary Statistics

Download

All | Individual | Harmonization

Population/Data Collection Event (DCE)

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

Data Collection Event (DCE)

	Data Collection Event (DCE)			Socio-demographic and economic characteristics ×			Lifestyle and health behaviours ×		Diseases ×
	Study	Population	DCE	Marital/partner status ×	Education ×	Income, possessions, and benefits ×	Tobacco ×	Alcohol ×	Mental and behavioural disorders (F00-F99)
	3D	3D - Mothers	3D Prenatal Visit 1 2010-08 to 2012-09	6	3	4	16	26	5
			3D Prenatal Visit 2 2010-08 to 2012-11	1	0	2	7	20	1
			3D Prenatal Visit 3 2010-10 to 2013-02	1	0	2	7	20	1
			3D Delivery Visit 2010-11 to 2013-03	0	0	0	0	0	0
			3D Immediate Postpartum Visit and Chart Review 2010-11 to 2013-03	0	0	2	0	0	0
			3D 3-Month Postpartum Visit 2011-03 to 2013-07	3	0	24			
			3D 12-Month Postpartum Visit 2011-12 to 2014-03	1	0	17			
			3D 24-Month Postpartum Visit 2013-03 to 2015-03	3	0	4			
		3D - Fathers/Partners	3D Prenatal Visit 1 2010-08 to 2012-09	1	2	0			
			3D 3-Month Postpartum Visit 2011-03 to 2013-07	1	0	0			
			3D 12-Month Postpartum Visit 2011-12 to 2014-03	1	0	0			
			3D 24-Month Postpartum Visit 2013-03 to 2015-03	1	0	2			
		3D - Infants	3D Delivery Visit 2010-11 to 2013-03	0	0	0			
			3D Immediate Postpartum Visit and Chart Review 2010-11 to 2013-03	0	0	2			
			3D 3-Month Postpartum Visit 2011-03 to 2013-07	0	0	0			
			3D 24-Month Postpartum Visit 2013-03 to 2015-03	0	0	0			

A1PA53

Overview

Label	Ever drank at least a drink per day more than 3 days a week
Description	Was there ever a time in your life when you regularly had at least one drink three or more days a week?
Study	Midlife in the U.S.
Dataset	MIDUS1
Value Type	integer
Variable Type	Study variable

Classification

Additional information	
Source	Questionnaire
Target	Participant
Areas of Information	
Lifestyle and health behaviours	Alcohol

Categories

Name	Label	Missing
7	DONT KNOW	✔
8	REFUSED/MISSING	✔
9	INAPPLICABLE	✔
1	YES	
2	NO	

Information collected on sleep duration (N=20 studies…)

Target Variable:
Number of hours of sleep
in a 24h period

FAMILY
Weekday: When did child get out of bed in the morning?
Weekday: At what time did child go to bed?
Weekday: How many hours of sleep does child get per night?
Saturday: When did child get out of bed in the morning?
Saturday: At what time did child go to bed?

TARGet Kids
How many hours does your child usually spend sleeping in a 24-hour period?

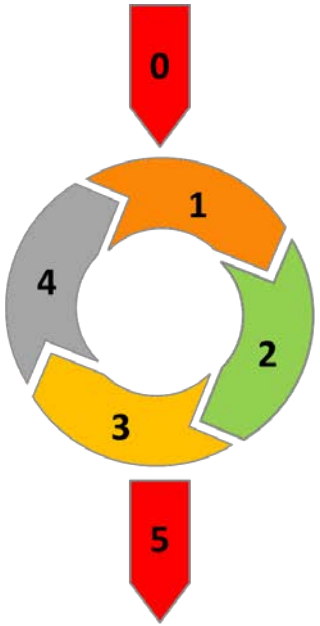
and if
Number of hours of sleep
During the night?

CHILD
What is your child's usual bedtime?
What is your child's usual wake up time?
How long does your child nap per 24-hour period?
Child's usual amount of sleep each WEEK DAY? Each WEEKEND DAY?
At what time does your child usually wake in the morning on WEEK DAYS? On WEEKEND DAYS?

START
Time sleeping during the Night?
Time sleeping during the Day?

Harmonization platforms/process

Harmonize, integrate and co-analyze data



Step 0: Define the research questions, objectives and protocol.

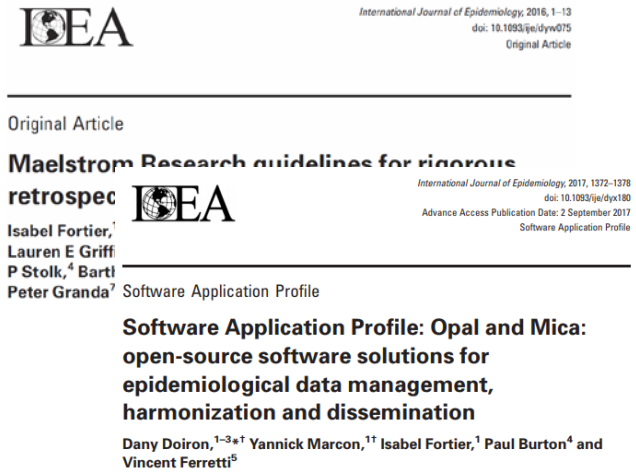
Step 1: Assemble information and select studies.

Step 2: Define variables and evaluate harmonization potential.

Step 3: Process data.

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

Step 5: Disseminate and preserve final harmonization products.



IEA

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

Original Article

Maelstrom Research guidelines for rigorous retrospec

IEA

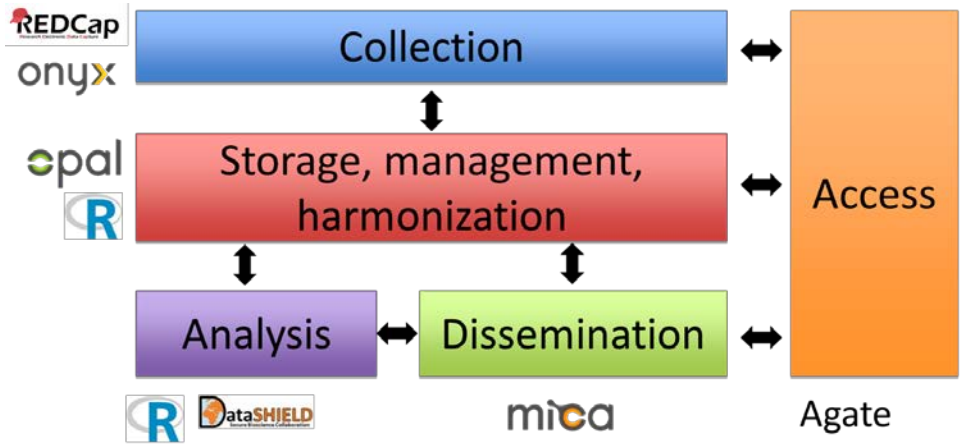
International Journal of Epidemiology, 2017, 1372–1378
doi: 10.1093/ije/dyx180
Advance Access Publication Date: 2 September 2017
Software Application Profile

Isabel Fortier,
Lauren E Griff
P Stolk,[4] Bartl
Peter Granda[7]

Software Application Profile

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

Dany Doiron,[1–3*†] Yannick Marcon,[1†] Isabel Fortier,[1] Paul Burton[4] and Vincent Ferretti[5]



Factors associated with profiles of alcohol consumption before, during and after pregnancy (A Bocking)

- Aim: Explore the variability of Canadian mothers' alcohol intake before, during and after pregnancy and the impact of alcohol consumption on birth weight and gestational age at delivery (2 papers planned) (2019-2020)
- Five Canadian mother-child cohorts; 10,263 mothers
- Information to be considered:
  - frequency and quantity of alcohol intake at multiple time point before, during and after pregnancy
  - maternal age, education, household income, occupation, ethnicity, smoking and parity
  - Birth weight, gestational age

ALL OUR FAMILIES STUDY

Design Develop Discover

FAMILY

ONTARIO BIRTH STUDY

APrON

Harmonization platforms

Retrospective

BioSHaRE .eu

MINDMAP
Promoting mental well-being and healthy ageing in cities

ReACH
Research Advancement through Cohort Cataloguing and Harmonization

CHPT
CROSS-COHORT HARMONIZATION PROJECT FOR TOMORROW

InterConnect

ATHLOS

NEAR

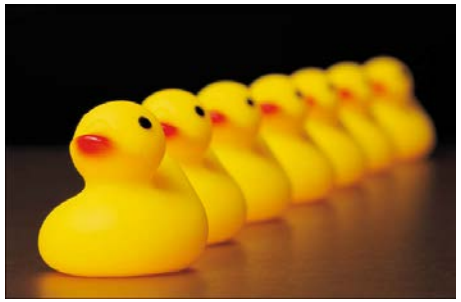
Prospective

HeLTI
HEALTHY LIFE TRAJECTORIES INITIATIVE

4 mother/child cohorts
Harmonized datasets: preconception, pregnancy follow-ups, at and after deli

Canadian Partnership for Tomorrow Project

5 adult cohorts (300 000 participants)
Harmonized datasets: baseline and follow-ups



Canadian Partnership for Tomorrow Project

Total:
307,017 Participants
30 to 74 years of age

BC	AB	ON	QC	ATL
28,808	39,116	163,463	42,393	33,529

1,636 core variables

55 study-specific datasets (18,291 study-specific harmonized variables)

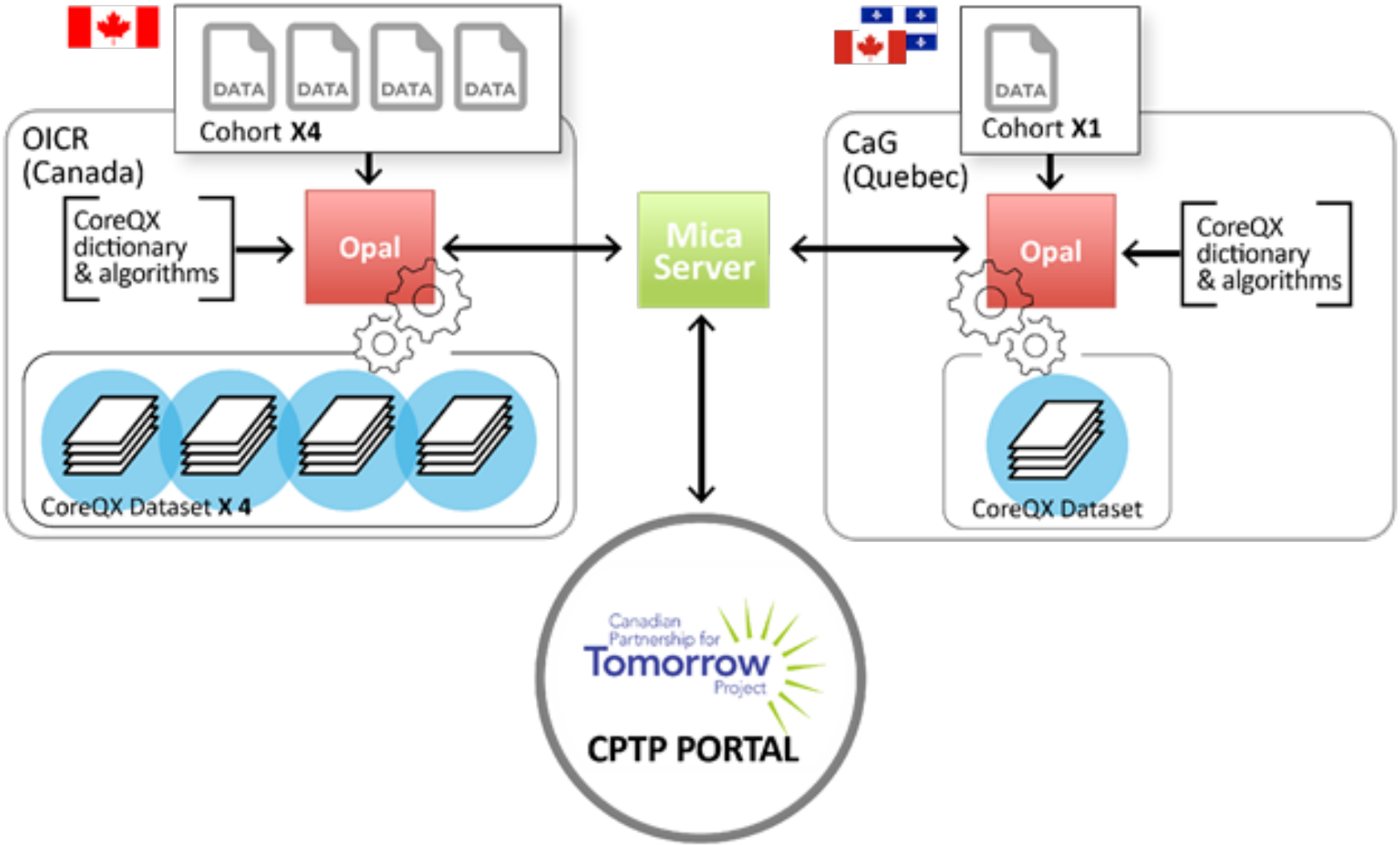
Define the research question(s)

Achieve consensus on compatible study designs, measures and collection procedures

Ensure quality and consistency of the study-specific data collections

Process study-specific data under the common data format and implement data sharing infrastructure

Provide access to users



Harmonization

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

Download

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

Complete - the study assessment item(s) (e.g. survey question, physical measure, biochemical measure) allow construction of the variable as defined in the dataset

Incomplete - there is no information or insufficient information collected by this study to allow the construction of the variable as defined in the dataset.

Showing 26 to 50 of 716 entries

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

A_SMK_CIG_CUR_FREQ

Overview

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

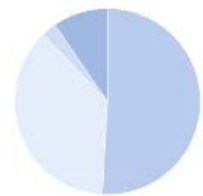
Classification

Areas of Information	
Lifestyle and health behaviours	Tobacco

Statistics

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

Valid values frequencies



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

Harmonization Algorithms

A_SMK_CIG_CUR_FREQ -- atlantic-path (1)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

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

A_SMK_CIG_CUR_FREQ -- atlantic-path (2)

Study variable(s)

[Current frequency of cigarette smoking]

Dataschema variable values

Value	Condition
0, 1, 2	If A_SMK_CIG_EVER = 1, ma
-7	If A_SMK_CIG_EVER = 0
	Missing

A_SMK_CIG_CUR_FREQ -- atp (1)

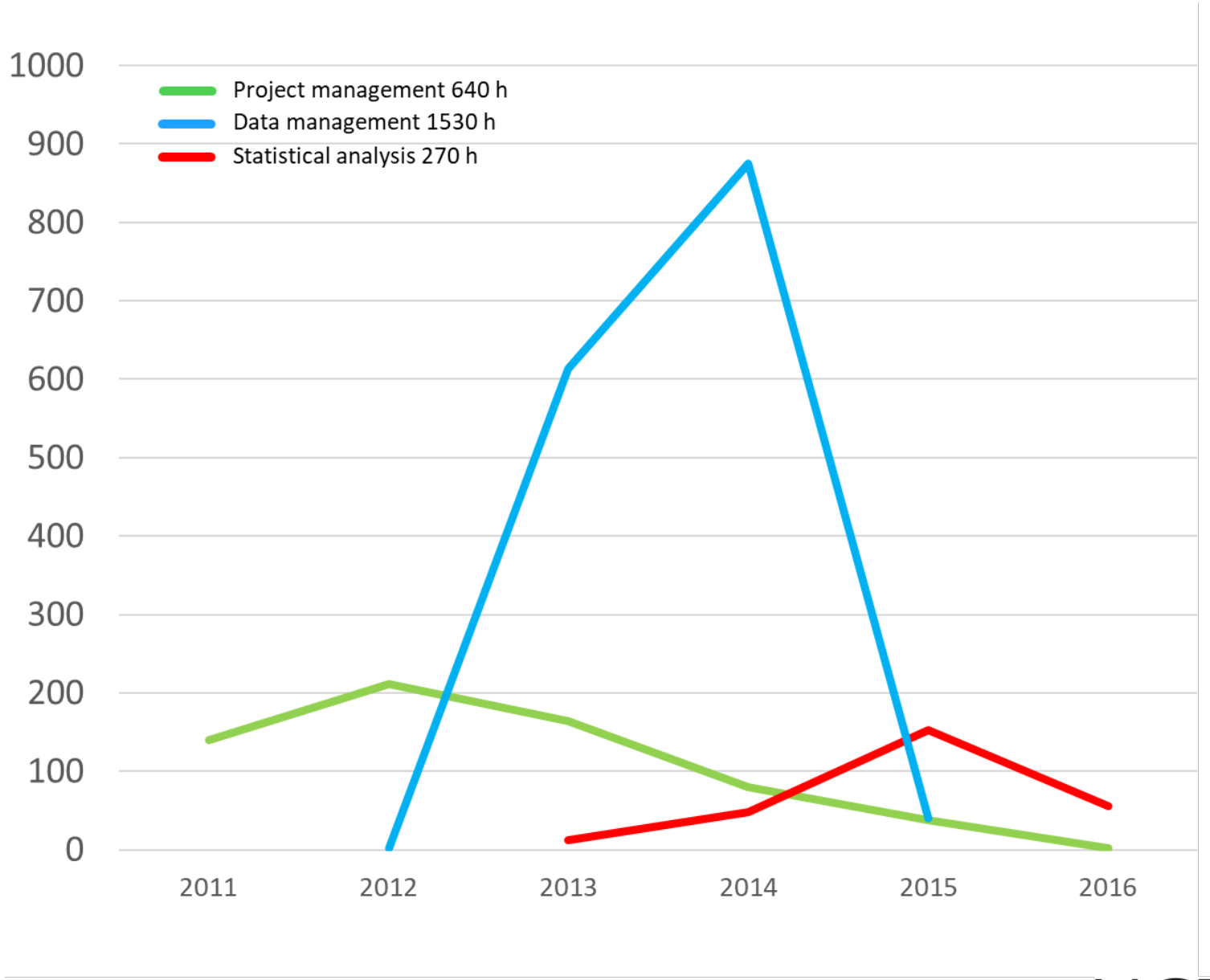
Study variable(s)

[Current frequency of cigarette smoking]

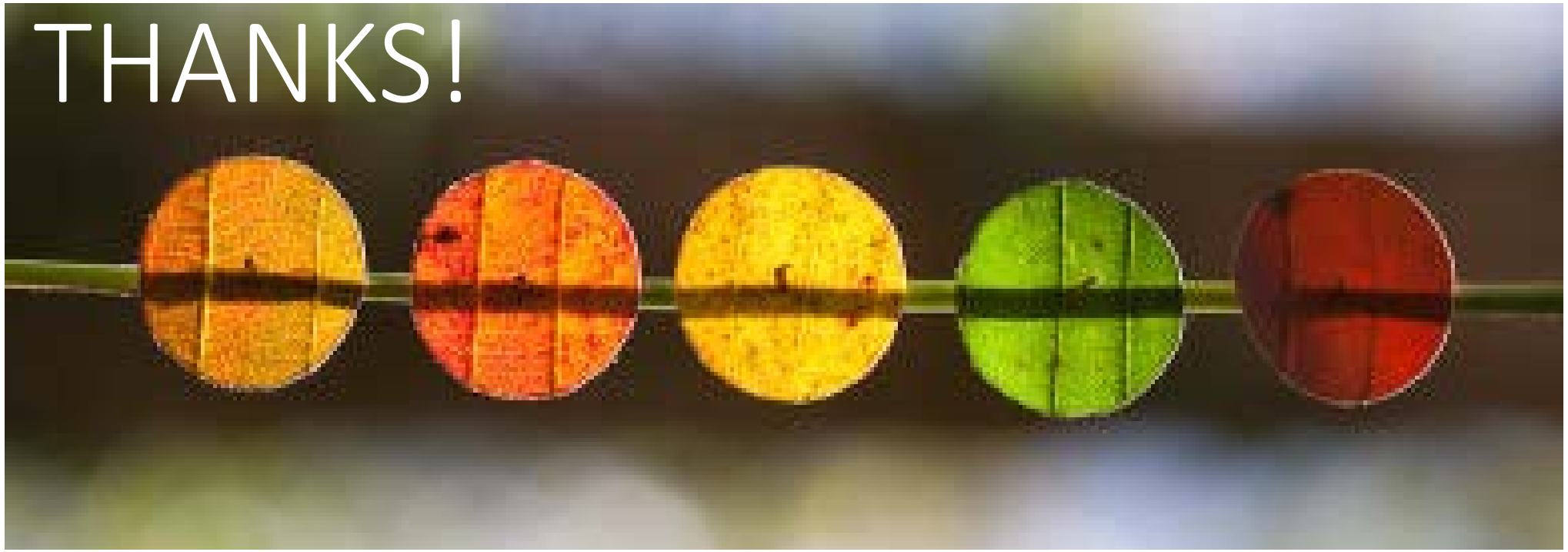
Dataschema variable values

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

Challenges and opportunities



THANKS!



Funding and support:

Centre universitaire de santé McGill Institut de recherche
McGill University Health Centre Research Institute
MUHC McGILL UNIVERSITY HEALTH CENTRE FOUNDATION
Économie, Innovation et Exportations Québec
INNOVATION.CA CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
CANADIAN PARTNERSHIP AGAINST CANCER
PARTENARIAT CANADIEN CONTRE LE CANCER
SEVENTH FRAMEWORK PROGRAMME
EUROPEAN COMMISSION
CIHR IRSC Canadian Institutes of Health Research | Instituts de recherche en santé du Canada
National Institute on Aging
canarie @25

www.maelstrom-research.org

Canadian Association for Population Therapeutics
Annual Conference
Toronto, ON
October 21, 2019

Data-Sharing, Reciprocity and Population Biobanks

Ma'n H. Zawati
Assistant Professor,
Executive Director
Centre of Genomics and Policy
McGill University

McGill

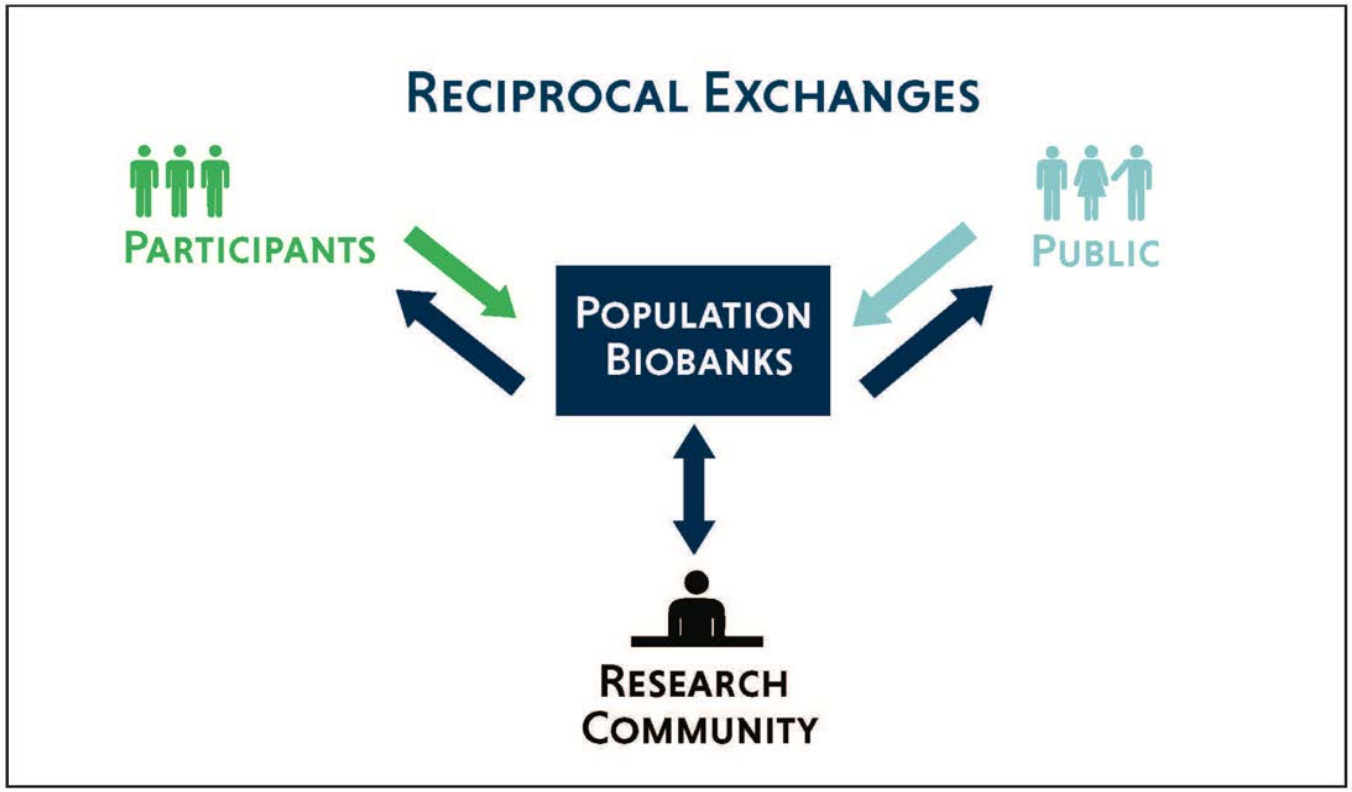
CGP Centre of Genomics and Policy
Centre de génomique et politiques

Prelude

http://whsc.emory.edu/_pubs/momentum/2006fall/biobank.html

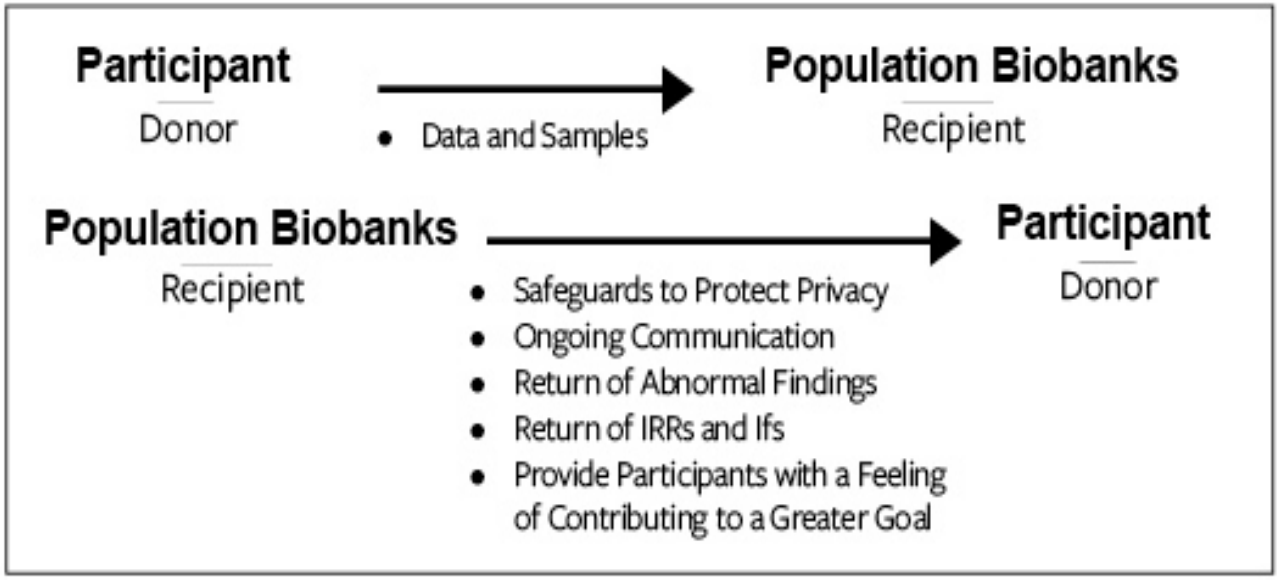
Prelude

Reciprocity in Population Biobanks

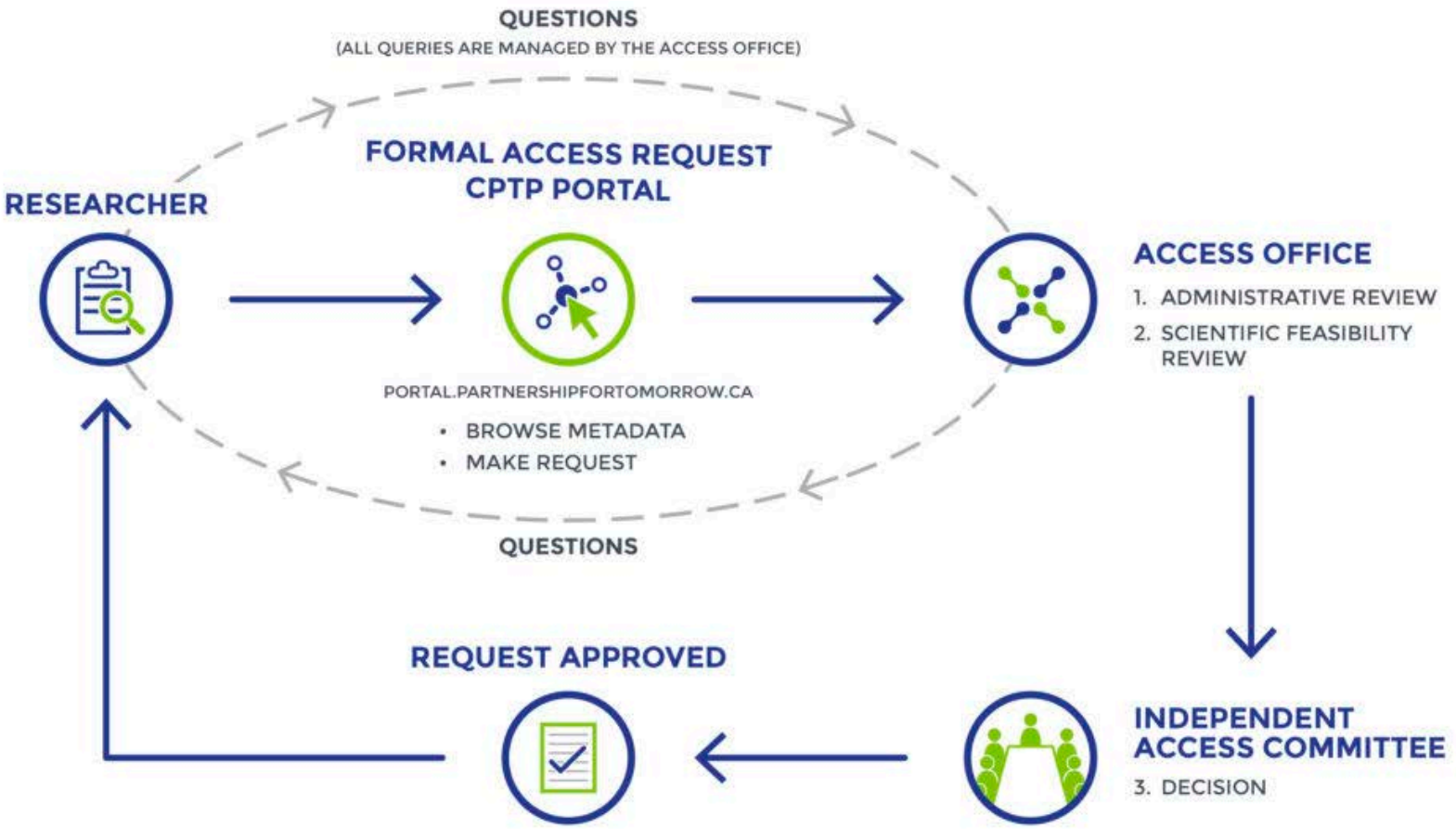


Reciprocity in Population Biobanks

- Population Biobank - Participant



CPTP Access Application Process



Conclusion: Maintaining the dynamic

Participant donates data and samples: altruism & reciprocity

⇩

Collaboration between researchers

⇩

Maximization of statistical power

⇩

Translation of knowledge

to the clinic

⇩

=

Better health for the population

Privacy, Data Sharing, and Access: A pan-Canadian perspective

Eric Sutherland, Executive Director
External Data and Information Governance

Canadian Institute for Health Information
Better data. Better decisions. Healthier Canadians.

Understanding Canada's data and information capabilities

Last year, CIHI engaged in a pan-Canadian 'listening tour' to discover practical steps to better use data and information to improve health.

Based on that work, CIHI published a Health Data and Information Capability Framework in September 2019.

The objective of the framework is to accelerate improvement of the appropriate use of data and information for a healthier Canada.

The Capability Framework provides:

- A checklist of items to consider for health information organizations,
- A common language to discuss capabilities within and across organizations,
- A tool for self-assessment to identify opportunities to improve, and
- A resource to harmonize practices across organizations.

CIHI

Reflections in today's context…

- Privacy
  - Optimizing outcomes while minimizing harm
  - Caldicott principles and Privacy by Design

- Data Sharing
  - Value of data is created when shared and used
  - Improving information (governance) literacy

- Access
  - Lessons from 1880's New York
  - Understanding public perceptions to be trustworthy