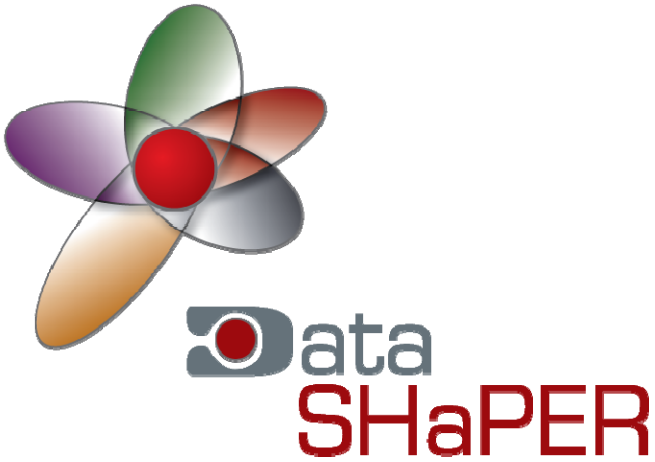
Hamonization Potential of 53 Large Population-Based Studies Using the DataSHaPER

Dany Doiron[1], François L'Heureux[1], Mylène Deschênes[1], Isabel Fortier[1]

[1] Public Population Project in Genomics (P³G)



Why harmonize data?

(1) Increase statistical power through data pooling,
(2) Maximize research potential of existing data,
(3) Facilitate multidisciplinary and cost-effective research strategies. There is growing interest amongst researchers and funders in the social and health sciences for data harmonization tools and methodologies.

How? DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) – www.datashaper.org

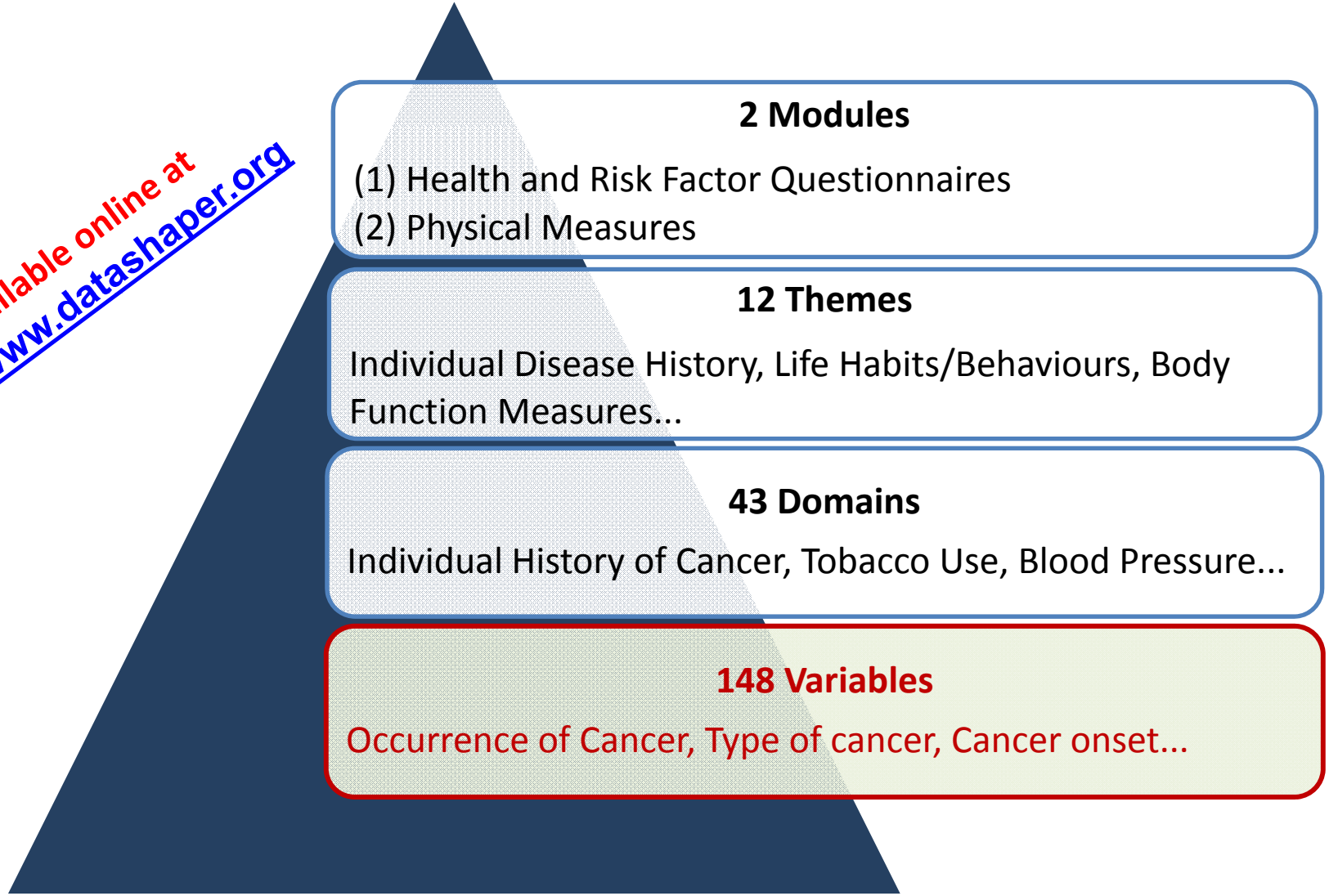
The DataSHaPER is both a scientific approach and a suite of practical tools. Its primary aims are to facilitate the prospective harmonization of emerging biobanks, provide a template for retrospective synthesis and support the development of questionnaires and information-collection devices, even when pooling of data with other biobanks is not foreseen.

It includes two primary components:

(1) The **DataSchema** documents and annotates sets of core variables, that each provide a concise but effective list of information to be harmonized in a specific scientific context.

(2) The **Harmonization Platform** provides a template for the formal estimation of the potential to synthesize information across networks of studies. Three-step process that entails: **(a)** the development of rules providing a formal assessment of the potential for each individual study to generate each of the variables in the DataSchema; **(b)** the application of these rules to determine and tabulate the ability of each study to generate each variable, thereby identifying the information that 'can' be shared; **(c)** where a variable can be constructed by a given study, the development and application of a processing algorithm enabling that study to generate the required variable in a common format and pool with other compatible studies.

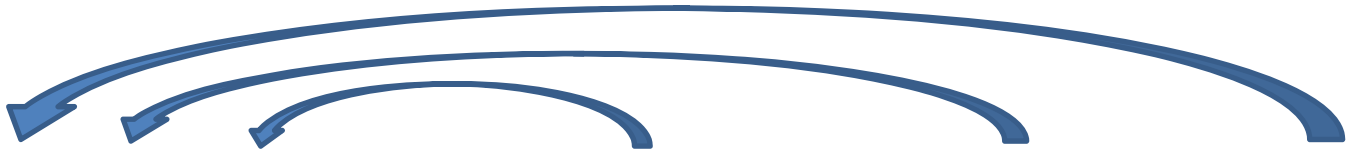
(1) Example of a DataSchema



(2a) Development of pairing rules

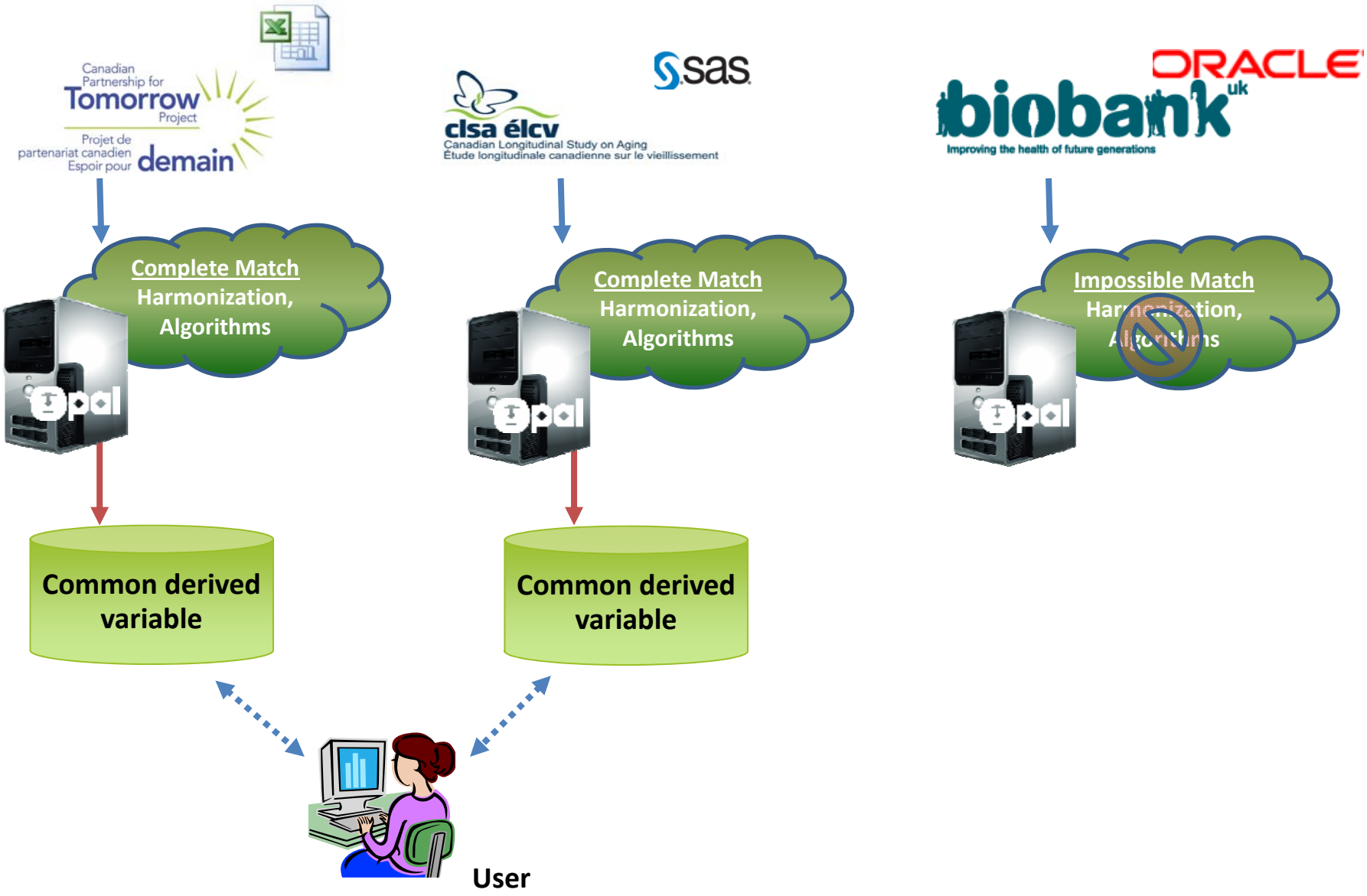
Pairing Result	Exemplar pairing rules for "Current quantity of red wine consumed"
Complete	• Number of drinks can be collected per day, week, month, etc; • Information can be collected for the entire week or for specific periods covering the whole week (week-days and week-end; Monday to Sunday, etc); • Question must target the current consumption (over the past 12 months or more contemporaneously)
Partial	• Categories are used for the number of glasses of red wine consumed
Impossible	• Impossible if only wine is mentioned without distinction between types of wine (red, white); • Impossible if relevant information is collected only for the consumption in the past (before the past 12 months); • Impossible if relevant information is collected at the same time for the current and the past consumption without distinction between the two

(2b) Application of pairing rules



	Study1	Study 2	Study 3
Cigarette	IMPOSSIBLE	PARTIAL	COMPLETE
Occurrence of Cancer	COMPLETE	COMPLETE	COMPLETE
Body Mass Index	PARTIAL	IMPOSSIBLE	IMPOSSIBLE
etc…			

(2c) Algorithm development and implementation (making use of Opal software)



P3G 53 Studies Harmonization Project

Objectives

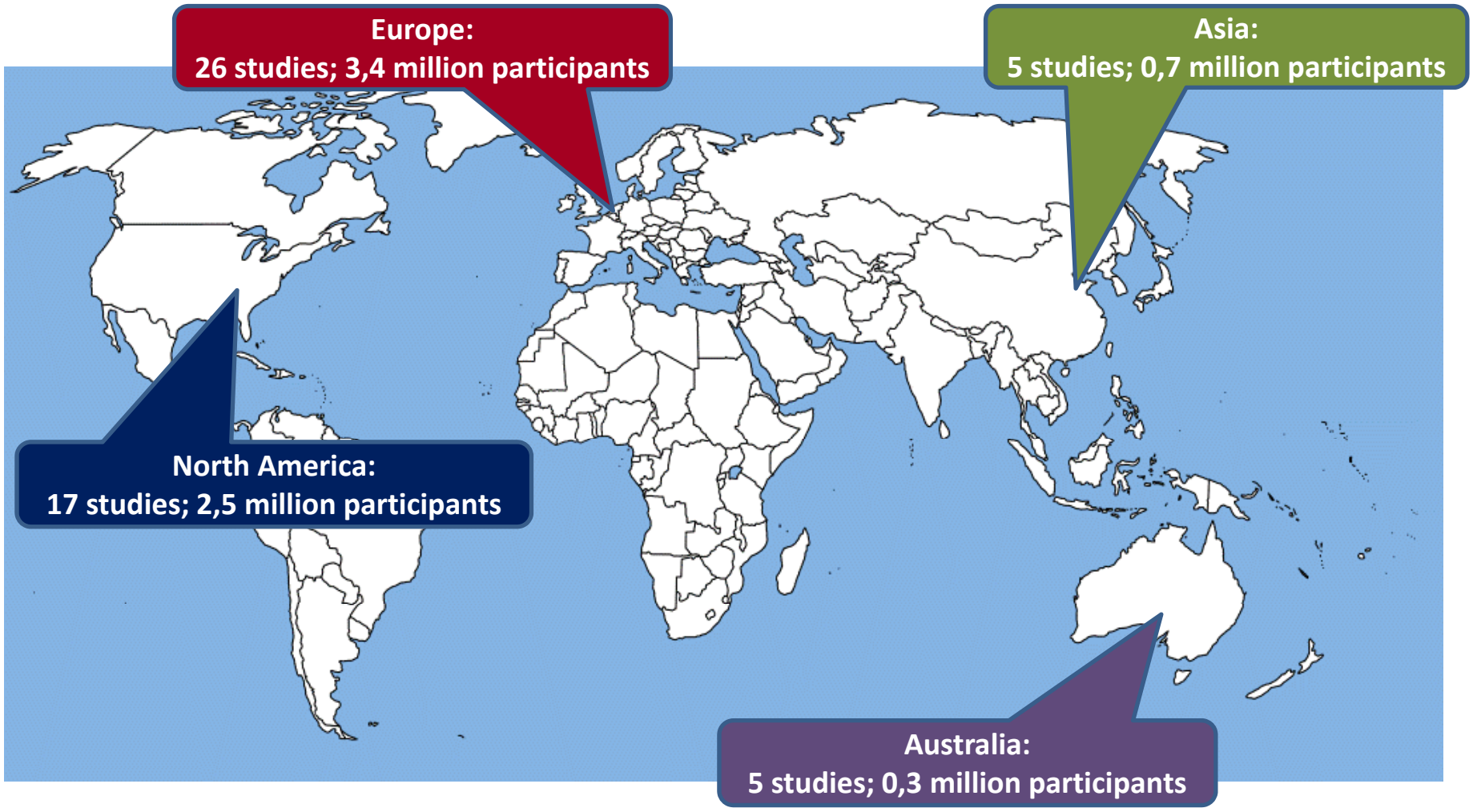
- Pilot the DataSHaPER approach
- Explore the harmonization potential between major population studies worldwide

Study Selection Criteria

- Recruit of planned to recruit at least 10 000 adult participants
- Collect biological samples enabling DNA extraction
- Collect comprehensive information on life habits, economic status, and health outcomes
- Provide access to the baseline questionnaire and standard operating procedures used

P³G 53 Studies Harmonization Project

6.9 Million potential participants in 20 countries



P³G 53 Studies Harmonization Project

Potential to synthesize data: Blood Pressure Case Study (n=53)

Co-analysed variable(s)	Total	Targeted number of Participants	Current number of participants
- Blood pressure	31 studies	3 731 000	1 816 000
- Blood pressure - Body Mass Index - Level of Physical Activity	28 studies	3 131 000	1 700 000
- Blood pressure - Body Mass Index - Level of Physical Activity - Use of alcohol - Quantity of cigarettes smoked - Post-secondary education	14 studies	2 019 000	1 335 000

P3G 53 Studies Harmonization Project

Results

- Out of all assessment items evaluated (148 variables for 53 studies), 38 percent could be harmonized
- Some key variables could be harmonized for most of studies: Occurrence of diabetes and Current use of alcohol (89%); Occurrence of Cancer, Occurrence of stroke, Employment status (81%); Type of cancer, Standing height and Weight (66%)
- Certain characteristics of variables (i.e. individual targeted, reference period) and of studies (i.e. mode of data collection, data collection start date) were associated with the potential for harmonisation

Conclusions

- DataSHaPER provides an effective and flexible approach for the harmonization of data across studies
- The approach offers the promise of collaborative projects and enhanced research potential though synthesized databases in in the health and social sciences
- To implement data synthesis, some additional scientific, ethico-legal, and technical considerations must be addressed

Data SHaPER

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