Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2012; all rights reserved. Advance Access publication 10 July 2012

International Journal of Epidemiology 2013;**42**:689–700
doi:10.1093/ije/dys084

COHORT PROFILE

Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness

Blair H Smith,[1,2]* Archie Campbell,[3] Pamela Linksted,[3] Bridie Fitzpatrick,[4] Cathy Jackson,[5] Shona M Kerr,[3] Ian J Deary,[6] Donald J MacIntyre,[7] Harry Campbell,[8] Mark McGilchrist,[1] Lynne J Hocking,[9] Lucy Wisely,[2] Ian Ford,[4] Robert S Lindsay,[4] Robin Morton,[3] Colin N A Palmer,[1] Anna F Dominiczak,[4] David J Porteous[3] and Andrew D Morris[1]

[1]Medical Research Institute, University of Dundee, Dundee, UK, [2] Division of Applied Health Sciences, University of Aberdeen, Aberdeen, UK, [3]Medical Genetics Section, Molecular Medicine Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK, [4]College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK, [5]School of Medicine, University of St. Andrews, St. Andrews, UK, [6]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK, [7]Division of Psychiatry, University of Edinburgh, Edinburgh, UK, [8]Division of Community Health Sciences, University of Edinburgh, Edinburgh, UK and [9]Division of Applied Medicine, University of Aberdeen, Aberdeen, UK

*Corresponding author. University of Dundee, Ninewells Hospital and Medical School, Mackenzie Building, Kirsty Semple Way, Dundee, DD2 4BF, UK. E-mail: b.h.smith@dundee.ac.uk

Accepted 8 May 2012

Summary GS:SFHS is a family-based genetic epidemiology study with DNA and socio-demographic and clinical data from about 24 000 volunteers across Scotland aged 18–98 years, from February 2006 to March 2011. Biological samples and anonymized data form a resource for research on the genetics of health, disease and quantitative traits of current and projected public health importance. Specific and important features of GS:SFHS include the family-based recruitment, with the intent of obtaining family groups; the breadth and depth of phenotype information, including detailed data on cognitive function, personality traits and mental health; consent and mechanisms for linkage of all data to comprehensive routine health-care records; and 'broad' consent from participants to use their data and samples for a wide range of medical research, including commercial research, and for re-contact for the potential collection of other data or samples, or for participation in related studies and the design and review of the protocol in parallel with in-depth sociological research on (potential) participants and users of the research outcomes. These features were designed to maximize the power of the resource to identify, replicate or control for genetic factors associated with a wide spectrum of illnesses and risk factors, both now and in the future.

Why was the cohort set up?

Generation Scotland (GS) is a multi-institutional, cross-disciplinary collaboration bringing together the Scottish medical schools and the National Health Service (NHS). Funded since 2003, GS includes studies with more than 30 000 volunteers who have provided samples and data for genetic epidemiology research. In 2004, GS bid successfully for the Scottish

government's genetics in health-care initiative and was funded to undertake the Scottish Family Health Study (SFHS).[1]

Family-based population genetic studies are important because the increased kinship among subjects enables the measurement of heritability and study of parent-of-origin effects. They allow use of approaches such as long-range haplotype phasing, imputation of rare alleles and haplotypes and linkage mappings. These will be increasingly important as genetic epidemiological studies turn their attention to identifying quantitative trait loci carrying multiple rare variants, causal trait-associated variants and *de novo* gene mutations, facilitated by recent advances in the speed and cost-effectiveness of DNA sequencing.[2–4] This work has the potential not only to lead to the discovery of genetic variants associated with health, disease and response to treatment but also to contribute to discovery of important disease pathways leading to future biomarker development and drug discovery.

Participants came from across Scotland, with some family members from further afield. Scotland has a high prevalence of common conditions including coronary heart disease, stroke, cancer, chronic obstructive pulmonary disease (COPD), diabetes and mental illness. Furthermore, the population in Scotland is relatively static and stable,[5] providing an ideal population in which to measure heritable and lifestyle influences on disease. Life expectancy is rising but still lags behind the rest of the UK[6] and many other Western European countries.[5] The leading causes of chronic illness and premature death are likely to become more prevalent in coming decades with the ageing population. The Scottish government, therefore, identified four main health priorities: cancer, coronary heart disease, stroke and dementia.[7,8] These and many other common diseases have complex aetiology, with genetic and environmental factors interacting to influence phenotype and disease expression.

Who is in the cohort?

The original protocol for GS:SFHS has been described previously.[1] A summary is presented here, with more details available on the study website (www.generationscotland.org). Potential participants ('probands') were identified at random from those aged 35–65 years from the lists of collaborating general medical practices, and invited to participate and also to identify at least one first-degree relative aged at least 18 years who would also participate, and preferably more. In the UK, ~96% of the population is registered with a general practitioner (GP),[9] so this approximates to a general population family sample. Initially (2006–10), GPs in the Glasgow and Tayside areas of Scotland collaborated; the study was extended during 2010 to include Ayrshire, Arran and Northeast Scotland. The age range for probands was also broadened to 18–65 years at this time. Besides recruitment by invitation, volunteers were welcomed if they met the criteria ($\geqslant$18 years, with at least one first-degree relative who could participate). A total of 21 476 participants attended a research clinic in Glasgow, Dundee, Perth, Aberdeen or Kilmarnock, having first completed a pre-clinic questionnaire (PCQ). A further 2484 postal participants who were unable to attend a clinic completed a PCQ and sent a saliva sample for DNA extraction, using an Oragene kit.[10] These methods of identification, approach and recruitment were developed in parallel with an extensive public consultation exercise, as were the consent form and the participant information leaflet (PIL).[11] They were all reviewed after further consultation with actual and potential participants.[12,13]

In total, 126 000 potential probands were invited to participate, and 12.3% of them volunteered and met the main study criteria. Not all were recruited, for practical reasons or because of inability to recruit family members. In total, 6665 of these probands completed appointments (overall response rate 5.3%), along with an additional 1288 individuals who volunteered without invitation, and 16 007 family members, giving a total of 23 960.

The sample was 59% female, with a wide range of ages and socio-demographic characteristics (Tables 1 and 2; Figure 1); 39% lived in areas with above average socio-economic deprivation in the Scottish index of multiple deprivation.[14] Most (87%) participants were born in Scotland and 96% in the UK or Ireland. Around 82% of the parents and 75% of the grandparents were also born in Scotland. Data are available on birthplaces by local council area for participants, parents and grandparents born in Scotland.

A shared family identity number was given to groups where each member was a first-degree relative of at least one other person. Mean family size (excluding 1400 singletons without any relations in the study) was 4.05 members; median was 3 (IQR 2–5). The largest family had 36 participating members, and participants were grouped in 5573 families (Figure 2).

Compared with the Scottish population, where recent data were available in comparable format,[15–18] our sample was generally healthier and wealthier, with a different age–sex profile, though important similarities were apparent (Table 3). Although not truly representative, the sample includes a wide range of socio-demographic and clinical features. The cohort size results in large amounts of data on participants from all socio-economic classes, with many or multiple disease traits (including extreme values) and identification of intensively matched control subjects.

Table 1 Some comparisons between GS:SFHS cohort and the Scottish population[15–18]

	GS:SFHS	Scottish population
Median age (years)		
Male	47	37
Female	48	39
Gender (% Male)	41	48
Ethnicity (% White)	99	98
Employment (those aged up to 75 years) (%)		
Unemployed	1.7	4.0[a]
Retired	15.1	12.9[a]
Employed (full or part-time, or self-employed)	62.8	58.0[a]
Education (%)		
Degree	33	20
No qualifications	5	33
Overweight or obese (BMI > 25) (%)		
Males	65	68
Females	53	61
Current smokers (%)		
Males	19	25
Females	16	25
Alcohol intake (mean/week) (units)		
Males	15.8	17.5
Females	7.1	7.8
Depression (%)	4 (2 or more (GHQ-d)	8 (2 or more symptoms)
Anxiety (%)	16.3 (2 or more (GHQ-b)	9 (2 or more symptoms)
Chronic pain (%)		
Any chronic pain	32	46
Severe chronic pain	2.7	5.7
Hypertension (%)		
Males	37.4 (Measured)[b]	35[c]
	14.0 (Self-reported)	
Females	24.7 (Measured)[b]	31[c]
	13.2 (Self-reported)	
Heart disease (%)	Total 3.9 (self-reported)	
Males	5.6	7.3
Females	2.7	5.2
Diabetes (%)	Total 3.3 (self-reported)	
Males	4.1	5.7
Females	2.8	4.3
Stroke (%)	Total 1.5 (self-reported)	
Males	1.7	2.7
Females	1.3	2.2

[a]People aged 16–74 years.
[b]Systolic > 140 or diastolic >90.
[c]Treated or untreated hypertension.

Table 2 Summary of phenotype and samples available, and percentage providing valid/useable data

Consent (%)	Consent to link data with medical and related records	100
	Consent for linkage with medical and related records, with CHI number available[a]	92.1
	Consent to be re-contacted for related research	98.2
Demographics (%)	Age	100
	Gender	100
	Education level	90.2
	Ethnicity	92.7
	Housing tenure	93.9
	Occupation	95.3
	Family structure	100
	Family history	95.6
	Postcode and Scottish index of multiple deprivation[a]	87.2
Health and lifestyle (%)	Smoking history	96.9
	Tobacco exposure	95.0
	Alcohol history	95.9
	Physical activity	96.5
	Diet history	95.7
	Personal medical history[b]	37.0
	Family medical history[b]	94.6
	Parents lifespan and cause of death	55.9
	Medication history[b]	74.2
	Surgical operations[c]	29.1
	Rose angina questionnaire	94.7
	Chronic pain—case identification, chronic pain grade	95.3
	Musculoskeletal history (fractures and arthritis)	95.6
	Women's health (% of females, $n = 14\ 046$)	95.3
	Mood disorder questionnaire (MDQ)[d]	55.8
	Schizotypal personality questionnaire (SPQ-B)[d]	56.1
Samples collected and results available (%)	Blood—DNA extracted and stored	84.5
	Saliva—DNA extracted and stored; Whatman cards stored	13.6
	Total with DNA available for analysis (after extraction)	98.1

Data for whole study (clinical and postal); $n = 23\,960$.
[a] Available for Scottish residents only, the CHI number is a unique identifying number allocated to every person registered with a GP, and used for all NHS procedures.
[b] Only completed where disease self-reported.
[c] Phase 1 PCQ only.
[d] Phase 2 PCQ only.

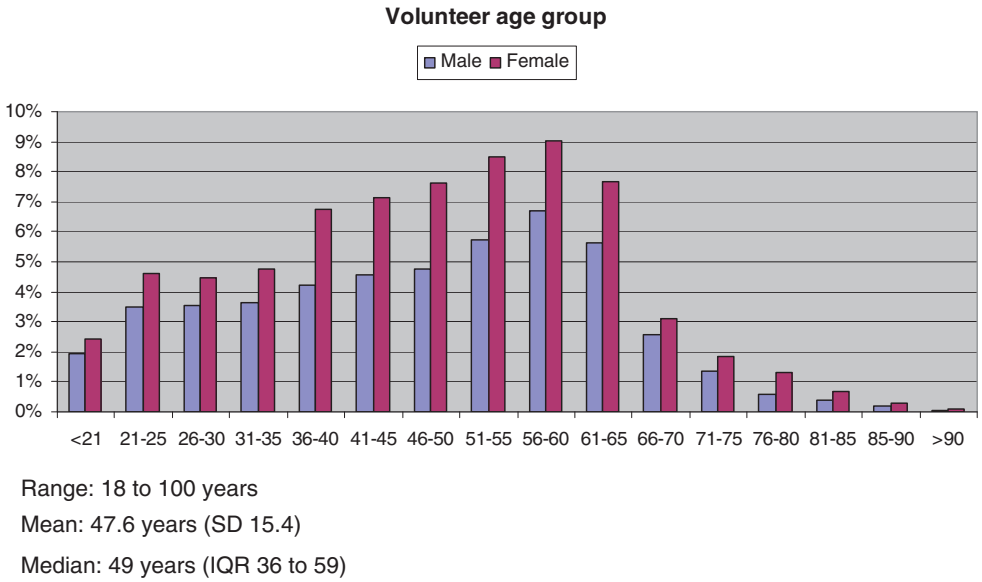

Figure 1 Age and gender distribution of GS:SFHS. Range: 18–100 years. Mean (SD) 47.6 (15.4) years. Median (IQR) 49 (36–59) years

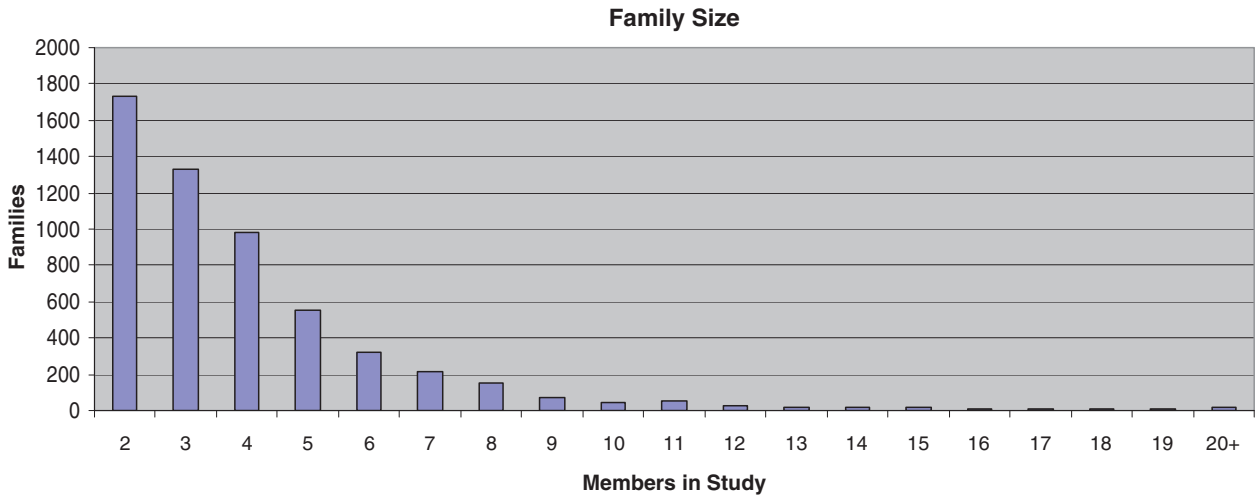

Figure 2 Family size of participating families. Families are groups in the study sharing a family identity number; a family identity number was given to groups where each member was a first-degree relative of at least one other person in the group. Singletons (1400) are excluded from this graph—most of these had relatives who were registered with the study but did not participate

What has been measured?

In addition to the PCQ, multiple measures were taken at the clinic, and samples of blood and urine collected (or saliva, if insufficient blood could be obtained for DNA extraction). In total, DNA was successfully extracted from 98.1% of the participants. A summary of all data collected and their completeness is shown in Tables 1 and 2.

All measures were taken in a standard order by trained clinical staff, according to rigorous standard operating procedures (SOPs) based on best supported practice (available on request or at www.generationscotland.org). Before any data collection, participants signed 'broad' consent forms,[19] with appropriate discussion, permitting use of data and samples for 'future medical research into health, illness and medical treatment' without specifying this further. It was made clear in the consent form, PIL and discussion that such research would be subject to appropriate ethical review, but might also lead to patent development and commercialization. They were also invited to provide consent for linkage of their data and samples to routine data sets (including

Table 3 Clinical measurements and data

Physical measurements (%)	Height	99.7
	Weight	99.3
	Waist circumference	98.4
	Hip circumference	98.4
	Body fat (bio-impedance)	97.2
	Blood pressure and heart rate	99.9
	Ankle brachial pressure index	93.2
	Spirometry—FEV_1, FVC, FEF	84.1
	ECG	98.6
Cognitive and mental health	GQH	98.7
	Eysenck personality questionnaire	99.4
	Digit symbol test	98.8
	Verbal fluency	98.7
	Mill Hill vocabulary scale	98.2
	Wechsler memory test (immediate and delayed)	99.3
	Screening for emotional and psychiatric problems using structured clinical interview for DSM-IV disorders (SCID)	99.6
	Mood sections of SCID completed after positive screen	18.8
	Impulsivity questionnaire BIS-11[a]	4.9
	Choice reaction time[a]	4.1
Samples collected and results available (%)	Blood in cryopreservation	85.7
	Serum—stored	92.7
	Cholesterol, urea and electrolytes	94.5
	Glucose	94.2
	Urine—stored	95.6

$n = 21\,476$.
[a]Tests introduced in final year of study.

NHS data), and to be re-contacted in the future for possible participation in research.

Minor revisions to the PCQ were introduced in 2009, and some additional instruments and questions included (as listed in Tables 1, 4 and 5); the period up to this was termed Phase 1 ($n = 9901$), and the period thereafter Phase 2 ($n = 14\,059$). At the time of collection, the PCQ, clinic report form (CRF) and biological samples were given barcode labels with unique study identification codes. No personal identifying information was attached to any data or samples at any stage. Ethical approval for the study was given by the NHS Tayside committee on research ethics (reference 05/s1401/89). Governance of the study, including public engagement, protocol development and access arrangements, was overseen by an independent advisory board, established by the Scottish government.

Validated questionnaires included those for chronic pain severity,[20] and angina.[21] Among other clinical assessments, blood pressure (BP) and heart rate were recorded twice, an electrocardiogram (ECG) was taken and respiratory function was measured using forced expiratory volume in 1 s (FEV_1), forced vital capacity (FVC) and forced expiratory flow (FEF) 3× each. All clinical participants were screened for a history of emotional and psychiatric disorders using the structured clinical interview for DSM-IV disorders (SCID); 21.7% screened positive and were invited to continue the interview that focused on mood disorders; of these, 88% completed the interview (19.0% of participants). Milder degrees of psychological distress were assessed using the general health questionnaire 28 (GHQ 28).[22] Individual domain scores gave information on (i) somatic symptoms, (ii) anxiety/insomnia, (iii) social dysfunction and (iv) severe depression. The personality traits of neuroticism and extraversion were assessed using the Eysenck personality questionnaire—revised.[23] In Phase 2, participants were asked to complete standard questionnaires on schizotypal personality (SPQ-B)[24] and mood disorders (MDQ),[25] and 1278 people also completed an impulsivity questionnaire (BIS-11).[26] The following domains of cognitive function were assessed: verbal declarative memory (Wechsler logical memory test immediate and delayed);[27] executive function (letter-based verbal fluency test);[28] speed of information processing (Wechsler digit symbol test)[29] and verbal ability (Mill Hill vocabulary test).[30] Finally, 1049 people were tested on four-choice reaction time (Deary–Liewald CRT test[31]), introduced later in the study.

How often have they been followed up?

Although data collection was cross-sectional, GS:SFHS becomes a cohort as a result of the ability to link to routine NHS data (Table 6). Linkage is possible using the community health index (CHI) number, a unique identifying number allocated to every person in Scotland registered with a GP, and used for all NHS procedures (registrations, attendances, samples, prescribing and investigations). Full prescribing data are available for participants in the Tayside area dating back to January 1989, in Fife from March 2008, and from the Glasgow area from January 2010.

A study looking at cardiovascular risk and assessment of familial clustering of risk involved obtaining hospital admissions and prescribing data for 15 787 participants, by linkage to their NHS records. In June 2011, these participants were found to have had a total of 27 061 hospital admissions, involving 6347 unique patients, since January 1981. Among those living in Tayside and Fife ($n = 7315$), a total

Table 4 Reported personal and family history of common illnesses

	Self-reported	Self and parent[a]	Self and sibling[a]	Self, parent and sibling[a]	Families with ⩾2 participants reporting
High blood pressure	3243	1694	1003	658	666
Heart disease	930	620	281	221	85
Diabetes	798	259	184	89	77
Stroke	351	111	47	27	16
Depression[b]	1496	552	384	196	328
Severe depression[c]	680	214	148	87	86
Asthma	2644	624	636	240	476
Osteoarthritis	1770	688	356	218	247
Rheumatoid arthritis	430	121	65	39	22
Cancer (various)[d]	567	229	60	17	29
COPD[c]	272	15	9	19	0

[a] With the exception of the final column, parents and siblings who were reported to have these illnesses were not necessarily participants in GS:SFHS.
[b] Phase 1 ($n = 9901$).
[c] Phase 2 ($n = 14\,059$).
[d] Breast cancer, bowel cancer, lung cancer and prostate cancer.

Table 5 Results of cognitive function and personality testing in GS:SFHS

	n	Theoretical maximum score	Mean	Range	Standard deviation
Eysenck personality test—neuroticism	21 340	12	3.9	0–12	3.2
Eysenck personality test—extraversion	21 333	12	7.8	0–12	3.5 (median 8, IQR 5–11)
Wechsler memory test—short-term	21 316	25	16.1	0–25	3.9
Wechsler memory test—delayed	21 200	25	14.9	0–25	4.3
Wechsler memory test—combined	20 969	50	31.0	0–50	8.0
Digit symbol test	21 219	133	72.2	0–133	17.2
Verbal fluency test—combined	21 207		39.7	0–97	11.7
Mill Hill vocabulary test	21 075	44	30.1	0–44	4.8
GHQ-a[a], median (IQR)	21 188	7	0.9	0–7	1.6 [0 (0–1)]
GHQ-b[a], median (IQR)	21 190	7	0.7	0–7	1.5 [0 (0–1)]
GHQ-c[a], median (IQR)	21 183	7	0.6	0–7	1.4 [0 (0–1)]
GHQ-d[a], median (IQR)	21 183	7	0.2	0–7	0.8 [0 (0–0)]
GHQ total[a], median (IQR)	21 186	28	2.4	0–24	4.1 [0 (0–3)]
Schizotypal personality questionnaire[b]	12 798	22	3.9	0–22	3.7
MDQ[b]	13 022	13	2.5	0–13	3.2

[a] GHQ domain scores give information on: (i) somatic symptoms; (ii) anxiety/insomnia; (iii) social dysfunction; and (iv) severe depression.
[b] Phase 2 only.

of 1 099 566 NHS prescriptions had been recorded, with 8283 individuals, including 9173 items, between January 1989 and March 2011. Until 30 June 2010, these included, for example, 3183 prescriptions of drugs for diabetes [British National Formulary (BNF)[32] Chapter 6.1] for 243 individuals; 7129 anti-hypertensives (BNF Chapter 2.5) for 1012 individuals and 7404 statins for 1128 individuals. In the same group and period, there were 1293 hospital admission involving 978 individuals, including 282 to general medicine (197 individuals), 188 to general surgery (157 individuals) and 182 to orthopaedics/trauma (161 individuals). The prescribing linkage was repeated in December 2011 for participants living in the Glasgow area, among whom 143 151 prescriptions had been recorded for 4575 individuals,

Table 6 Examples of routine and research data sets to which GS:SFHS data and samples can be linked anonymously, using the CHI number

Data set	Description
Scottish morbidity record 1 (SMR 01)[a]	Acute inpatient and daycase hospital admissions—disease and procedure codes. Scotland-wide, around 750 000 admissions per annum since 1981. http://www.isdscotland.org/health-topics/hospital-care/
The acute linked database (SMR 01/SMR 04/SMR 06/deaths)[a]	SMR 01 linked with cancer registration (SMR 06), mental health inpatients (SMR 04) and death records for the period 1981 onwards and holds data on >5 million patients with >40 million contacts within the acute hospital sector
Scottish morbidity record 2 (SMR 02)[a]	Pregnancies and births—disease and procedure codes. Scotland-wide, around 58 000 deliveries per annum since 1975. http://www.isdscotland.org/health-topics/maternity-and-births/births/
Maternity and neonatal linked data set	Linked maternity, neonatal and stillbirth and infant death records from 1975, with records pertaining to mother and baby held together
The Scottish birth record (SBR)[a]	The Scottish birth record system for all of a baby's neonatal care in Scotland collected since 2002. http://www.isdscotland.org/products-and-services/scottish-birth-record/
Prescribing information systems[a]	Patient-specific—dispensed prescription items. Approximately 92 million items/annum since 2008 (95% complete), Scotland-wide. CHI is available on around 93% of all prescriptions dispensed in the community in Scotland from August 2009 onwards. For prescriptions prescribed by GPs (i.e. using the GP10 form), the figure is around 96%. http://www.isdscotland.org/health-topics/prescribing-and-medicines/
Scottish care information, diabetes collaboration (SCI-DC)	Core clinical data set on all people with diabetes in Scotland (around 238 000)
Scottish cancer registry[a]	Register of cancer diagnoses since 1958: 45 000 registrations/annum; total 1.4 million cases, Scotland-wide. ISD hold cancer registrations back to 1981, indexed with the CHI number where possible. http://www.isdscotland.org/health-topics/cancer/
Hepatitis C diagnosis and HIV diagnosis databases	Data set from laboratory data on individuals newly diagnosed with (i) hepatitis C virus and (ii) HIV infection: 30 000 hepatitis C diagnoses in total during 1991–11, 7000 HIV diagnoses in total during 1981–11, Scotland-wide
Scottish stroke care audit	Demographic information, stroke pathway data, outcome predictors, anti-thrombotics on admission, medication on discharge, brain imaging, stroke classification: >40 000 patients in Scotland since 2002 with comprehensive data from 2005
Walker birth cohort	All births in Tayside, including maternal and child phenotype, 1952–66: $n = 46\,000$; targeted recruitment of ~1200 into GS family study
Aberdeen children of the 1950s	Cohort comprising individuals born in Aberdeen during 1950–56. The extensive early life information including Aberdeen child development survey supplemented by follow-up information collected around 1999. Targeted recruitment of ~570 plus their family members into GS family study

[a]Data sets held by Information Services Division (ISD) of NHS National Services Scotland (www.isdscotland.org).

including 3281 items. Detailed analysis of these data is underway.

A further project linking study data with a range of NHS Scottish morbidity records (SMR 01, 04, 06) and prescribing records for all participants who have consented and for whom a CHI number is available allowing linkage [~22 000 (92%)] is currently under way. This will allow validation of the self-reported events recorded in the study, and provide information on clinical endpoints and follow-up. The mechanism for governance of these and further linkages has been established and a number of other research proposals involving linkage of study data from selected subsets of the cohort to a range of other NHS data sets have been approved.

In addition, consent for re-contact allows follow-up studies or related research. One study is in the process of doing this, having invited 500 participants to volunteer for further research on stroke; to date, around 150 have responded positively and meet the entry criteria. Several other studies involving re-contact of GS:SFHS participants have been approved and will begin soon.

What has it found?

Baseline epidemiology

Baseline clinical results are summarized in Tables 3–5 and Figures 3 and 4. Based on self-reporting, illnesses

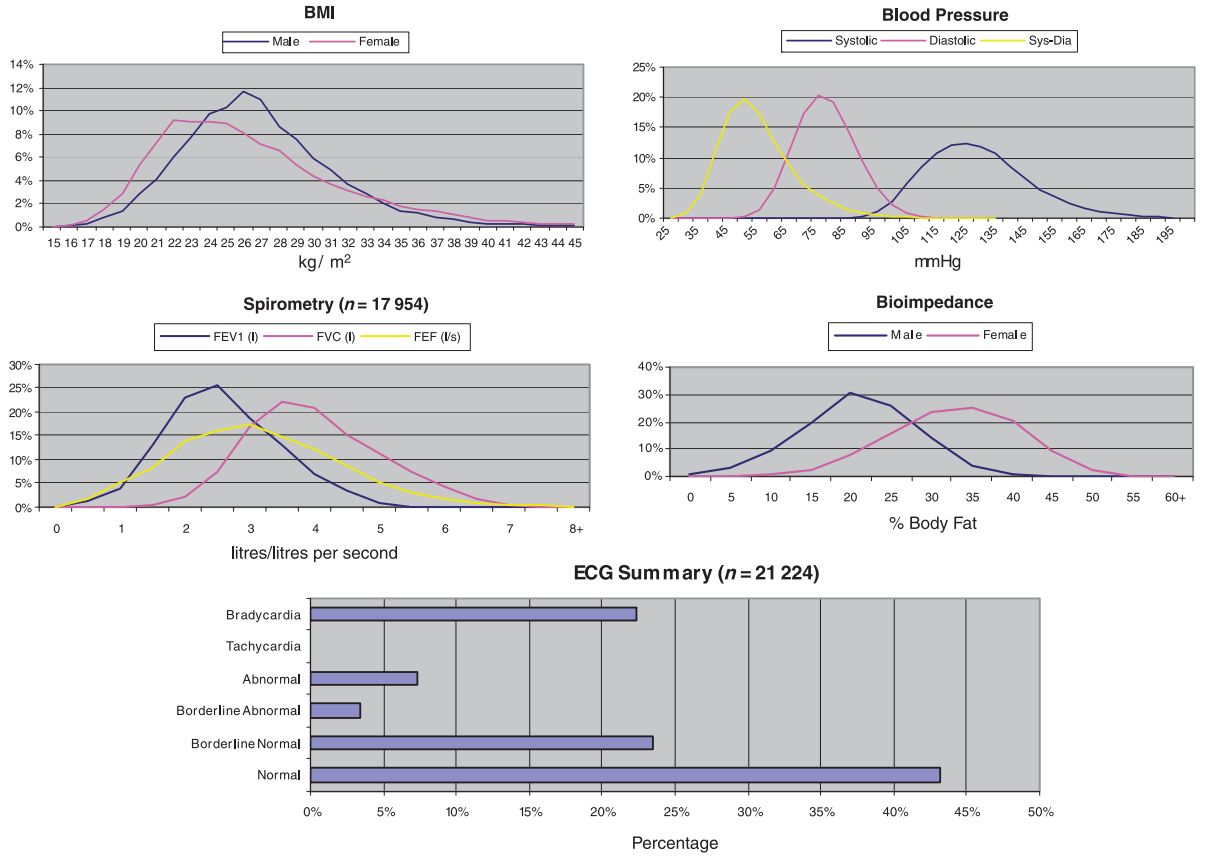

Figure 3 Distribution of results for selected physical measures

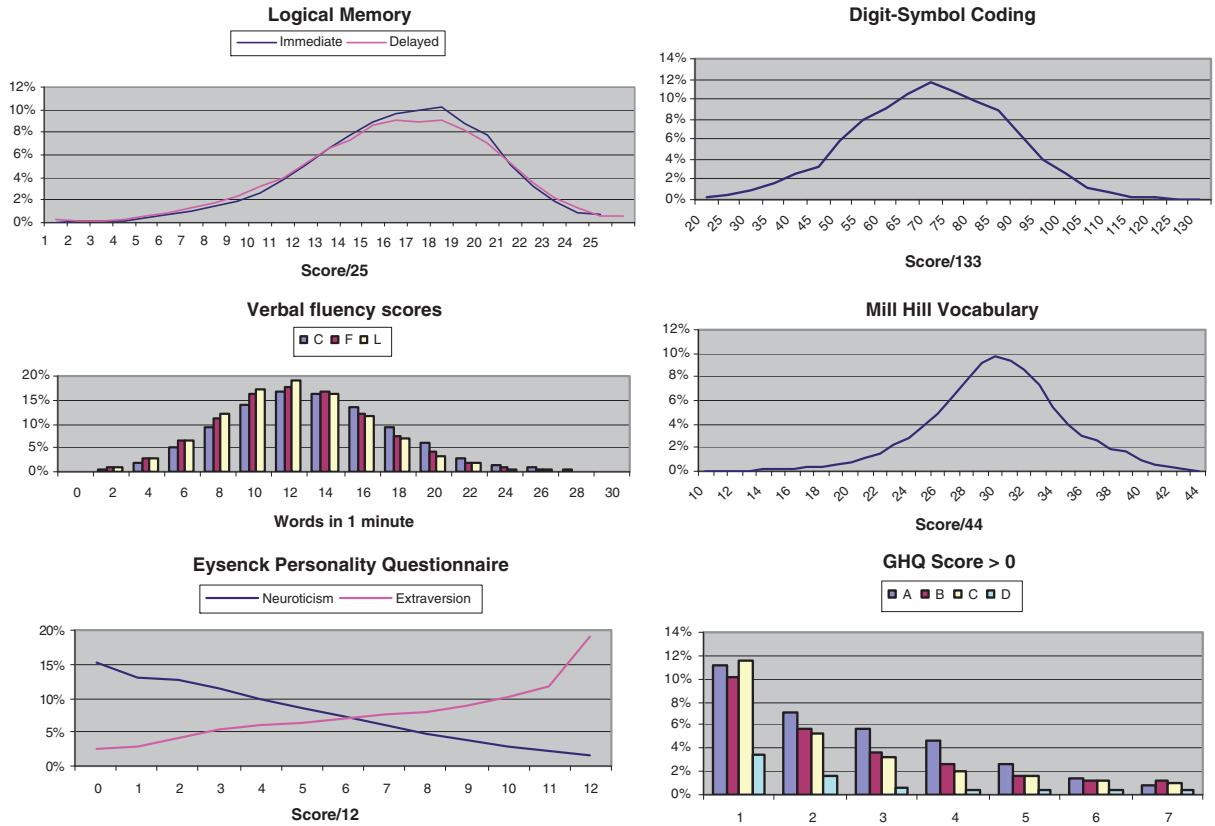

Figure 4 Distribution of responses to personality and cognitive function tests

were generally less prevalent than in the Scottish population (Table 3), but were reported more frequently by participating family members than by respondents themselves (Table 4). Almost half (41%) were taking at least one prescribed medication. From the Rose angina questionnaire,[21] 4% of the individuals were classified with a previous possible myocardial infarction, and 4% with possible angina. Only 28% of the former and 23% of the latter reported heart disease in their personal medical history. The cohort tended to be overweight, 21% having body mass index (BMI) >30. BP measurement found 31% to be 'hypertensive' (with systolic BP >140 mmHg, or diastolic BP >90 mmHg), compared with only 13.7% who self-reported hypertension. Similarly, around 6% of the participants who completed respiratory tests had evidence of COPD (FEV_1 <0.8 of expected value, and FEV_1/FVC ratio <0.7),[33] whereas only 1.1% self-reported COPD. Analysis of ECGs found that 68% were classified as 'normal', 10% 'abnormal' and 22% with bradycardia.

A history of a single major depressive disorder was diagnosed in 6.6%, and recurrent major depressive disorder in 6.3% (total with major depression 12.9%); the self-reported prevalence was only 6.2%. A further 0.3% were diagnosed with a history of bipolar disorder. Although comparable data for the UK are not available, epidemiological studies in the USA found rates of 17.1% in the national comorbidity survey in 1994[34] and 16.6% in a replication study a decade later.[35] The number (%) who had GHQ 28 scores >4 (indicating potentially significant psychological distress) was 4009 (18.9%) ($n = 21\,186$). The sample was very slightly higher in extraversion, and more markedly lower in neuroticism than the data reported on adults by Eysenck *et al*.[15]

Cognitive function testing produced a wide range of results, with most showing a near-normal distribution (Table 5 and Figure 4). Logical memory immediate and delayed correlated at 0.864, and so was combined to give a total score. Pearson's correlations among the cognitive tests were all significant ($P < 0.001$): logical memory total correlated 0.309 with digit symbol, 0.152 with verbal fluency and 0.221 with Mill Hill vocabulary; digit symbol correlated 0.233 with verbal fluency, and 0.066 with Mill Hill vocabulary; verbal fluency correlated 0.397 with vocabulary.

Research outputs

By February 2012, nearly 100 applications from investigators across the UK interested in a wide range of clinical and scientific questions had been processed. These include completed and currently active studies, and those awaiting funding. For example, a family-based analysis has provided the first estimate of the heritability (h^2) of chronic pain (0.29 unadjusted, 0.16 adjusted for other variables) and severe chronic pain (0.44 and 0.30).[36] This provides a basis for seeking the genes involved. Similarly, the heritability of a number of cardiovascular disease (CVD) risk factors and of cognitive function was calculated in the sample, finding that not only were CVD and cognitive function strongly associated but also that these associations were caused predominantly by genetic rather than environmental factors.[37]

A genetic analysis found that two variants of the complement receptor 1 gene (*cr1*), known to be associated with Alzheimer's disease, were also associated with major recurrent depression and with psychological distress, particularly among women.[38] Other analyses of genes associated with depression are completed or under way, including genome-wide association and exome sequencing. GS:SFHS was the largest contributor of cases and controls ($n = 5474$) to the SpiroMeta consortium that identified and characterized five new genetic variants associated with lung function, three of which (*tns*, *gstcd* and *htr4*) were associated with COPD.[39,40] GS:SFHS subsequently made the largest contribution ($n = 10\,399$) to the large-scale follow-up (total $n = 46\,411$) of a genome-wide association study that identified 16 further new loci associated with lung function.[41]

Serum from participants was used in immunosurveillance of H1N1 influenza exposure, finding that 34% of the Scots had been naturally infected by March 2010, at a peak rate of 4.3 new infections per 1000 people/day.[42] Other studies are under way in a range of clinical areas and will report soon (www.generationscotland.org/collaborations.htm).

What are the main strengths and weaknesses?

GS:SFHS data have been rigorously and robustly collected, on a wide range of clinical parameters, focusing on quantitative traits that are well established as risk factors for disease. The cohort includes substantial numbers representing the full adult spectrum of ages, lifestyle and demography and important phenotypes and quantitative traits to allow population-based genetic and epidemiological research on many important diseases and risks. Data can be linked anonymously to routinely collected NHS data sets, including prescribing records, hospital attendances and admissions and cancer and death registration.[1] This linkage effectively converts this cross-sectional study into a longitudinal study, with pluripotential outcomes.

GS:SFHS can contribute as a major partner to genome-wide association study meta-analysis consortia enabling the study of SNPs of low minor allele frequency (1–10%).[39–41] Study power increases with sample size and percentage of variance explained, and needs to be estimated for each individual study; however, GS:SFHS is one of the largest family-based genetic epidemiology studies in existence

internationally and is well placed to make an important contribution.

Besides its size, the family-based approach of GS:SFHS enables the study to take a flexible approach to gene discovery, encompassing association and linkage approaches, and gives the potential to study aspects such as heritability of disease-related traits, and parent-of-origin effects. The combination of linkage and association approaches has proven very effective in studies of various diseases/disease traits, including Crohn's disease and HbF levels,[4] especially where several rarer alleles act at a single locus. The study by Luciano *et al.*[37] illustrates the combined strengths of family structure and breadth of phenotypes available in GS:SFHS: by incorporating cognitive measures into the analysis, they were able to explain a significant fraction of cardiovascular risk not explained by established factors.

Medical and medication history are self-reported and therefore subject to recall error. We have noted above the discrepancy between measured and self-reported prevalence of hypertension, COPD and depression, though elsewhere it has been shown that there is substantial agreement between self-reporting and actual presence of disease (as confirmed in medical records).[43] Confirmation of much of the information is possible through linkage of the research data with routine NHS data, and further validation of self-reported medical history within our sample is currently under way. Meanwhile, the use of validated objective assessments is recommended where available.

Although the family-based nature of the cohort is an important strength of the study, providing among other things an efficient strategy for DNA sequence-based studies, relatedness of cohort members will be a confounding factor in some analyses, and may require statistical adjustment.

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

GS has successfully implemented an access process with high governance standards, to which researchers can submit collaboration proposals. Access to data and samples is governed by the GS access committee, and managed through a web-based procedure (www.gsaccess.org). Anonymized data and/or samples are released if the proposed research is consistent with ethical approval and participant consent, confidentiality, data and sample management, and subsequent publication requirements established in the management, access and publications policy (www.generationscotland.org/access.htm). Researchers and their host institutions sign data and/or material transfer agreements to this effect. The collection is now approved as a tissue bank (reference 10/s1402/20), including ethical approval for all research using the data and samples, within the terms of the original ethical approval and participant consent. GS welcomes research proposals from academic and commercial researchers.

Further information is available at www.generationscotland.org.

Funding

The Chief Scientist Office of the Scottish Government and the Scottish Funding Council provide core support for Generation Scotland. GS:SFHS was funded by a grant from the Scottish Government Health Department, Chief Scientist Office, Number CZD/16/6.

Acknowledgements

We are grateful to the families who took part in GS:SFHS, the GPs and Scottish School of Primary Care for their help in recruiting them, and the whole GS team, which includes academic researchers, clinic staff, laboratory technicians, clerical workers, statisticians and research managers.

Conflict of interest: None declared.

KEY MESSAGES

- Analysis of GS:SFHS has found that the association between cardiovascular risk and cognitive function is predominantly genetically mediated.
- GS:SFHS has provided the first estimates of the heritability of (severe) chronic pain, as a single clinical entity.
- GS:SFHS provided the largest contributions to consortia that identified and characterized up to 16 new genetic loci associated with lung function and COPD.
- Around 100 collaborations using the resource are completed or under way, and more are invited, including long-term follow-up via consented data linkage and re-contact.

References

[1] Smith BH, Campbell H, Blackwood D *et al*. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* 2006;**7:**74.

[2] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467:**1061–73.

[3] O'Roak BJ, Deriziotis P, Lee C *et al*. Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat Genet* 2011;**43:**585–89.

[4] Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. *Nat Rev Genet* 2011;**12:** 465–74.

[5] Scottish Public Health Observatory. *Population Dynamics: Overview*. http://www.scotpho.org.uk/home/populationdynamics/populationdynamics.asp; (3 April 2012, date last accessed).

[6] Office for National Statistics. *Topic Guide to Life Expectancies*. http://www.statistics.gov.uk/hub/population/deaths/life-expectancies (3 April 2012, date last accessed).

[7] Scottish Government. *NHS Scotland Chief Executive's Annual Report 2009/10*. http://www.scotland.gov.uk/publications/2010/11/09154028/0 (3 April 2012, date last accessed).

[8] Scottish Government. *Better Care, Better Health: Action Plan*. http://www.scotland.gov.uk/publications/2007/12/11103453/9 (27 March 2012, date last accessed).

[9] *Attribution Dataset GP Registered Populations* 2010. http://www.ic.nhs.uk/statistics-and-data-collections/population-and-geography/gp-registered-populations/attribution-dataset-gp-registered-populations-2010 (3 April 2012, date last accessed).

[10] Oragene kit. http://www.dnagenotek.com/dna_genotek_product_oragene_dna_a_overview.html (3 April 2012, date last accessed).

[11] Haddow G, Cunningham-Burley S, Bruce A, Parry S. Generation Scotland Preliminary Consultation Exercise 2003–04: Public and Stakeholder Views from Focus Groups and Interviews. 2004; Innogen Working Paper 20. Edinburgh: University of Edinburgh. Available at http://www.genomicsnetwork.ac.uk/innogen/publications/workingpapers/title,21143,en.html (Accessed 26 May 2012).

[12] Haddow G, Cunningham-Burley S, Bruce A, Parry S. Generation Scotland: consulting publics and specialists at an early stage in a genetic database's development. *Crit Public Health* 2008;**18:**139–49.

[13] Haddow G, Cunningham-Burley S, Murray L, Myant K, Carlsson A. Can the governance of a population genetic data bank effect recruitment? Evidence from the public consultation of Generation Scotland. *Public Underst Sci* 2011;**20:**117–29.

[14] *Scottish Index of Multiple Deprivation*. http://www.scotland.gov.uk/topics/statistics/simd/ (3 April 2012, date last accessed).

[15] Corbett J, Dobbie F, Doig M *et al*. *Scottish National Health Survey 2009: Volume 1—Main Report*. http://www.scotland.gov.uk/publications/2010/09/23154223/0 (3 April 2012, date last accessed).

[16] General Register Office for Scotland. *Census: Standard Area Statistics (Scotland)*. ESRC/JISC Census Programme, Census Dissemination Unit, Mimas (University of Manchester)/ Centre for Interaction Data Estimation and Research (University of Leeds).

[17] Elliott AM, Smith BH, Penny KI, Smith WC, Chambers WA. The epidemiology of chronic pain in the community. *Lancet* 1999;**354:**1248–52.

[18] Smith BH, Elliott AM, Chambers WA, Smith WC, Hannaford PC, Penny K. The impact of chronic pain in the community. *Fam Pract* 2001;**18:**292–99.

[19] Hansson MG, Joakim D, Bartram CR, Carlson JA, Helgeson G. Should donors be allowed to give broad consent to future biobank research? *Lancet Oncol* 2006;**7:** 266–69.

[20] Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. *Pain* 1992;**50:**133–49.

[21] Rose GA. The diagnosis of ischaemic heart pain and intermittent claudication in field surveys. *Bull World Health Organ* 1962;**27:**645–58.

[22] Goldberg D. *General Health Questionnaire*. Windsor: NFER Publishing Company, 1978.

[23] Eysenck SBG, Eysenck HJ, Barrett P. A revised version of the psychotocism scale. *Pers Individ Dif* 1985;**6:**21–29.

[24] Raine AADB. The SPQ-B: a brief screening instrument for schizotypal personality disorder. *J Pers Disord* 1995;**9:** 346–55.

[25] Hirschfeld RM, Williams JB, Spitzer RL *et al*. Development and validation of a screening instrument for bipolar spectrum disorder: the mood disorder questionnaire. *Am J Psychiatry* 2000;**157:**1873–75.

[26] Patton JH, Stanford MS, Barratt ES. Factor structure of the BARRATT impulsiveness scale. *J Clin Psychol* 1995;**51:** 768–74.

[27] Wechsler D. *Wechsler Memory Scale III UK*. London: Psychological Corporation, 1998.

[28] Lezak MD. *Neuropsychological Testing,* 3rd edn. Oxford: Oxford University Press, 1995.

[29] Wechsler D. *Wechsler Adult Intelligence Scale III UK*. London: Psychological Corporation, 1998.

[30] Raven JC, Court JH, Raven J. In: Lewis HK (ed.). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London: H.K.Lewis, 1977.

[31] Deary IJ, Liewald D, Nissan J. A free, easy-to-use, computer-based simple and four-choice reaction time programme: the Deary–Liewald reaction time task. *Behav Res Methods* 2011;**43:**258–68.

[32] BMA, Royal Pharmaceutical Society. *British National Formulary, 60 (September 2010)*. London: BMJ Group and Pharmaceutical Press, 2010.

[33] Rabe KF, Hurd S, Anzueto A *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: gold executive summary. *Am J Respir Crit Care Med* 2007;**176:**532–55.

[34] Kessler RC, Mcgonagle KA, Zhao S *et al*. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the national comorbidity survey. *Arch Gen Psychiatry* 1994;**51:**8–19.

[35] Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Arch Gen Psychiatry* 2005;**62:**617–27.

[36] Hocking LJ, Generation Scotland, Morris AD, Dominiczak A, Porteous D, Smith BH. Heritability of chronic pain in 2,195 extended families. *Eur J Pain* 2012;**16:**1053–63.

[37] Luciano M, Batty GD, Mcgilchrist M *et al*. Shared genetic aetiology between cognitive ability and cardiovascular disease risk factors: Generation Scotland's Scottish Family Health Study. *Intelligence* 2010;**38:**304–13.

[38] Hamilton G, Evans KL, Macintyre D *et al*. Alzheimer's disease risk factor complement receptor 1 is associated with depression. *Neurosci Lett* 2012;**510:**6–9.
[39] Repapi E, Sayers I, Wain LV *et al*. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2010;**42:**36–44.
[40] Artigas MS, Wayne LV, Repapi E *et al*. Effect of five genetic variants associated with lung function on the risk of chronic obstructive pulmonary disease, and their joint effects on lung function. *Am J Respir Crit Care Med* 2011;**184:**786–95.
[41] Artigas MS, Loth DW, Wain LV *et al*. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* 2011;**43:** 1082–90.
[42] Mcleish NJ, Simmonds P, Robertson C *et al*. Sero-prevalence and incidence of A/H1N1 2009 influenza infection in Scotland in winter 2009–2010. *PLoS One* 2011; **6:**e20358.
[43] Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, Rodeheffer RJ. Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. *J Clin Epidemiol* 2004;**57:** 1096–103.