Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2013;42:1285–1299
doi:10.1093/ije/dys160

COHORT PROFILE

Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics

Philip Awadalla,[1]* Catherine Boileau,[2] Yves Payette,[2] Youssef Idaghdour,[1] Jean-Philippe Goulet,[2] Bartha Knoppers,[3] Pavel Hamet,[4] Claude Laberge[5] on behalf of the CARTaGENE Project[†]

[1]Department of Pediatrics, Faculty of Medicine, Université de Montreal, Montreal, Canada, [2]Centre de Recherche du Centre Hospitalier Universitaire (CHU) Sainte-Justine, CARTaGENE, Canada, [3]Department of Human Genetics, Faculty of Medicine, McGill University, Montreal, Canada, [4]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Canada and [5]Centre de Recherche du Centre Hospitalier Universitaire du Québec (CHUQ), Université Laval, Québec, Canada

*Corresponding author. CARTaGENE, 3333 Queen Mary road suite #100, Montreal (Québec), H3V 1A2, Canada.
E-mail: Philip.awadalla@umontreal.ca
[†]The members of the CARTaGENE Project are provided in the Supplementary Appendix.

Accepted 20 August 2012

The CARTaGENE (CaG) study is both a population-based biobank and the largest ongoing prospective health study of men and women in Quebec. In population-based cohorts, participants are not recruited for a particular disease but represent a random selection among the population, minimizing the need to correct for bias in measured phenotypes. CaG targeted the segment of the population that is most at risk of developing chronic disorders, that is 40–69 years of age, from four metropolitan areas in Quebec. Over 20 000 participants consented to visiting 1 of 12 assessment sites where detailed health and socio-demographic information, physiological measures and biological samples (blood, serum and urine) were captured for a total of 650 variables. Significant correlations of diseases and chronic conditions are observed across these regions, implicating complex interactions, some of which we describe for major chronic conditions. The CaG study is one of the few population-based cohorts in the world where blood is stored not only for DNA and protein based science but also for gene expression analyses, opening the door for multiple systems genomics approaches that identify genetic and environmental factors associated with disease-related quantitative traits. Interested researchers are encouraged to submit project proposals on the study website (www.cartagene.qc.ca).

Why was the cohort set up?

Much of what is known about the causes of chronic disorders comes from large epidemiological studies, especially prospective cohorts or population studies from the USA and Europe.[1–4] Because of the significant investment for initial recruitment, specimen collection and storage, cohorts of large sizes did not exist in Canada until recently.

CARTaGENE (CaG) is the largest ongoing prospective health study of men and women in Quebec, Canada. Developed since 2003, CaG obtained funding

from Genome Canada and Genome Quebec in 2007 to provide a scientific platform for the investigation of modifiable environmental and lifestyle factors and the genomic determinants of chronic diseases. Based on the results of public consultations, the community engagement in the project was promising when CaG was being set up.[1] A few years later, momentum was building in other Canadian provinces for developing population-based cohorts, and efforts were being put towards the creation of a prospectively and collaboratively designed national cancer cohort under the Canadian Partnership for Tomorrow Project (CPTP). CaG was instrumental in these efforts to ensure interoperability among regional initiatives while keeping a broad scientific focus on all major chronic diseases of public health relevance.

A total of 20 007 men and women were enrolled from August 2009 to October 2010 and are tracked based on linkage to governmental health administrative databases and direct reassessment. The content of the data encompasses a broad range of chronic conditions, clinical phenotypes and their potential determinants, making this a versatile research platform for the investigation of the role of genes, the environment and lifestyle on various health conditions.

CaG stands apart from a number of large population-based biobanks in many ways. A primary distinction is that in addition to collecting health-related information and biospecimens, it also collected substantial physiological measures. Secondly, the study was also designed to be representative of metropolitan areas to facilitate translation into health promoting policies and interventions. Lastly, via the Public Population Project in Genomics (P_3G), it was designed to maximize the ability to harmonize with Canadian and international cohorts.[2] Given these distinguishing features, CaG represents one of the most powerful tools for investigating determinants of multiple chronic disorders, providing research scientists substantial flexibility in designing research programmes so that they can move away from standard paradigms of genetic or epidemiological analyses while also having substantial numbers in Canada (through CPTP) and internationally (P_3G) to ensure sufficient power.

Who is in the cohort?

The CaG cohort consists of men and women aged between 40 and 69 years, residing in metropolitan areas representing a total of 55.7% of the Quebec population (Montreal, Quebec, Sherbrooke and Saguenay). Based on population density from the 2006 Census, expected numbers of recruits in each area were as follows: 15 271 for Montreal, 3224 for Quebec City, 804 for Sherbrooke and 701 for Saguenay. Figure 1 represents the actual numbers of recruits in each metropolitan area. Participants were randomly chosen to be broadly representative of the population based on provincial health insurance registries—FIPA files [fichier administratif des inscriptions des personnes assurées de la Régie de

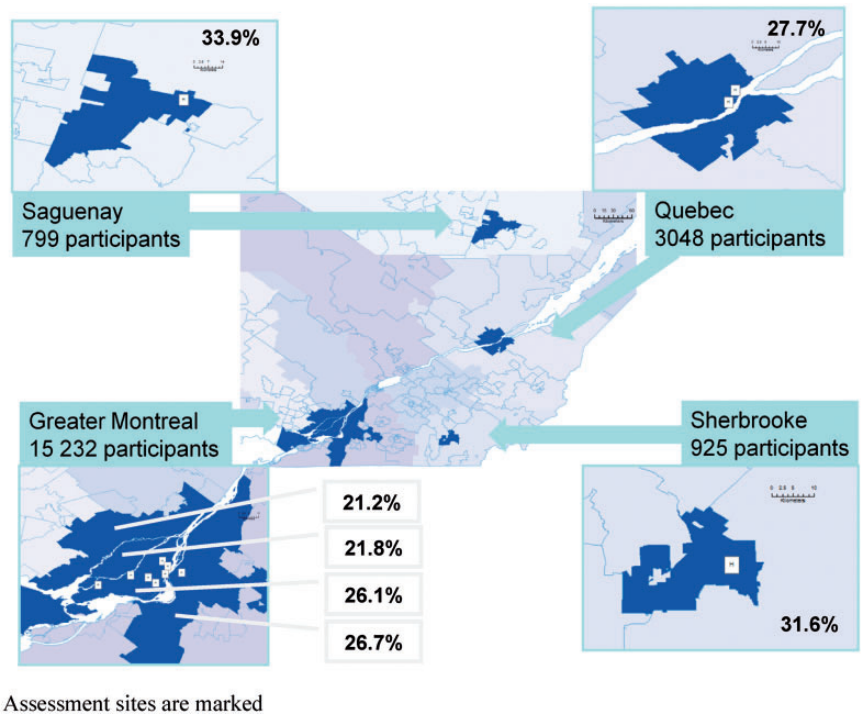

Figure 1 Geographic postal regions coverage, location of assessment sites and collaboration rates in the CARTaGENE study, Quebec, Canada

l'assurance maladie du Québec (RAMQ)]. Survey design was defined by two age groups, sex and forward sortation area (defined by 3-digit postal codes). Probability proportional to size was used to define quotas for each of these strata. Participants were excluded if they were not registered in the FIPA files, if they resided outside the selected regions, lived in First Nations Reserves or long-term health care facilities or were in prison.

Recruitment

Several strategies were used to obtain adequate response rates and minimize attrition during follow-up phases.[3,4] These included the following: (i) the use of a well-trusted governmental body to contact participants and handle identifying information (RAMQ); (ii) the use of systematic methods for contact, scheduling and sending reminders; and (iii) financial compensation of 45$.[5] The recruitment of participants was done through a call centre at the RAMQ such that no identifying information would be transferred to CaG. Information packages were first sent by mail, and potential participants were then contacted by telephone to enrol them and schedule an interview date in one of the clinical assessment sites. A total of 12 assessment sites participated in the project (Figure 1). Around 35% of the individuals in the FIPA extraction files did not have a phone number. Another 13–15% of the files had incorrect phone numbers. Only files with phone numbers were included in the extraction files as of January 2010 up to the end of recruitment.

Figure 2 represents the flow diagram of each stage of the recruitment process. After the initial phone contact, 38% of participants refused to be included, and 2% were found to be ineligible. A total of 20 007 persons came to the assessment site and signed a consent form. From these, 98% gave blood and consented to be re-contacted for follow-up. Only

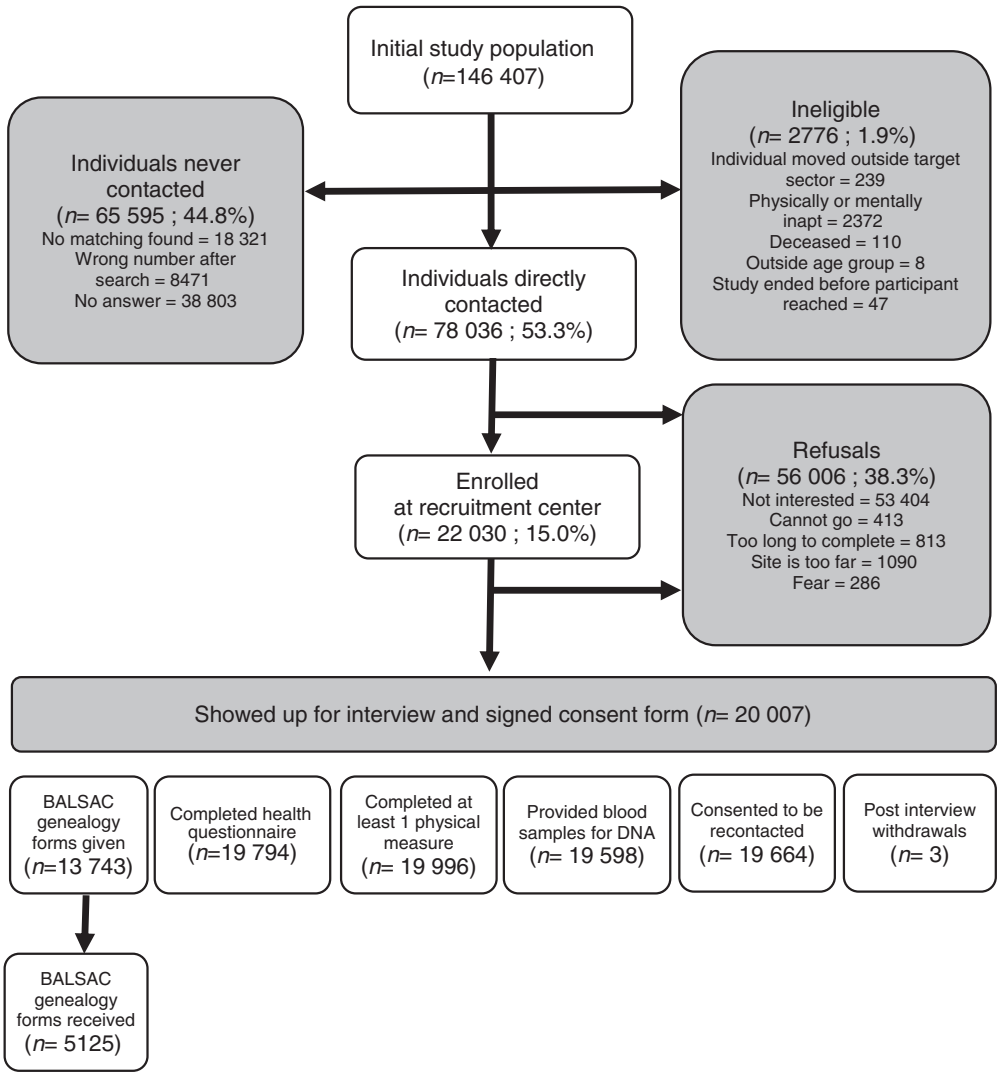

Figure 2 General recruitment process in the CARTaGENE study

three participants withdrew from the study after completion. A quarter (25.6%) of the participants (5125) completed and returned the genealogical questionnaire.

The overall collaboration rate was 25.6%. There were important regional variations in the collaboration rates, with Saguenay having the highest collaboration rate (33.9%) and Montreal northern suburbia having the lowest (21.8% for Laval and 21.2% for the North Shore). As typically observed in population health studies, men of the youngest age subset were the least likely to participate (23.3%) compared with older men (28.2%) or women of all ages (26.1%).

Questionnaire modules were completed by most participants (99%). Completion of physiological measurements was also high (>89%) except for spirometry (81.5%), which had strict contraindication criteria and Electrocardiogram (ECG) (39.8%), which was not implemented in all sites. The saliva tubes were used only if DNA could not be extracted from blood. Ninety-nine percent of participants answered at least one questionnaire or undertook one physical measurement.

Statistical power

CaG was primarily designed for studying the genetic and environmental determinants of quantitative traits (QTs) and the rate of change of QTs over time. The statistical power associated with the analysis of QTs is generally higher than that of dichotomous traits, and the required sample size can be smaller than that required for studying dichotomous traits.[6] The minimum effect sizes detectable with 80% power were computed for selected QTs under different scenarios, assuming that 18 000 subjects were recruited into CaG (Table 1). The data presented in this table demonstrate for instance, that with 18 000 participants, an increase or decrease of 1.24 mm/Hg in systolic blood pressure that may be attributable to a genetic determinant (modelled as SNPs using an additive genetic model) can be detected with 80% power at $P < 10^{-4}$ in a candidate gene study.

How often have they been followed-up?

CaG is a young cohort, and although 98% of participants have consented to being contacted in the future, no health reassessment has been conducted so far. Regular updates of health status will be collected using web-based questionnaires in the coming years. Participants may also be tracked for the next 50 years based on linkage to governmental health administrative databases. Data owned or managed by the RAMQ, including diagnosis, ambulatory and inpatient care, and prescribed medication may be used in conjunction with participant data. Plans to validate self-reported health status with information from government data are currently underway. CaG has also launched two surveys to enhance baseline

Table 1 Power profiles for selected physiological measurements in the CARTaGENE study

Scenarios	Genetic main effect (candidate gene study) $P < 10^{-4}$	Genetic main effect (GWAS) $P < 10^{-7}$	Environment main effect $P < 0.01$	G:E interaction (candidate gene study) $P < 10^{-4}$	G:E interaction (GWAS) $P < 10^{-7}$
Systolic BP: mean ≈ 126 mmHg, SD ≈ 18.2					
Common determinants	1.24 mmHg	1.42 mmHg	2.29 mmHg	4.05 mmHg	4.62 mmHg
Moderately common determinants	1.92 mmHg	2.19 mmHg	3.1 mmHg	6.21 mmHg	7.09 mmHg
Uncommon determinants	2.66 mmHg	3.04 mmHg	4.41 mmHg	10.96 mmHg	12.51 mmHg
Fat mass by bioimpedance: mean ≈ 22.38 kg, SD ≈ 10.0					
Common determinants	0.68 kg	0.78 kg	1.3 kg	2.2 kg	2.5 kg
Moderately common determinants	1.1 kg	1.2 kg	1.7 kg	3.4 kg	3.9 kg
Uncommon determinants	1.5 kg	1.7 kg	2.4 kg	6 kg	6.9 kg
Total cholesterol: mean ≈ 5.27 mmol/l, SD ≈ 0.986					
Common determinants	0.061 mmol/l	0.07 mmol/l	0.11 mmol/l	0.2 mmol/l	0.22 mmol/l
Moderately common determinants	0.093 mmol/l	0.11 mmol/l	0.15 mmol/l	0.29 mmol/l	0.34 mmol/l
Uncommon determinants	0.13 mmol/l	0.14 mmol/l	0.21 mmol/l	0.51 mmol/l	0.58 mmol/l
Glucose (in non-diabetics): mean ≈ 4.84 mmol/l SD ≈ 0.925					
Common determinants	0.057 mmol/l	0.066 mmol/l	0.1 mmol/l	0.18 mmol/l	0.21 mmol/l
Moderately common determinants	0.087 mmol/l	0.1 mmol/l	0.14 mmol/l	0.28 mmol/l	0.31 mmol/l
Uncommon determinants	0.12 mmol/l	0.14 mmol/l	0.2 mmol/l	0.48 mmol/l	0.54 mmol/l

Common determinants: minor allele frequency (under a dominant genetic model) = 0.3 and prevalence of at-risk environment = 0.5; moderately common determinants: minor allele frequency = 0.10 and prevalence of at-risk environment = 0.2; uncommon determinants: minor allele frequency = 0.05 and prevalence of at-risk environment = 0.1.
SD, standard deviation; GWAS, Genome-Wide Association Studies; G:E, Gene by environment.

Table 2 Selected socio-demographic characteristics of CARTaGENE participants compared with the general population (Statistics Canada, Census 2006[a])

Socio-demographic domains		CARTaGENE cohort *n* (%)	General population *n* (%)
Gender	Men	9689 (48.4)	816 580 (48.4)
	Women	10 315 (51.6)	870 365 (51.6)
Age (years)	40–44	2672 (13.4)	341 425 (20.2)
	45–49	4259 (21.3)	352 775 (20.9)
	50–54	4599 (23.0)	318 410 (18.9)
	55–59	3009 (15.0)	281 680 (16.7)
	60–64	2947 (14.7)	225 800 (13.4)
	65–69	2518 (12.6)	166 855 (9.9)
Regions	Montreal (Mtl)	8265 (41.3)	637 085 (37.8)
	Laval	1603 (8.0)	143 185 (8.5)
	North shore Mtl	2638 (13.2)	268 080 (15.9)
	South shore Mtl	2726 (13.6)	236 895 (14.0)
	Quebec city	3048 (15.2)	280 750 (16.6)
	Sherbrooke	925 (4.6)	58 920 (3.5)
	Saguenay	799 (4.0)	62 030 (3.7)
Country of birth	Canada	16 704 (83.5)	1 389 165 (82.3)
First language learned	French only	15 657 (78.6)	1 270 655 (75.3)
	English only	1251 (6.3)	144 985 (8.6)
	French and English	316 (1.6)	8865 (0.5)
	French and other	133 (0.7)	7275 (0.4)
	Not French	2554 (12.8)	255 265 (15.2)
Education	<High school	462 (2.3)	296 780 (17.6)
	High school	4725 (23.7)	654 590 (38.8)
	College	6326 (31.8)	378 735 (22.5)
	University	5899 (29.6)	261 370 (15.5)
	Graduate degree	2463 (12.4)	95 485 (5.7)
	No answer	40 (0.2)	n.a.
Working status	Employed	12 993 (65.5)	1 131 690 (67.1)
	Retired	4437 (22.4)	492 410 (29.2[a])
	Unable to work	806 (4.1)	36 380 (2.2)
	Unemployed	1048 (5.3)	26 470 (1.6)
	Home (caregiving)	562 (2.8)	n.a.
Marital status	Married	12663 (63.7)	1 150 270 (68.2)
	Single	2945 (14.8)	241 835 (14.3)
	Divorced/ separated/ widowed	4265 (21.5)	294 845 (17.5)
Living arrangement	Couple with at least one child	3388 (20.9)	641 325 (38.1)
	Single with at least one child	959 (5.9)	153 335 (9.1)
	Couple with no children	5389 (33.2)	513 910 (30.6)
	Adults living together with or without children	2497 (15.4)	74 625 (4.4[a])
	Living alone	4016 (24.7)	298 960 (17.8)
Household annual income	<25 000$	2466 (12.4)	102 095 (6.1)

(continued)

Table 2 Continued

Socio-demographic domains		CARTaGENE cohort *n* (%)	General population *n* (%)
	25 000–49 999$	4361 (21.9)	277 945 (16.5)
	50 000–74 999$	4059 (20.4)	304 235 (18.0)
	75 000–149 999$	5834 (29.3)	491 785 (29.2)
	>150 000$	1961 (9.9)	132 495 (7.9)
	No answer	1234 (6.2)	378 385 (22.4)
Self-reported ethnicity (available for only 10% of the participants)	White	1710 (83.3)	1 530 130 (90.7)
	Black	74 (3.6)	38 845 (2.3)
	Arab	77 (3.8)	21 490 (1.3)
	Latino	71 (3.5)	17 990 (1.1)
	Southeast Asian	22 (1.1)	15 245 (0.9)
	East Asian	22 (1.1)	
	West Asian	10 (0.5)	3560 (0.2)
	South Asian	9 (0.4)	16 615 (1.0)
	Jewish	8 (0.4)	390 (<0.1)
	Other	51 (2.5)	35 445 (2.1)

[a]The available statistic from the Statistics Canada 2006 census is not equivalent to the CARTaGENE categories. Number of persons rounded up to the nearest 5. Although the research and analysis are based on data from Statistics Canada, the opinions expressed by the authors do not represent the views of Statistics Canada.
n.a.: similar statistic from Statistics Canada not available.

data with information on nutrition and environment. These are still ongoing, and so far nearly 11 000 participants have completed a full residential and occupational history questionnaire, and >7500 have completed a food frequency questionnaire. Data from these surveys will be available in the fall of 2012.

What has been measured?

Interviews

The interview process is composed of five modules: (i) identification module; (ii) the consent module; (iii) the self-administered socio-demographic and lifestyle questionnaire; (iv) the interviewer-administered health questionnaire; (v) physical measurements and contra-indication questionnaire; and (vi) biospecimen collection. A genealogical questionnaire to be done at home was also included (http://balsac.uqac.ca/).

Questionnaire content

The core of the questionnaire is based on the P_3G DataSHaPER[2] and has been revised by more than 30 experts from various scientific fields. It has also been pre-tested with 223 respondents. Topics cover the following: socio-demographic factors, lifestyle, mental state, psychosocial environment, personal and family history of disease, health care utilization, medication use and women and men's reproductive health and history (Supplementary Appendix 1, available as Supplementary data as *IJE* online). Declared health conditions had to have been diagnosed by a physician. Scales included in the questionnaire were previously validated and extensively used, including the Patient Health Questionnaire,[7] the General Anxiety Scale,[8] the Job Content Questionnaire[9] and the International Physical Activity Questionnaire (www.ipaq.ki.se).[10]

Physical measures

Participants underwent non-invasive physical measurements that included anthropometry, body composition, physical strength, lung function, bone density, blood pressure, cardiac function, peripheral and central blood pressures and cognitive function (Supplementary Appendix 2, available as Supplementary data as *IJE* online).

Biochemical and haematological analysis

Haematological and biochemical tests include immediate assessment of blood cell counts and biochemical analysis (Supplementary Appendix 3, available as Supplementary data as *IJE* online). The latter was done in one central laboratory in Chicoutimi. Quality assurance tests in the optimization phase demonstrated that all parameters were measured with test–retest reliability well in excess of 90% (not shown).

Biospecimens

A total of 106.5 ml of blood was drawn in Vacutainers® tubes from each participant. Part of the samples was sent to clinical diagnostic laboratories for immediate haematological and biochemical analysis, whereas the rest was sent to the Genome Quebec and Saguenay

hospital/ECOGENE-21 Biobank (GQ Biobank) for storage. Whole blood, plasma, serum and urine were collected and stored in various conditions at the biobank. Whole blood (10 ml) was aliquoted in 384-well GenPlates by an automated liquid handler dispensing 10 µl in each well, and then stored at room temperature for extraction of DNA (average DNA concentration per GenPlate element is 60 ng/µl). Whole blood was also transferred in 0.5 ml straws (2 ml) and in 2-ml cryovials (5 ml) and stored at −176°C and −80°C for RNA extraction, respectively. A total of 19.5 ml of plasma was collected using Ethylenediaminetetraacetic acid (EDTA), Na-Citrate and PST tubes and transferred into 0.5-ml straws at −176°C. Plasma was also store in 2-ml cryovials at −80°C for toxicological analyses (supelco tubes) and −176°C for cell transformation and production of immortalized cell lines (peripheral blood mononuclear cell (PBMC) isolates). Serum (5 ml) was collected in SST tubes and transferred into 0.5-ml straws at −176°C for metabolomics studies. Red blood cells (4 ml) were collected and stored in 0.5-ml straws at −176°C. Finally, urine (12 ml) was stored in 2-ml cryovials at −80°C for biochemistry and analyses of the metabolome.

Biobanking

The Biobank has developed storage technologies and procedures that minimize costs, maximize versatility and monitor continuous quality assurance. A robust multi-level quality system monitors the integrity, duration, quality of life and security of stored samples. It provides short- and long-term storage with flexibility, regarding storage conditions and sample types, and has a capacity of tens of millions of samples. Liquid nitrogen cryocontainers for long-term storage, refrigerators and freezers for shorter-term storage are also part of the technological platform. Diverse automated sample handler systems were used to process and track hundreds of samples every day. Half of the samples are stored in a 'mirror site' with the same quality and safety standards to maximize their safety.

What has been found?

We assessed representativity by comparing socio-demographic characteristics of cohort participants with the general population using data from the 2006 Canadian Census. There is an overall concordance in the distribution of socio-demographic characteristics between the cohort and the general population (Table 2). The most striking difference is the fact that CaG participants are generally more educated. Surprisingly, ethnic minorities are also slightly over-represented amongst CaG participants.

The distribution of risk behaviours and chronic conditions reported by participants is outlined in Table 3. Because risk behaviours and disease tend to co-exist within individuals, we also looked at rates of

Table 3 Lifestyle and chronic conditions reported by CARTaGENE participants (missing excluded)

Lifestyle and chronic conditions		*n* (%)
Smoking status	Never smoked	8123 (40.9)
	Past-smoker	7950 (40.0)
	Occasional smoker	886 (4.5)
	Daily smoker	2911 (14.7)
Passive exposure to smoke at home	Never	15 585 (78.6)
	Once a month	1069 (5.4)
	Once a week	685 (3.5)
	Almost daily	961 (4.9)
	Daily	1527 (7.7)
Alcohol intake (past 12 months)	Never	2237 (11.2)
	< once a month	2541 (12.8)
	2–3 times a month	1415 (7.1)
	Once a week	4843 (24.3)
	2– 3 times a week	4325 (21.7)
	Almost every day	4493 (22.6)
Sleep	3 h	45 (0.2)
	4–6 h	4243 (21.4)
	7–8 h	13 496 (68.0)
	9–10 h	1906 (9.6)
	11+ h	157 (0.8)
Exposure to UV radiation	<30 min/week	3577(18.0)
	30–59 min/week	4977 (25.0)
	1–<2 h/week	5056 (25.4)
	2–<3 h/week	3517 (17.7)
	4–<5 h/week	1730 (8.7)
	5–<6 h/week	839 (4.2)
	>7 h/week	210 (1.1)
Physical activity (IPAQ)	Low	3377(17.4)
	Moderate	7618 (39.3)
	High	8377 (43.2)
Sedentary time	⩽3 h	4975 (25.8)
	4–5 h	5169 (26.8)
	6–8 h	5204 (26.9)
	9–10 h	2351 (12.2)
	⩾11 h	1616 (8.4)
Daily fruit, vegetable and juice consumption	None	130 (0.7)
	1	503 (2.5)
	2–3	3474 (17.5)
	4–6	8293 (41.8)
	7–10	6205 (31.3)
	⩾11	1241 (6.3)
Depression score	Mild	16 434 (82.6)

(continued)

Table 3 Continued

Lifestyle and chronic conditions		*n* (%)
	Moderate	2448 (12.3)
	Moderately severe	655 (3.3)
	Severe	363 (1.8)
Anxiety score	Mild	16811 (84.5)
	Moderate	2112 (10.6)
	Moderately severe	634 (3.2)
	Severe	343 (1.7)
Chronic conditions		
Endocrine/ Metabolic diseases	Type II diabetes	1365(6.9)
	Type I diabetes	127 (0.7)
	Thyroid disorder	2174(10.9)
	Hypercholesterolaemia	5670 (29.2)
Diseases of the circulatory system	Hypertension	4960 (25.0)
	Angina	654 (3.3)
	Stroke	329(1.7)
	Myocardial infarct	559 (2.8)
Diseases of the respiratory system	COPD	1002 (5.0)
	Asthma	2590 (13.0)
Diseases of the genitourinary tract	Renal failure	110 (0.6)
	Kidney stones	1099 (5.5)
	Kidney infection	329 (1.7)
Diseases of the musculoskeletal system and connective tissue	Osteoarthritis	3169 (16.1)
	Rheumatoid arthritis	564 (2.9)
	Osteoporosis	1294 (6.5)
Diseases of the eyes	Glaucoma	553 (2.8)
	Cataracts	1138 (5.7)
	Macular degeneration	125 (0.6)
Liver disorders	Cirrhosis	55 (0.3)
	Chronic hepatitis	213 (1.1)
	Cholescystitis	2035 (10.2)
Diseases of the digestive system	Stomach ulcer/acid reflux	4650 (23.4)
	Irritable bowel syndrome	891 (4.5)
	Colorectal polyps	884 (4.5)
	Diverticulitis	562 (2.8)
	Crohns disease	120 (0.6)
	Helicobacter pylori	551 (2.8)
Diseases of the skin and subcutaneous tissue	Eczema	2298 (11.6)
	Psoriasis	1205 (6.1)
	Other skin diseases	536 (2.7)
	Pollen	3515 (17.7)
Allergies and food intolerances	Medication	2960 (14.9)
	Metal	274 (1.4)
	Latex	193 (1.0)
	Insect bite	465 (2.3)
	Food	1378 (6.9)
	Animals	2393 (12.1)
	Other	2292 (11.5)
Autoimmune diseases	Systemic lupus erythematosus	55 (0.3)
Depression	Minor depression	2694 (14.8)
	Major unipolar depression	80 (0.4)
	Major bipolar depression	272 (1.5)
Diseases of the nervous system	Epilepsy	112 (0.6)
	Migraines	2171 (10.9)
	Multiple scterosis	82 (0.4)
	Parkinsons	24 (0.1)
	Schizophrenia	75 (0.4)
Neoplasms (initial and subsequent)[a]	All cancers	1622 (8.2)
	Breast	351 1.90
	Skin	334 (1.80)
	Prostate	178 (1.0)
	Cervix	131 (0.70)
Comorbidities (total number of diseases)	0	4542 (22.8)
	1	5255 (26.3)
	2	4045 (20.8)
	3	2735 (13.7)
	3+	3370 (16.9)

[a]Participants with other types of cancer are excluded from the denominator. UV, ultra-violet.

multimorbidity and clustering of risk behaviours in the cohort (Table 4). Daily smoking (14.7%) and low physical activity (17.4%) were the most commonly reported unhealthy habits, either as single or combined behaviours (2.0%). Nevertheless, co-occurrence of unhealthy habits was low, and most of the participants did not seem to have high-risk profiles. High cholesterol (29%) and high blood pressure (25%) were the most commonly reported health conditions and also the most common comorbidities. Rates of multimorbidity were high (30.6% overall), especially in the oldest age group where 54% reported three or more chronic conditions (not shown). This rate far exceeded the prevalence of each individual disease. Comorbidity is associated with a decline in many health outcomes and increases in mortality and use of health care resources.[11–14]

Table 4 Occurrence of the combinations of four of the most common unhealthy habits and most common single or combinations of two or three chronic conditions reported by CARTaGENE participants ($n = 20\,004$, including missing)

Number of unhealthy habits	Unhealthy habit[a]			*n* (%)
None				12 295 (61.5)
One	Low physical activity			2290 (11.4)
	Smoking			1725 (8.6)
	Drinking			1152 (2.8)
	Low fruit and vegetable			804 (4.0)
Two	Low physical activity	Smoking		400 (2.0)
	Smoking	Low fruit and vegetable		363 (1.8)
	Low physical activity	Low fruit and vegetable		260 (1.3)
	Low physical activity	Drinking		191 (1.0)
	Smoking	Drinking		183 (0.9)
	Drinking	Low fruit and vegetable		72 (0.4)
Three	Low physical activity	Smoking	Low fruit and vegetable	147 (0.7)
	Low physical activity	Smoking	Drinking	54 (0.3)
	Smoking	Drinking	Low fruit and vegetable	30 (0.1)
	Low physical activity	Drinking	Low fruit and vegetable	29 (0.1)
Four	Low physical activity	Smoking Drinking	Low fruit and vegetable	6 (0.0)
Number of conditions	**Chronic conditions[b]**			***n* (%)**
One	High cholesterol			929 (4.64)
	Stomach disease			704 (3.52)
	High blood pressure			637 (3.18)
	Arthritis			561 (2.80)
	Asthma			454 (2.27)
Two	High cholesterol	High blood pressure		343 (1.71)
	High cholesterol	Stomach disease		212 (1.06)
	High cholesterol	Arthritis		157 (0.78)
	Stomach disease	Arthritis		146 (0.73)
	High blood pressure	Arthritis		143 (0.71)
Three	High cholesterol	High blood pressure	Diabetes	90 (0.45)
	High cholesterol	High blood pressure	Stomach disease	85 (0.42)
	High cholesterol	High blood pressure	Arthritis	84 (0.42)
	High cholesterol	Upper GI tract	Arthritis	68 (0.34)
	High blood pressure	Upper GI tract	Arthritis	49 (0.24)

[a]Smoking: daily smoking. Drinking: >10 drinks per week. Low physical activity—Low IPAQ score: ⩽3 days of vigorous activity of at least 20 min/day or <5 days of moderate activity and/or walking for ⩾30 min/day or <5 days of any combination of walking, moderate or vigorous activity of <600 MET-min/week. MET = Metabolic equivalent of task. Low fruit and vegetable consumption: ⩽2 servings per day.
[b]Self-declared chronic conditions that have been diagnosed by a physician. Stomach disease includes gastric reflux and ulcers.

We took a closer look at the distribution of cancer and heart diseases in different subsets. Together, these diseases account for >50% of deaths in Canada.[15] As expected, cancer and heart diseases were found to be lower in subgroups with higher education and higher income (Figure 3). Finally, Spearman correlations were computed to investigate the typologies of comorbid conditions (Figure 4).[14] Our observations were as expected. Correlations among cardiovascular conditions were the highest. Similarly, anxiety

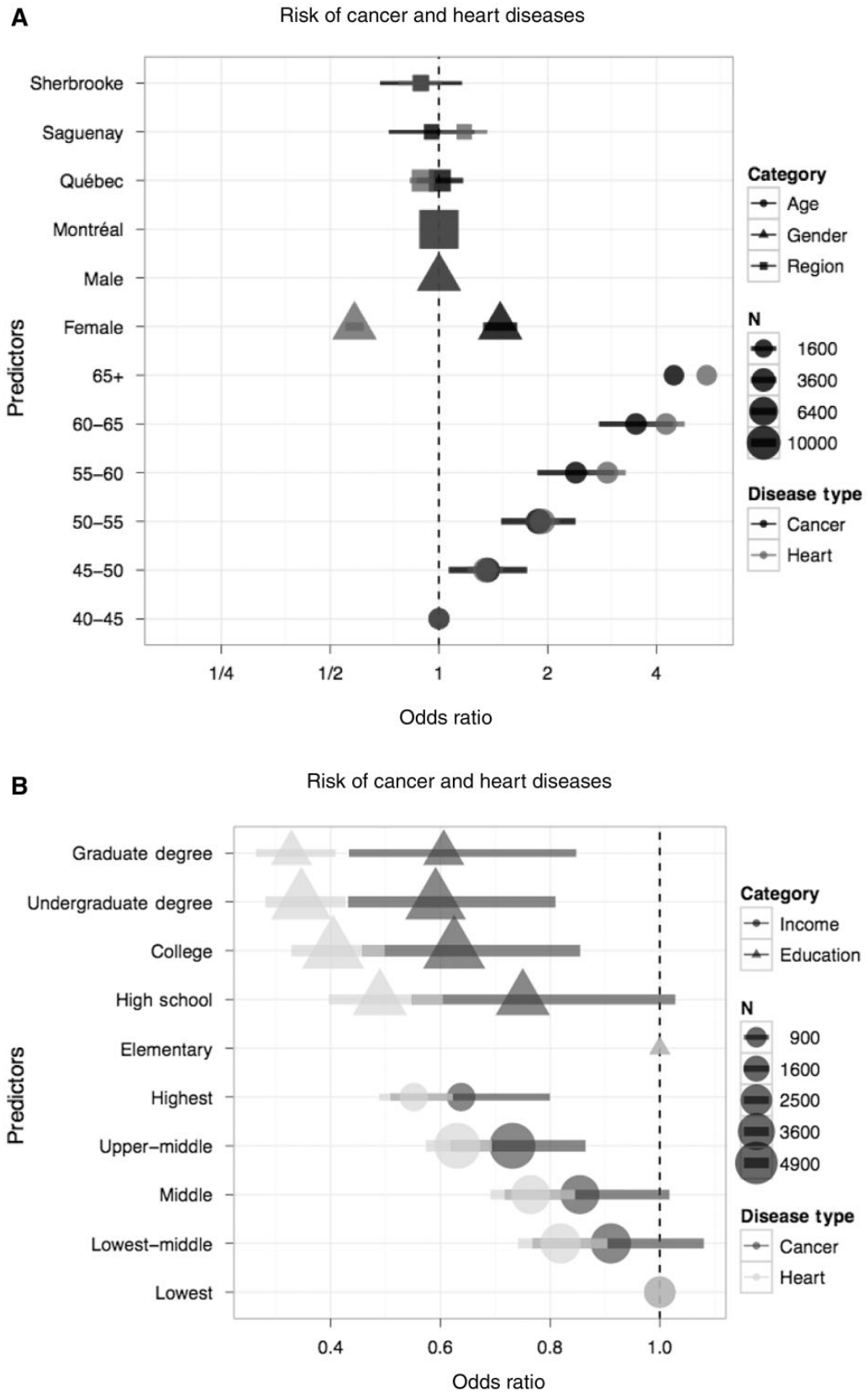

Figure 3 Odds ratio for all cancers and heart diseases in the CARTaGENE study. (**A**) Odds by regions, gender and age groups and (**B**) Odds by education and income

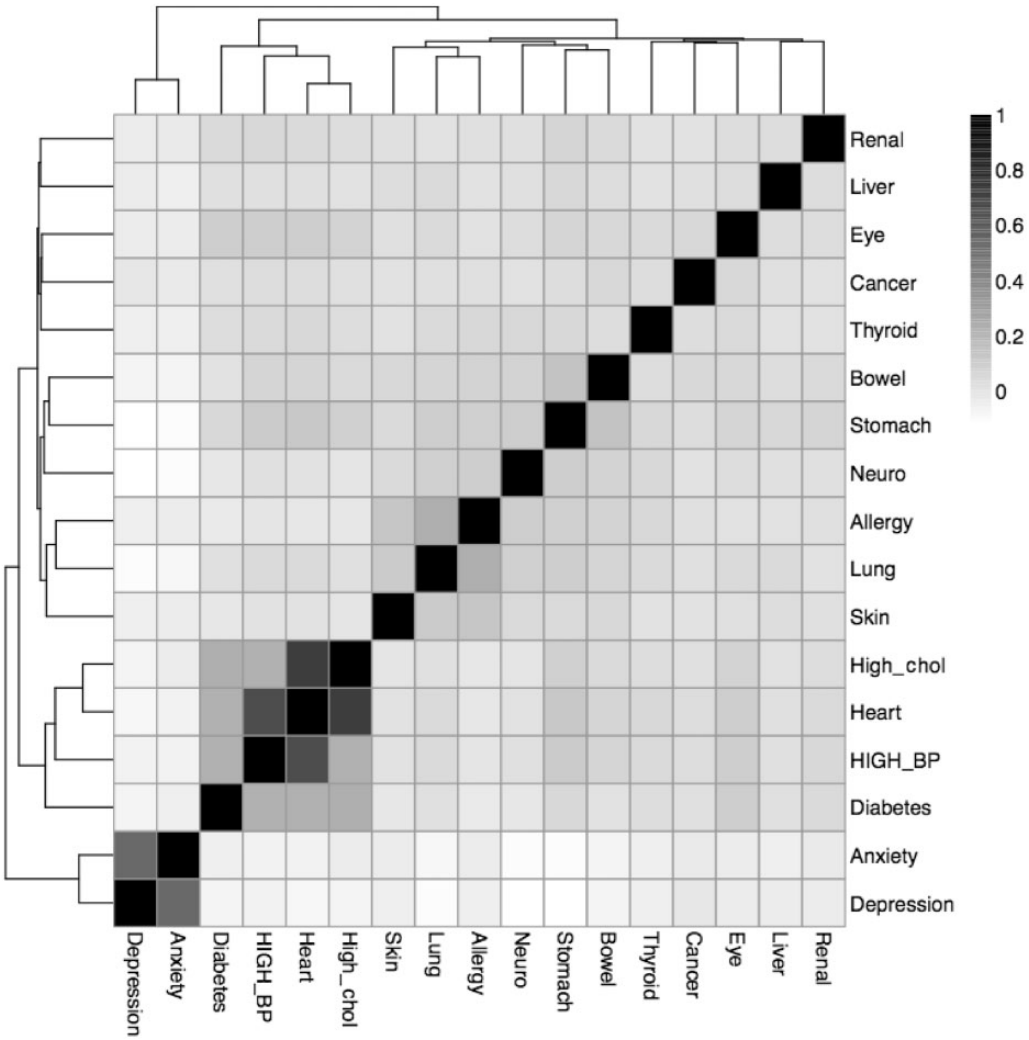

Figure 4 Correlations heatmap among major chronic diseases evaluated in the CARTaGENE study

symptoms were closely correlated related with depressive symptoms. Moderate correlations were also observed between conditions affecting the digestive system (stomach and bowel) and between allergies and diseases of the respiratory system.

Table 5 shows the means of physical measurements made on participants. As expected from reported health conditions, high proportions of men and women had suboptimal arterial fitness, low pulmonary capacity and were overweight. Suboptimal health conditions in low-risk-taking individuals may be the result of increased awareness. Nevertheless, currently, cohort participants' health status is a reflection of their habits earlier in their life.

What are the main strengths and weaknesses?

CaG has prioritized depth of information and versatility of use over widespread geographic coverage and large numbers, making it an attractive platform for researchers and public health practitioners. Health parameters collected during physical assessment provide a series of valuable QT phenotypes that are meaningful in their own right as complex traits that are worthy of aetiological study or QTs that reflect intermediate traits that lie on the causal pathways leading to a number of complex binary traits that are of considerable scientific interest. QTs on which CaG focuses require smaller sample sizes for adequate power, thus removing any issues associated with sample size for QT mapping.

CaG is designed to maximize the ability to harmonize with other international large-scale cohorts through the P_3G platform. Moreover, prospective harmonization of content and methods was applied in the design of the five cohorts within the CPTP recruiting hundreds of thousands of participants across Canada, thus enhancing its potential to generate a high-quality synthesized database for the study of complex chronic diseases.[16] Combining data from

Table 5 Mean, SD and comparative values for selected physical, haematological and biochemical measures in the CARTaGENE study, Quebec, Canada

Measures	Means				Thresholds	
	All	SD	Men	Women	Value	*n* (%)
Peripheral blood pressure						
Systolic blood pressure (mmHg)	123.9	15.7	128.3	119.8	⩾140	2930 (14.9)
Diastolic blood pressure (mmHg)	73.7	10.3	75.5	72.0	⩾90	1307 (6.7)
Central blood pressure						
Aortic pulse height	39.3	10.1	39.8	38.8	⩾40	7282 (40.5)
Aortic augmentation index (%)	27.1	11.3	22.5	31.6	⩾28.01	8640 (48.2)
Systolic aortic pressure (mmHg)	114.2	15.0	116.5	112.0	⩾117.01	6969 (38.8)
Anthropometry[a]						
Waist to hip ratio	0.9	0.1	1.0	0.9	⩾1.0	7166 (37.0)
BMI (kg/m^2)	27.5	5.3	28.0	27.1	⩾25.0	12 277 (65.7)
% body fat	30.7	8.8	25.4	35.5	W ⩾ 34.01	11 850 (63.4)
					M ⩾ 22.01	
Lung function[b]						
Forced vital capacity FEV (l)	3.8	1.1	4.5	3.2	W ⩽ 3.69	11 642 (58.2)
					M ⩽ 4.79	
Forced expiratory volume FEV1 (l)	3.0	0.8	3.5	2.5	⩽2.4	3826 (23.4)
FEV1/FVC ratio (FEV %)	78.8	8.6	78.0	79.5	⩽69.99%	1795 (11.0)
Bone density						
T score[c]	0.2	1.2	0.3	0.1	⩽−2.51	97 (0.5)
Risk of fracture	1.1	0.4	1.0	1.2	3	311 (1.9)
Haematology						
White blood cells (10^9 cells/l)[d]	6.9	2.2	6.8	7.0	⩽4.09	573 (3.1)
Red blood cells (10^{12} cells/l)[d]	4.6	3.8	4.8	4.3	W ⩽ 4.4	9145 (48.7)
					M ⩽ 4.49	
Haemoglobin (g/dl)[d]	139.5	12.5	147.3	131.9	⩽129	4079 (21.7)
Haematocrit[d]	0.4	0.0	0.4	0.4	W ⩽ 0.349	1756 (9.4)
					M ⩽ 0.399	
Mean corpuscular volume (fl)[d]	90.7	4.5	90.7	90.8	⩾96.01	1779 (9.5)
Mean corpuscular haemoglobin (pg)[d]	30.9	1.8	31.0	30.8	⩾33.21	1159(6.2)
Mean corpuscular haemoglobin concentration (g/l)[e]	340.2	7.8	341.7	338.7	⩾361	72 (0.4)
Red cell distribution width (RDW)[e]	13.4	1.1	13.4	13.4	⩾16.51	232 (1.2)
Platelets (10^9 cell/l)[e]	243.8	58.5	228.6	258.4	⩽149	520 (2.8)
Lymphocytes (10^9 cell/l)[e]	2.0	1.0	1.9	2.0	⩾4.01	114 (0.7)
Monocytes (10^9 cell/l)[e]	0.5	0.2	0.5	0.5	⩾1.81	2 (0.0)
Neutrophils (10^9 cell/l)[e]	4.2	1.4	4.1	4.3	⩾7.01	710 (4.2)
Eosinophils (10^9 cell/l)[e]	0.2	0.1	0.2	0.2	⩾0.41	540 (3.2)
Basophils (10^9 cell/l)[e]	0.0	0.0	0.0	0.0	⩾0.31	7 (0.0)
Biochemistry						
Glucose (mmol/l)[d]	5.8	2.3	6.0	5.6	⩾6.11	4694 (24.8)
Uric acid (μmol/l)[d]	301.7	80.3	343.5	261.5	⩾451	781 (4.2)
Creatinine (μmol/l)[f]	77.2	18.8	86.7	68.2	⩾16	429 (2.2)
Albumin (g/l)[f]	42.7	3.0	43.2	42.2	⩽33	46 (0.2)

(continued)

Table 5 Continued

Measures	Means				Thresholds	
	All	SD	Men	Women	Value	*n* (%)
Total calcium (mmol/l)[d]	2.4	0.1	2.4	2.4	⩽2.16	195 (1.0)
Sodium (mmol/l)[f]	139.0	2.4	139.0	138.9	⩾146	79 (0.4)
Potassium (mmol/l)[f]	4.3	0.5	4.3	4.3	⩽3.49	330 (1.7)
Chloride (mmol/l)[f]	103.7	2.6	103.7	103.7	⩾111	72 (0.4)
Aspartate aminotransferase (AST) (IU/l)[g]	24.8	10.4	26.7	23.0	⩾39	908 (4.7)
Alanine aminotransferase (ALT) (IU/l)[g]	25.2	14.9	29.2	21.4	W ⩾ 41	786 (4.0)
					M ⩾ 61	
Gamma glutamyltransferase (GGT) (IU/l)[g]	27.4	31.5	32.9	22.1	⩾65	1044 (5.4)
Triglycerides (mmol/l)[d]	1.9	1.3	2.2	1.6	⩾2.16	5628 (29.0)
Total cholesterol (mmol/l)[h]	5.1	1.0	5.0	5.3	⩾5.21	8570 (44.1)
HDL-cholesterol (mmol/l)[e]	1.2	0.4	1.1	1.4	W ⩾ 2.40	6475 (33.3)
					M ⩾ 0.91	
LDL-cholesterol (mmol/l)[e]	3.0	0.9	3.0	3.1	W ⩾ 4.96	2937 (15.6)
					M ⩾ 3.41	
Glycated haemoglobin HbA1C (0 to 1)[e]	0.1	0.0	0.1	0.1	⩾0.0611	2156 (11.2)
Thyroid stimulating hormone(TSH) (UI/l)[e]	1.9	1.8	1.9	1.9	⩾6.01	274 (1.4)
Free-T4 (pmol/l)[e]	11.5	2.6	11.5	11.4	⩾18.21	363 (1.9)

Reference values depend on diagnostic method used, as well as demographic characteristics of the reference population.
Value of threshold tested based on normal values (see footnotes below for references).
[a]Carmel M *et al*. Normal vascular aging: differential effects on wave reflection and aortic pulse wave velocity. *J Am Coll Cardiol* 2005;46:1754–60.
[b] Salzman SH. Pulmonary function testing: tips on how to interpret the results. *J Resp Dis* 1999;20:812.
[c]Kanis JA, Glüer CC. An update on the diagnosis and assessment of osteoporosis with densitometry. Committee of Scientific Advisors, International Osteoporosis Foundation. *Osteoporos Int*. 2000;**11**:192-202; Siminoski K, Leslie WD, Frame H, Hodsman A, Josse RG, Khan A, Lentle BC, Levesque J, Lyons DJ, Tarulli G, Brown JP. Recommendations for bone mineral density reporting in Canada: a shift to absolute fracture risk assessment. *J Clin Densitom*. 2007;**10**:120-3
[d]Henry JD (ed). Clinical diagnosis and management by laboratory methods. 19th edn. Philadelphia, PA: WB Sounders Co., 1996.
[e]Services de laboratoire du CHUM and Centre de Santé et des Services Sociaux de Saguenay.
[f]http://www.jrank.org/health/pages/5108/Biochemical-reference-values-blood.html (Accessed May 2008).
[g]Services de laboratoire de Calgary. WebLink http://www.calgarylabservices.com/LabTests/ (Accessed May 2008).
[h]Programme national de cholestérol (recommendation).
HDL, high density lipoprotein; LDL, low density lipoprotein.

multiple cohorts offers great potential when investigating the interactions between genetic and environmental factors, which underlie almost all human diseases.[17] Studies of genomic association typically require several tens of thousands of cases to properly investigate gene–gene or gene–environment interactions.[6] Given this feature, CaG is potentially well-positioned for GWAS on QTs. In such studies, effect sizes are typically very small,[18,19] and sample sizes must necessarily be very large.

The collaboration rate of the study was 25.6%. Although this might be seen as low, it is high relative to other population programmes such as the UK Biobank.[20] Long-term follow-up potential is also high (98%). Information on health outcomes and use of health care services, including prescribed medication, may also be obtained from government health and administrative databases. The combined use of population surveys and health services registers is a powerful tool for public health, as their respective limitations and assets can balance each other. CaG also has a broad consent that foresees use in future unspecified research, linkage to health and administrative databases and an access oversight committee that evaluates requests for samples and data from researchers from Canada and abroad.

Finally, the richness of having detailed genealogical information through the BALSAC project on a large proportion of the Quebec founder population is highly attractive for population geneticists, especially given that an estimated 3380 founders contribute ∼70% of the present gene pool.[21,22]

Population representativity was targeted to ensure the use of the platform for public health research. The obvious benefit of designing CaG as such is to provide unbiased estimations of risk exposure and prevalence of health outcomes. The obvious pitfall is that it limits probable representation of minority subgroups and does

not allow self-enrolment of potentially long-term members. Although significant efforts were made to ensure representativity, differences were found between CaG participants and the general population. Proper weighting and post-stratification is required for further population prevalence estimations.[23]

Can I get hold of the data? Where can I find more?

CaG is a public resource that may be used for health research. Participants have consented to the use of their data and samples for research at the national or international level, and access policies are in place. In addition to obtaining ethics approval, researchers must submit their request for data and samples to an independent Sample and Data Access Committee (SDAC). The SDAC, in collaboration with the scientific management of CaG, also defines what results or data ought to be returned to the project.[24] Interested researchers are encouraged to submit project proposals on the study website (www.cartagene.qc.ca). Further information may be obtained from the access coordinator (access@cartagene.qc.ca).

Conclusion

The CaG study has a number of distinguishing features that makes it an important research platform: (i) it is broadly representative of the population; (ii) it is harmonized with other cohorts; (iii) it includes detailed genealogical information on >25% of participants; and (iv) it has an explicit focus on intermediate/quantitative deep phenotyping and physical measurements.

Supplementary Data

Supplementary Data are available at *IJE* online.

Funding

This work was supported by Genome Canada supporting P_3G and CARTaGENE, as well as a Genome Québec Recruitment Award to P.A.

Acknowledgements

Most importantly, we are grateful to all participants who have kindly provided information and samples and who believe in this project. We would also like to thank the dedicated staff at the RAMQ and from the assessment sites for their work. Finally, we are greatly in debt to collaborators and members of the scientific advisory board and scientific working groups for their invaluable help during the conception of the data collection tools and methods.

Conflicts of Interest: None declared.

KEY MESSAGES

- CARTaGENE is a unique platform for investigating genetic and environmental determinants of multiple chronic disorders.
- It is the largest prospective study in the Quebec population that has measured such a broad range of phenotypes associated with chronic diseases
- Participant follow-up is secured for 50 years through health and administrative databases. The vast majority of participants have also agreed to be directly contacted in the future.
- The burden of multimorbidity was high with over a third of the cohort reporting three or more chronic conditions.
- Patterns of non-random associations between chronic conditions were observed suggesting common underlying pathophysiological phenomena that may be explored.

References

[1] Godard B, Marshall J, Laberge C. Community engagement in genetic research: results of the first public consultation for the Quebec CARTaGENE project. *Community Genet* 2007;**10:**147–58.

[2] Fortier I, Burton PR, Robson PJ *et al*. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39:** 1383–93.

[3] Willett WC, Colditz GA. Approaches for conducting large cohort studies. *Epidemiol Rev* 1998;**20:**91–99.

[4] Doll R. Cohort studies: history of the method. I. Prospective cohort studies. *Soz Praventivmed* 2001;**46:**75–86.

[5] Robinson KA, Dennison CR, Wayman DM, Pronovost PJ, Needham DM. Systematic review identifies number of strategies important for retaining study participants. *J Clin Epidemiol* 2007;**60:**757–65.

[6] Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J *et al*. Size matters: just how big is

BIG?: quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2008; 2009; **38:**263–73.

[7] Spitzer RL, Kroenke K, Williams JB, Lowe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;**166:**1092–97.

[8] Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999;**282:** 1737–44.

[9] Karasek R, Brisson C, Kawakami N, Houtman I, Bongers P, Amick B. The job content questionnaire (JCQ): an instrument for internationally comparative assessments of psychosocial job characteristics. *J Occup Health Psychol* 1998;**3:**322–55.

[10] Craig CL, Marshall AL, Sjostrom M *et al*. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* 2003;**35:**1381–95.

[11] Fortin M, Bravo G, Hudon C *et al*. Relationship between multimorbidity and health-related quality of life of patients in primary care. *Qual Life Res* 2006;**15:**83–91.

[12] Fortin M, Dubois MF, Hudon C, Soubhi H, Almirall J. Multimorbidity and quality of life: a closer look. *Health Qual Life Outcomes* 2007;**5:**52.

[13] Glynn LG, Valderas JM, Healy P *et al*. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Fam Pract* 2011;**28:**516–23.

[14] Prados-Torres A, Poblador-Plou B, Calderon-Larranaga A *et al*. Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis. *PLoS ONE* 2012;**7:**e32190.

[15] Statistics Canada. *Leading causes of death in Canada 2008*, Available from: http://www.statcan.gc.ca/daily-quotidien/111101/dq111101b-eng.htm (7 September 2012, date last accessed).

[16] Fortier I, Doiron D, Little J *et al*. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol* 2011;**40:** 1314–28.

[17] Smith-Warner SA, Spiegelman D, Ritz J *et al*. Methods for pooling results of epidemiologic studies: the pooling project of prospective studies of diet and cancer. *Am J Epidemiol* 2006;**163:**1053–64.

[18] Luan JA, Wong MY, Day NE, Wareham NJ. Sample size determination for studies of gene-environment interaction. *Int J Epidemiol* 2001;**30:**1035–40.

[19] Hattersley AT, McCarthy MI. What makes a good genetic association study? *Lancet* 2005;**366:**1315–23.

[20] Watts G. UK Biobank gets 10% response rate as it starts recruiting volunteers. *BMJ* 2007;**334:**659.

[21] Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vezina H. Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* 2011;**144:**432–41.

[22] Laberge AM, Michaud J, Richter A *et al*. Population history and its impact on medical genetics in Quebec. *Clin Genet* 2005;**68:**287–301.

[23] Korn E, Graubard B. *Analysis of Health Surveys*. New York: John Wiley and Sons, 1999.

[24] Fortin S, Pathmasiri S, Grintuch R, Deschenes M. 'Access arrangements' for biobanks: a fine line between facilitating and hindering collaboration. *Public Health Genomics* 2011;**14:**104–14.