Genetic Epidemiology 6

Population-based family studies in genetic epidemiology

John L Hopper, D Timothy Bishop, Douglas F Easton

Designs that involve families (the traditional strength of genetic epidemiology) and population-based sampling (the traditional strength of environmental epidemiology) allow investigation of both genes and environment, separately or together, and allow valid inference to the population. These case-control-family designs (including those involving twin pairs), can be regarded as retrospective cohort studies of relatives, and can be used for: determining familial risks and genetic models; estimating risk (penetrance) for measured genotypes; genetic association studies; stratifying risks by family history and known mutation status; and studying modifiers of risk in genetically susceptible individuals. Follow-up of families allows genetic and environmental risks to be studied prospectively. We discuss statistical methods, theoretical and practical strengths, limitations, and other issues. Given their versatility, population-based family studies could become a principal framework in epidemiology, and move genetics from its traditional focus on high-risk families to give it a wider clinical and population health relevance.

Lancet 2005; 366: 1397–406

This is the sixth in a Series of seven papers on genetic epidemiology.

University of Melbourne, Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, 723 Swanston Street, Carlton, Victoria 3053, Australia (Prof J L Hopper PhD); Cancer Research UK Genetic Epidemiology Division, St James's Hospital Leeds, Leeds, UK (Prof D T Bishop PhD); and Cancer Research UK Genetic Epidemiology Unit, Strangeways Laboratories, Cambridge, UK (Prof D F Easton PhD)

Correspondence to:
Prof John L Hopper
j.hopper@unimelb.edu.au

Family studies have a central role in genetic epidemiology. Although epidemiology generally involves studies of unrelated individuals, often using population-based sampling, genetic epidemiology focuses on related individuals in the form of family histories or opportunistically identified and sampled pedigrees. In this article we discuss designs that involve both population-based sampling and groups of relatives who are both interviewed and provide biological material. For simplicity, we focus on diseases and do not consider continuously distributed characteristics.

Genetic epidemiology seeks to dissect the relative contributions of genes and environment and to identify genes determining susceptibility. For the former, studies have been based on specific relationships (especially twin pairs) or on statistical analysis of the patterns of disease aggregation between and within the families of systematically identified cases (probands), and analysed to identify the most plausible explanation of the family aggregation (segregation analysis). For the latter, studies of opportunistically sampled multiple-case families have been critical because such families are those most likely to carry a strong genetic predisposition. In practice, studies of such families have been successful at identifying genes that have a large effect on individual risk. This is traditional gene discovery, and it has worked even for complex diseases such as breast cancer,[1] colorectal cancer,[2] melanoma,[3] Alzheimer's disease,[4] and one form of diabetes.[5]

Once disease-associated genes have been identified, or for candidate genes, researchers try to measure the increased risk for individuals with the putative genetic susceptibility (penetrance) and study interactions with other genetic and environmental risk factors (modifiers of genetic risk). For these tasks, especially the latter, any lack of systematic ascertainment of the multiple-case families affects analysis and interpretation. Analyses must adjust for sampling to avoid bias, and because this involves strong conditioning, limits statistical power. Also, findings from multiple-case families might not apply outside that context, since these families could be enriched for unmeasured risk factors or the mutations with highest risk.

Population-based sampling means sampling from a defined subset of the population using a predetermined scheme with high response rates as required by case-control studies, as distinct from volunteers from the community with possibly low response rates as is typical of cohort studies. It is an attractive alternative, or at least a complementary approach, for gene characterisation and permits valid extrapolation to larger groups within the population. Family studies built around population-based sampling of both cases and controls can in some circumstances be more powerful and robust than a case-control or multiple-case family-based approach alone. By selecting for characteristics that are associated with a putative underlying genetic cause, such as family history of disease or early age at onset, these studies have increased power to address hypotheses related to genetic determinants, and not only those that are rare and associated with strong risks.[6]

Population-based family designs

Case-control families

Case-control family studies include data about disease status and other characteristics of relatives of cases and controls but the information is obtained only from interviews of the cases and controls (figure 1). The data are often restricted to first-degree relatives since information about others is less reliable. The accuracy of the information depends on the disease. For example, reports about first-degree relatives can be quite accurate for breast cancer but negative family histories for ovarian and endometrial cancers are less useful.[7] Furthermore, family history is often only recorded crudely as yes or no. More reliable and extensive data can be obtained in populations with genealogical record linkage, such as Iceland,[8] Sweden,[9] and Utah, USA.[10] The

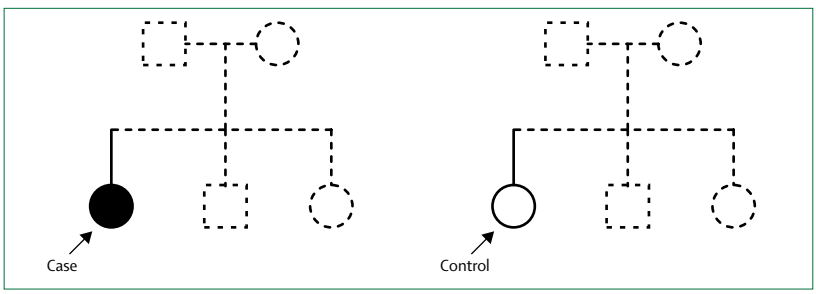

Figure 1: **Example of a case-control family design**
Dotted lines indicate that, although information about disease status is sought for indicated first-degree relatives, they are not interviewed.

lack of information about relatives' risk factors can lead to biased estimates of the dependence between relatives for ages at onset.[11]

Case-families with or without population controls

Studies of case-families recruit relatives of cases in a predefined way (figure 2). This allows comparisons between cases and unaffected family members. The most typical comparison is with siblings, and a special example is disease discordant twin pairs. In case-family studies, as in twin studies, family members are contacted, interviewed, and perhaps sampled for DNA. This contrasts with case-control family studies in that the information on relatives is now based on interviewing the family members as well as the case. Case-family studies allow all comparisons achievable within traditional population-based case-control studies—namely genetic effects alone, effects of environmental exposures alone, and gene-environment interactions.[12] Overmatching within the family for both genetic and shared environmental factors results in reduced power when compared with a similar sized population-based case-control study, but only for some exposure prevalences. For other exposures, such as rare genetic exposures, this argument is immaterial; it is unrealistic to study them using a population-based case-control design because too few, or even no, controls may be carriers. Studies of gene-environment interaction have comparable power for the population-based case-control study as for the case-sibling control.

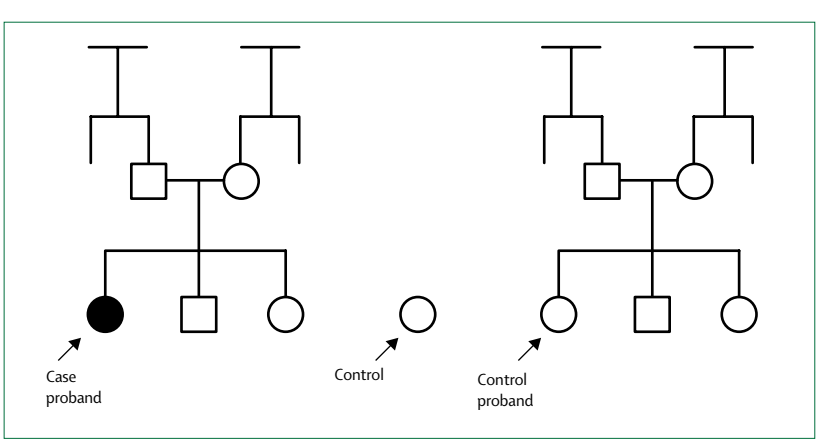

Figure 2: **Example of case-control-family designs**
Solid lines indicate that attempts are made to interview all indicated relatives.

The addition of population-based controls allows further comparisons than the case-family design (figure 2). The additional comparison with the general population means that, for a broad range of exposures and effects, the optimum sample structure has been achieved. Specifically, if there are no shared environmental factors then using a sibling control is more powerful than using a population-based control when trying to estimate genetic effects.[13]

One concern for genetic epidemiological studies is unrecognised genetic stratification within a population. Stratification is much discussed even though there is little prima facie evidence for its existence.[14,15] Nevertheless, the concern has drawn extra attention to family-based studies,[16] because in designs such as the case-family study, cases and controls share parents and hence single gene pools.

Case-control-families

This design, which subsumes the two previous ones by recruiting cases and their relatives as well as controls and their relatives, is consistent with epidemiological principles in that the approach is the same for cases and controls, and therefore for case relatives and control relatives. The hyphens in "case-control-families" indicate this equivalence, just as single hyphens do in the case-control and case-family designs.[17] As well as the opportunities afforded by the family designs above, the case-control-family design allows additional comparisons between sets of relatives of cases and corresponding sets of relatives of controls, provided the two sets are studied with the same protocols and care. If population-based incidences are available (eg, from cancer registries) the quality of the family data can be assessed by comparison with the disease incidence recorded for the control relatives.[18] By interviewing relatives, and thereby obtaining the cancer histories of relatives from multiple sources, the family history of the proband can be extended beyond first-degree relatives and could well become more trustworthy via the cross-validation. Given the increasing difficulty of obtaining high response rates when sampling controls from the population, control groups based on spouses or partners or their relatives (or both) are being used more often. The population-based case-control-family design has now been used in various forms, especially in cancer studies where population complete registries facilitate recruitment (panel 1).[19–24]

What questions can be addressed?

There are three different perspectives on the information collected in case-control-family studies depending on whether the focus is on the cases (and

controls), the disease history of the relatives up until the diagnosis of the case (a retrospective cohort[25]), or the future disease incidence in relatives of cases (a prospective cohort study of sets of relatives[20,21]).

Determining familial risks

The contribution of known disease genes allows an upper estimate of the familial risk due to other factors, and hence informs the prospects for identifying other susceptibility genes. For example, several studies have screened early-onset breast cancer cases for mutations in *BRCA1* and *BRCA2*. Examination of the family history of the mutation carriers detected suggested that only about 15% of the first-degree familial risk of breast cancer was explicable by mutations in these two genes, providing an impetus for the identification of further susceptibility genes.[18,26] Such an estimate could not be derived directly, because the population frequencies of *BRCA1* and *BRCA2* mutations are small and not accurately known.

Determining genetic models

Population-based family data can be used to estimate familial relative risks, the ratio of the risk of the disease for a relative of an affected individual to that for the general population.[27] Familial relative risks (or recurrence risk ratios) can vary by the type of relative (see first paper in this series[28]). For most common cancers, the familial relative risk for first-degree relatives (parents, siblings, and offspring) is 2–4. For many non-malignant diseases it can be much larger; examples are multiple sclerosis, schizophrenia, type I diabetes, and inflammatory bowel disease, where values in the range 5–20 have been recorded.

Familial relative risks quantify the extent of familial aggregation across the population, and reflect the combined effect of all susceptibility genes and environmental risk factors that cluster in families. The pattern of familial risks can provide clues to the existence and mode of inheritance of an underlying disease gene. For example, a higher risk to the co-twin of an affected twin in monozygotic rather than dizygotic pairs suggests that genetic effects may be explaining familial aggregation (provided one can argue or show that this is not due to a greater similarity within monozygotic pairs of the environmental factors associated with the disease); a higher risk to siblings of cases than to their offspring suggests recessive or X-linked genetic components; a greater risk in the maternal aunts or uncles in the absence of increased risks between paternal aunts or uncles indicates X-linked effects; and a rapid decline in familial relative risk with degree of relationship may indicate a polygenic model.[26]

One application of family studies has been to determine genetic models for susceptibility. This approach is known as segregation analysis and asks what model best explains the pattern of familial aggregation of the disease. It involves specification of the mode of inheritance, the population frequency of individuals at high genetic risks, and the associated genetic risks themselves. Segregation analysis was developed in the 1960s and 1970s,[29] but the technique has fallen out of fashion. One weakness is that unless the underlying true model is simple, segregation analysis lacks power to distinguish one model from another. However, as genes are identified, segregation analysis can be used to model the joint effect of known (measured) genes and unmeasured effects.[30,31] This approach can provide a rational basis for genetic counselling,[32–34] genetic testing programmes, and public-health initiatives.

Estimating risk (penetrance) for measured genotypes

An important application of family-based studies is estimation of the risk associated with a particular genetic variant, especially if it is rare. Relatives of a case with such a variant constitute a subpopulation with increased likelihood of carrying the same genetic variant. Including relatives within a systematic study improves the estimation of the effect of being a carrier of such a variant by examining the incidence of disease in the cohort of carrier relatives. In the context of a retrospective cohort of relatives this has been called a kin-cohort design.[35] The carrier status for the variant may be known exactly (eg, if DNA is available from those relatives) or probabilistically by applying the rules of mendelian inheritance to the constellation of known carriers and non-carriers among the case and other relatives. The comparable analysis of the risk associated with a variant for a case-control design is

***Panel 1:* Cancer family registries—infrastructures for studies of the genetic epidemiology of cancers based on population-based and clinic-based families**

Large cancer family registries typically involve both clinic-based multiple-case families and population-based families sampled through cancer registries. They also involve controls, and in some instances control families, and even targeted sampling of twin pairs and their relatives. Population sampling of case families is either irrespective of family history or through a two-stage sampling scheme that over-samples cases with a self-reported family history. Oversampling of cases with earlier onset is often a feature, given that familial and hence most likely genetic factors are more pronounced in those case families. Population-based inference can be achieved by applying appropriate sampling weights. The US National Cancer Institute has supported these novel research infrastructures for breast cancer,[23] colorectal cancer, prostate cancer, and melanoma. Typically these involve a collaboration of research institutions in several countries. The sites have developed core family history and epidemiology questionnaires, data dictionaries, and common protocols for the collection and processing of biospecimens and pathology review, and established centralised informatics support. Many thousands of families have been recruited and some of these resources are also available to external researchers. Detailed information can be found at www.epi.grants.cancer.gov/CFR for the Breast and Colon Cancer Family Registries, www.genomel.org for the Melanoma Genetics Consortium and www.icpcg.org for the International Consortium for Prostate Cancer Genetics.

only informative when sufficient controls carry this variant.

This family-based approach has been widely used to estimate the risks of breast and colorectal cancer for carriers of mutations in *BRCA1* and *BRCA2*[36–38] and *MSH2* and *MLH1*.[39] It can also be used to determine if variants in candidate genes, identified from normal biological function of known genes or by being in regions of interest identified by genome scanning, are associated with increased risk. For rare variants a case-family design can be much more powerful than a case-population control one, and in some instances may be the only feasible design. For commoner, low risk variants, a case-control approach may be more powerful. Irrespective of allele frequencies, case-family analysis of the relatives of variant-carrying cases is informative[40] and provides an essentially independent source of information or confirmation. Risk estimates based on a cohort of relatives may not be strictly comparable with those based on a standard case-control or cohort approach. This is because there may be other causes of familial aggregation.[30,40]

If relatives of cases are genotyped for a variant, it is possible to assess the evidence that the variant is causal by investigating its segregation with disease within families.[41] This approach has the advantages of not being susceptible to population stratification, and of being immune to preferential selection of cases with a family history.

Genetic association studies

Family history may influence the design and size of association studies. Antoniou and Easton[6] showed that selection for genotyping of cases with an affected first-degree relative, as opposed to unselected cases, at least halves the required sample size in genetic association studies to detect a dominant disease allele, and the number is halved again if the selection is with two affected first-degree relatives. The use of such enriched samples may be particularly important in reducing the cost of whole-genome association studies of common diseases.

Stratification of risk by family history

Family history is often used in epidemiological studies to stratify analyses of the effects of other risk factors. For example, the odds ratios associated with many breast cancer risk factors, such as parity and age at menarche, are similar in women with and without a family history.[42] However, the Minnesota Breast Cancer Family Study,[43] which analysed the incidence of breast cancer over more than 40 years in families and considered risk factors in subsets defined by family history, has reported some intriguing findings. For example, the association of alcohol use with breast cancer risk may be limited to women with a family history of the disease;[43] family history may modify the association between obesity in early adolescence and subsequent breast-cancer risk;[44] and women who have ever used the early high-dose formulations of oral contraceptives available up until the mid-1970s, and who have a first-degree relative with breast cancer, may be at particularly high risk.[45] A case-control-family study in Germany recorded that parity was less protective in women with a strong family history.[24] These findings should not be over-interpreted, as overviews have not confirmed them.[42]

Although such analyses may help when advising individuals with a family history, they do not necessarily predict the effects of risk factors in individuals with any particular genotype. The reason is that family history is usually only a weak surrogate for genetic susceptibility,[46,47] and for most diseases familial risks will reflect the effects of many genetic and environmental factors.

Studying modifiers of risk in genetically-susceptible individuals

A major issue for genetic epidemiology is being able to identify factors, environmental and genetic, that modify risk in individuals who have, or are strongly suspected of having, inherited a disease-predisposing mutation. Without knowing how to reduce risk, or at least not increase risk, it is difficult to justify wide-scale genetic testing. Obtaining this information requires large numbers of known carriers and their relatives, but valid statistical inferences about modifiers of risk are difficult if ascertainment of the families, and the collection from them of information on potential modifiers, was not systematic and well designed. Population-based case-control-family studies are likely to give clearer answers than analyses of members of mutation-carrying families ascertained through opportunistic sampling from genetics clinics.

For example, for known *BRCA1* and *BRCA2* mutation carriers it is an open question whether lifestyle factors are positively or negatively associated with cancer risks. One clinically important example is oral contraceptive use. A population-based study found that current formulations of oral contraceptives were associated with a reduced risk of breast cancer in *BRCA1* carriers[48] and a clinic-based study found that that use of such contraceptives was associated with a reduced risk of ovarian cancer in *BRCA1* and *BRCA2* carriers combined.[49]

Prospective studies of risk

In time, case-control-family designs generate a cohort of unaffected individuals with wider distributions of underlying familial risk than in the usual cohort studies, where sampling is often based on widening the variation in environmental or lifestyle exposure(s) of interest and not on disease or family history. A family-based cohort of cases and their relatives will also, on average, be at higher risk for the disease of interest than an unselected or representative cohort, and may be enriched for

individuals at genetic risk. Such cohorts should be therefore more informative for studying genetic effects and for studying how environmental or lifestyle factors might modify genetic risks (often referred to as gene-environment interactions). Care must be taken to ensure that loss to follow-up is not biased, especially with respect to family history.[50]

Some analytical issues

Most statistical analyses of family data use likelihood methods. In this approach, a model is described which defines the distribution of the genetic factors in the general population and the way in which these factors are inherited as a function of unknown parameters (such as allele frequencies and hazard ratios[38]). The likelihood, which is simply the probability of recording the actually observed configuration of disease phenotypes, is computed for all families. The parameter values for this model are estimated as those that result in the greatest likelihood (this procedure is called maximum likelihood estimation). The fitted models are usually oversimplifications and the estimates may not be robust to incorrect model specifications.

One complexity in case-control-family designs is when not all genotypes of family members are available (eg, when some blood samples have not been obtained). The results can be biased if genotyping is not random because, for example, affected individuals were more likely to be genotyped.[51,52] The likelihood approach accommodates such complexity provided the calculations sum over all possible combinations of unmeasured genotypes among the family members consistent with the observed genotypes.

The statistical methods used include those most familiar to epidemiologists, such as logistic regression and Poisson regression models. With such models, log-odds of disease would be determined by contributions from genes, environment and their joint effect, and including adjustment for age, sex, and any other confounding factors. Because families include multiple generations and ages, simple stratification is less readily accommodated without further modelling.

An important principle in the analysis of family studies is ascertainment. Likelihoods must take into account the process by which the families included in the analysis were sampled. Failure to do so can lead to bias. For example, if the analysis is of relatives of cases carrying a specific variant, unselected for family history, the likelihood of each family must be conditional on the index case carrying the variant. The naive practice of simply excluding the variant-carrying index case does not guarantee unbiased estimates. Analysis of relatives of cases studied specifically because they have a family history is more problematic, though not intractable.[53] In theory, the likelihood approach provides a natural method for calculating unbiased estimates, though it may not always be straightforward.

In most analyses, likelihoods are calculated conditional on the observed family structure—ie, the precise configuration of age, sex, and genetic relationships. This approach therefore assumes that family structure per se does not contain information about the disease. Thus a study that involved genotypes associated with early loss in pregnancy would not readily be accommodated within this framework.

Statistical inference is limited to statements relevant to the sub-population sampled. Extrapolation requires untestable assumptions. A study of families based solely on selecting incident cases of a specific disease diagnosed in a given age range allows valid inferences about mutations that cause the disease in that age range. A study of families based on random sampling the whole population, irrespective of disease status, allows inferences about all mutations that exist in the population.[54] In practice, however, if the mutations are rare the latter might be nigh impossible and the results of little or no practical concern.[55]

Traditionally, genetic testing and counselling has been limited to multiple-case families. There are now emerging possibilities that cases can be efficiently targeted for mutation screening through their phenotypic characteristics. For example, there are pathological features that seem to be typical of breast cancers in *BRCA1* mutation carriers,[56–59] and immunohistochemistry and microsatellite instability testing of colorectal tumours can be used to target the testing of early-onset cases, irrespective of family history, for mutations in the DNA mismatch repair genes.[60] Therefore it is becoming much more relevant to study the characteristics of mutations that cause disease in a wider setting than multiple-case families attending a genetics clinic.

There is also the challenging statistical issue of how to obtain maximum information from case-control analyses that combine information from cases, family controls and population controls.[61] The risk estimated from a comparison of cases with population-controls is not necessarily the same as that from a comparison of cases with family controls, although in practice the difference may not be substantial.

Special case: twin pairs and twin families

Twin pairs represent a special, and historically widely used, family design. By studying both monozygotic and dizygotic pairs, and invoking the assumptions of the classic twin model,[62] the null hypothesis that genetic factors do not have a role in explaining variation in a trait can be tested. By studying the relatives of twins important extra information can be obtained.

The twin family design is an efficient way of teasing apart the effects of shared genes and shared environment. It is often built around twin pairs identified either at birth or a young age, and involves studying the biological parents of twins and sometimes

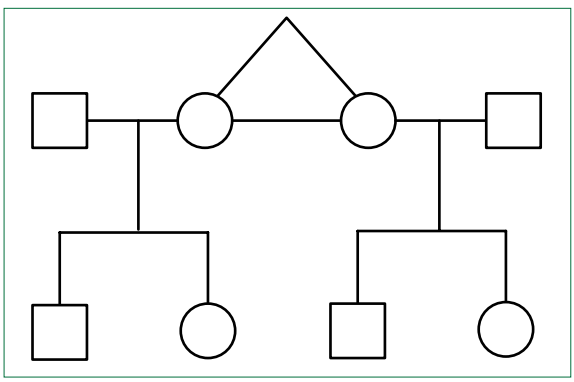

Figure 3: Twin family design, based on a monozygous pair and their offspring

their siblings, as in the Minnesota Twin Family Study.[63] Another design is to start with adult twins and study their offspring (figure 3).[64] Here, the offspring of one twin are genetically full siblings of one another and are presumably raised in the same environment, so they share on average half their genes as well as the family environment. The offspring of the other monozygotic twin are genetically half-siblings of the other children, so they share on average a quarter of their genes but do not share the same family environment. Bringing other pairs of relatives into such a design, or including the offspring of a dizygotic twin, will generate different patterns of shared genes and shared environment.

As discussed in the first paper of this series,[28] by analysing the correlations between different relative pairs under the paradigm first enunciated by Fisher,[65] the population variance can be broken down into both genetic and environmental components, common household effects, and maternal effects. The inclusion of differing relative types permits the examination of the consistency of familial aggregation with the assumed model of inheritance. For example, comparison of dizygotic twin pair concordance and sibling concordance allows quantification of the non-genetic effects shared by dizygotic twins.[66] Another example of this approach, this time for a continuous trait, is the study of twin and non-twin families in the Victorian Family Heart Study.[67] Analysis of a number of cardiovascular risk factors revealed evidence for environmental effects shared in childhood that persisted into adulthood, and consequently gave heritability estimates substantially less than those derived from studies of monozygotic and dizygotic twin pairs alone.[68] An example for survival times measured in families is the Busselton Cohort Study.[69]

The within-pair matching means that twin studies do not need to rely on external controls. Monozygotic pairs discordant for a disease provide the basis for perfectly matched case-control studies of environmental factors. Same-sex dizygotic pairs discordant for a disease, being matched for age and sex, are also useful for case-control analyses. Twin studies can be designed and implemented with several variations: (1) comparison of disease concordance allows estimation of the relative contribution of shared genes, shared family environment, and individual specific environment; (2) comparison of monozygotic pairs discordant for disease allows estimation of the effects of non-genetic exposures; (3) comparison of same-sex dizygotic pairs discordant for disease allows estimation of the effects of both genetic and environmental exposures, strengthened by the precise matching of the dizygotic twins for potentially confounding factors such as age and sex; and (4) comparison of within-pair differences in environmental and lifestyle factors in relation to within-pair differences in age at onset in monozygotic pairs concordant for disease, allows estimation of the effects of these risk factors in individuals at a higher than average genetic susceptibility.[70]

Sampling issues

Population-based sampling is an essential feature of these designs. This partly explains why the major examples to date have been studies of common cancers ascertained through population-complete cancer registries. The principles discussed above could be applied to hospital-based or other epidemiological studies, with the usual caveats.

A key issue with population-based family studies is obtaining a high and/or non-differential response from the relatives. Even within the same generation, relatives might differ in age and live in very different environments, so careful consideration needs to be given to whether restrictions should be made. This care applies to recruitment, blood sampling, genotyping, and eligibility for different statistical analyses. In practice, sensitivity analyses could be important for establishing the robustness of findings.

Many twin registries are available for, and even encourage, research from external users (eg, the special issue of *Twin Res* 2002; **5:** no 5). Few are population-based, most notably the Scandinavian registries,[71–75] although opportunities have been taken in other countries to establish population-based registries.[76–78] Population-based cohorts of young twins can also be generated from studies of birth cohorts, such as the Avon Longitudinal Study of Pregnancy and Children[79] and the Western Australian Twins and Child Heath Study.[80]

Volunteer-based registries can still be regarded as good approximations to population-based with respect to a particular trait if ascertainment is independent of factors related to that trait.[61] In practice, one can never be sure that this condition is satisfied for any given trait, and might be especially problematic for behavioural studies, and even if it proved true for some traits it may not necessarily hold for any other traits. Nevertheless, these ascertainment issues might not be important for the within-pair comparisons (1–5 above). The

implications of using volunteer twin samples need to be thought through for each application.

Practical advantages

Designs that involve interviewing relatives provide data of high quality with respect to family history. This finding was apparent for Parkinson's disease where self-reported family history overestimated familial risks by a factor of 2 (panel 2).[81] The inclusion of relatives also brings people at increased risk of disease into the research and adds a novel dimension, enabling better characterisation of the role of genes known to be or suspected of being involved with disease susceptibility by viewing relatives as a retrospective cohort defined from birth, and by generating a prospective cohort of individuals enriched for genetic risk who could themselves be the focus of research based on observational or intervention studies (panel 1).

The sample collection process has the potential for recruitment of case relatives, and is likely to be less affected than the recruitment of population-based controls is by data protection legislation which impinges on the identification of controls and thus on interpretability. Within the same framework, and indeed within the same sample, the effect on risk of both rare and common variants can be estimated. Studies need not be restricted to analyses of genetic effects. Analyses of non-genetic exposures (ie, environmental or lifestyle factors), as well as joint effects of multiple genes and environment, can all be accommodated within the same framework.

Potential limitations

A major limitation of population-based family studies is recruitment since family members can be spread widely across the country or even live abroad. This can pose challenges for collection of data and specimens. Cost will be a major consideration. In many circumstances, relatives may be more willing to participate than population-based controls and care needs to be taken to ensure that undue pressure is not placed on relatives because of their genetic relatedness and convenience. Other ethical issues must be considered, including whether and how genetic results should be offered to participants.[82] These issues are compounded by the fact that new knowledge is rapidly accumulating, so that an individual's opinion at the time of study entry may be less relevant later on when results become available.

Family designs have their most major limitations in populations with small family sizes, and especially for older-onset disease, although some of these can be ameliorated by the inclusion of more distant relatives. For instance, in such populations, studies of sex-specific disease will be limited by the lack of availability of unaffected (or affected) siblings of cases and hence recruitment might need to focus on both offspring (who would not be very useful for older age disease) and cousins (who are likely to be of the same generation but with less of a family focus).

Another potential limitation is analytical, due to the non-independence and possible incompleteness of data within families and the way families have been chosen for study. Likelihood theory can handle both incomplete genetic data, provided maternity and paternity are correctly assigned, and non-random ascertainment, provided the sampling rules are known and consistently applied.

So there are many practical and scientific issues that need to be considered in deciding on the family design for a given study. Well-designed and ethically approved pilot studies before any commitment to large investigations would seem an essential starting point.

The future

We have described the strengths and limitations of family study designs that build on population-based studies for making inferences relevant to the general population (table). These designs take advantage of the genetic structure of families to facilitate the investigation of genes and their role in disease. Although the large-scale application of these designs (eg, in breast and colon cancer family registries; panel 1) is relatively new, the motivation behind extending case-control studies to relatives is much older. Woolf recognised in 1955 that

Panel 2: The value of obtaining family disease history by interviewing relatives, not just probands

Many studies have addressed the accuracy, sensitivity, and specificity of self-reports of family history of disease. For common cancers, the positive predictive value (the probability that a self-reported family history is true) was estimated by reviewing the medical records of relatives.[7] This was done for reports both by case probands with the particular cancer and for reports from control probands without cancer. Murff et al[7] provide a breakdown by degree of relationship and cancer (and 95% CIs also) and the following is a simplified summary table restricted to first-degree relationships:

	Positive predictive value	
	Case proband	Control proband
Breast	93%	74%
Prostate	85%	68%
Colon	81%	71%
Ovarian	69%	25%
Endometrial	37%	17%

For common cancers the positive predictive values were high, but this was not so for the rarer cancers of ovary and endometrium. Case probands reported the family history of the cancer better than control probands did.

A study of Parkinson's disease found that cases (or their proxies) were more aware of the disease in their first degree relatives than were controls (or their proxies), and that case probands over-reported Parkinson's disease in relatives.[81] The odds ratio for the effect on Parkinson's disease risk associated with having a first degree relative with the disease was 4·34 when based on self-report of the proband or proxy alone but only 1·86 when based on validated information.

	Participants	Some uses	Limitations	Strengths
Case-control families	Population-based cases and controls only	Case-control comparisons of effects of family history. Stratification of effects by family history	Self-report of family history usually limited to first-degree relatives; often not validated. Limited power due to small proportion with family history	Not overly difficult, or resource intensive, to ask for a self report of family history
Case-families with or without population-based controls	Population-based cases and relatives; population-based controls	Case-control comparisons using family-based controls matched for unmeasured familial risks. Retrospective and prospective cohort study of case relatives	Relatives may live in different environments. Proportion of cases may not have any controls, and it may be difficult to obtain full participation from controls. May be difficult to interpret different estimates from using different control groups	Greater validity of data about relatives. Controls often well matched for potential confounders. Reduces issues around population-stratification in genetic association studies. May have increased power by combining control groups
Case-control-families	Population-based cases and controls, and their relatives	As above, plus cohort of relatives of controls for comparisons with cohort of relatives of cases	As above. May be difficult to get high participation from relatives of controls	As above. Symmetry in design can be used to address potential biases
Twin families	Twin pairs (monozygotic or dizygotic or both), with or without relatives	Test hypothesis of no genetic association. Tease apart shared genes from shared environment. Use of co-twins as controls	Population-based sampling difficult. Reduced power due to overmatching	Large twin registries exist and are available for research

Table: **Population-based family designs**

"ideally each living member of every family should be contacted",[83] and most of the major design points were identified by Clemmesen in 1965.[84] The Minnesota Breast Cancer Family Study of 1944–52 is one of the earliest applications.[19]

The best design for any particular situation is difficult to determine. The willingness of patients to identify and contact relatives and the willingness of those relatives to participate might be unknown when a study is designed, so a pilot study could be useful. Willingness might be population-specific and dependent upon the nature of the disease. Failure to enrol sufficient relatives will weaken the interpretability of the results just as data protection legislation and public scepticism about unwanted telephone calls and letters restrict the participation of true population-based controls. Contacting relatives may not be straightforward; the natural approach is to ask an enrolled participant to make that initial contact but this must be done within the ethical consideration that relatives do not feel compelled to participate. A family study design can also be influenced by costs. For example, ideally DNA would be collected from all relatives, but in practice non-random collection (eg, prioritising those affected) is more likely, and this must be taken into consideration in data analysis.

As discussed in the first paper of this series,[28] the combined effects of all the risk factors—correlated between relatives—that cause familial aggregation of a disease on a population-basis must be an order of magnitude greater than the average increased risk to first-degree relatives of affected individuals.[46,47] For most common diseases, this means that the group of individuals in the upper quartile of familial (and possibly mostly genetic) risk are at least 20 times the risk of the group of individuals in the lower quartile.

Uncovering all the familial risks due to shared genes and/or shared environment, and understanding how they interact, in a biological sense, with the classic environmental risk factors should increase our understanding of the causes of complex diseases. However, really important extra information may best be obtained from novel innovative genetic epidemiology studies using, for example, a variation on the case-control-family design, even though these could mean many more participants and much more financial support. High response rates from population-based sampling of controls are becoming more difficult to achieve; even in Utah the effort needed to achieve the same or slightly lower response rates has doubled over the past decade.[85] Therefore population-based case-family designs could be the future of epidemiology, not just genetic epidemiology. Because of their versatility, retrospective and prospective population-based family studies may become the principal framework for epidemiology in the future and move genetics from its traditional focus on so-called high-risk families to give it a wider clinical and population health relevance.

References

1 Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994; **266:** 66–71.

2 Kolodner RD, Hall NR, Lipford J, et al. Structure of the human MSH2 locus and analysis of two Muir-Torre kindreds for msh2 mutations. *Genomics* 1994; **24:** 516–26.

3 Hussussian CJ, Struewing JP, Goldstein AM, et al. Germline p16 mutations in familial melanoma. *Nat Genet* 1994; **8:** 15–21.

4 Tanzi RE, Gusella JF, Watkins PC, et al. Amyloid beta protein gene: cDNA, mRNA distribution, and genetic linkage near the Alzheimer locus. *Science* 1987; **235:** 880–84.

5 Vionnet N, Stoffel M, Takeda J, et al. Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. *Nature* 1992; **356:** 721–22.

6 Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: implications for design of association studies. *Genet Epidemiol* 2003; **25:** 190–202.

7 Murff HJ, Spigel DR, Syngal S. Does this patient have a family history of cancer? An evidence-based analysis of the accuracy of family cancer history. *JAMA* 2004; **292:** 1480–89.

8 Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, et al. Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Med* 2004; **1:** e65.

9 Hemminki K, Rawal R, Chen B, Bermejo JL. Genetic epidemiology of cancer: from families to heritable genes. *Int J Cancer* 2004; **111:** 944–50.

10 Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *J Natl Cancer Inst* 1994; **86:** 1600–08.

11 Hsu L, Prentice RL, Stanford JL. Some further results on incorporating risk factor information in assessing the dependence between paired failure times arising from case-control family studies: an application to prostate cancer. *Stat Med* 2002; **21:** 863–76.

12 Haile RW, Siegmund KD, Gauderman WJ, Thomas DC. Study-design issues in the development of the University of Southern California Consortium's Colorectal Cancer Family Registry. *J Natl Cancer Inst Monogr* 1999; **26:** 89–93.

13 Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002; **155:** 478–84.

14 Cordell HJC, Clayton DG. Genetic association studies. *Lancet* 2005; **366:** 1121–31.

15 Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361:** 598–604.

16 Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 2002; **11:** 505–12.

17 Hopper JL. Case-control-family designs: a paradigm for future epidemiology research? *Int J Epidemiol* 2003; **32:** 48–50.

18 Dite GS, Jenkins MA, Southey MC, et al. Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations. *J Natl Cancer Inst* 2003; **95:** 448–57.

19 Anderson VE, Goodman HO, Reed SC. Variables related to human breast cancer. Minneapolis: University of Minnesota Press, 1958.

20 Sellers TA, Anderson VE, Potter JD, et al. Epidemiologic and genetic follow-up study of 544 Minnesota breast cancer families: design and methods. *Genet Epidemiol* 1995; **12:** 417–29.

21 Sellers TA, King RA, Cerhan JR, et al. Fifty-year follow-up of cancer incidence in a historical cohort of Minnesota breast cancer families. *Cancer Epidemiol Biomarkers Prev* 1999; **8:** 1051–57.

22 Hopper JL, Chenevix-Trench G, Jolley DJ, et al. Design and analysis issues in a population-based, case-control-family study of the genetic epidemiology of breast cancer and the Co-operative Family Registry for Breast Cancer Studies (CFRBCS). *J Natl Cancer Inst Monogr* 1999; **26:** 95–100.

23 John EM, Hopper JL, Beck JC, et al. The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res* 2004; **6:** R375–89.

24 Becher H, Schmidt S, Chang-Claude J. Reproductive factors and familial predisposition for breast cancer by age 50 years. A case-control-family study for assessing main effects and possible gene-environment interaction. *Int J Epidemiol* 2003; **32:** 38–48.

25 Susser E, Susser M. Familial aggregation studies. A note on their epidemiologic properties. *Am J Epidemiol* 1989; **129:** 23–30.

26 Peto J, Collins N, Barfoot R, et al. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J Natl Cancer Inst* 1999; **91:** 943–49.

27 Risch N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol Biomarkers Prev* 2001; **10:** 733–41.

28 Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005; **366:** 941–51.

29 Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Hum Hered* 1971; **21:** 523–42.

30 Antoniou AC, Pharoah PD, McMullan G, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br J Cancer* 2002; **86:** 76–83.31.

31 Cui J, Antoniou AC, Dite GS, et al. After BRCA1 and BRCA2-what next? Multifactorial segregation analyses of three-generation, population-based Australian families affected by female breast cancer. *Am J Hum Genet* 2001; **68:** 420–31.

32 Antoniou AC, Pharoah PP, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br J Cancer* 2004; **91:** 1580–90.

33 Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am J Hum Genet* 1998; **62:** 145–58.

34 Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 2004; **23:** 1111–30.

35 Wacholder S, Hartge P, Struewing JP, et al. The kin-cohort study for estimating penetrance. *Am J Epidemiol* 1998; **148:** 623–30.

36 Struewing JP, Hartge P, Wacholder S, et al. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N Engl J Med* 1997; **336:** 1401–08.

37 Hopper JL, Southey MC, Dite GS, et al. Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study. *Cancer Epidemiol Biomarkers Prev* 1999; **8:** 741–47.

38 Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 2003; **72:** 1117–30.

39 Dunlop MG, Farrington SM, Carothers AD, et al. Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet* 1997; **6:** 105–10.

40 Cui JS, Spurdle AB, Southey MC, et al. Regressive logistic and proportional hazards disease models for within-family analyses of measured genotypes, with application to a CYP17 polymorphism and breast cancer. *Genet Epidemiol* 2003; **24:** 161–72.

41 Thompson D, Easton DF, Goldgar DE. A full-likelihood method for the evaluation of causality of sequence variants from family data. *Am J Hum Genet* 2003; **73:** 652–52.

42 Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *Lancet* 2001; **358:** 1389–99.

43 Vachon CM, Cerhan JR, Vierkant RA, Sellers TA. Investigation of an interaction of alcohol intake and family history on breast cancer risk in the Minnesota Breast Cancer Family Study. *Cancer* 2001; **92:** 240–08.

44 Cerhan JR, Grabrick DM, Vierkant RA, et al. Interaction of adolescent anthropometric characteristics and family history on breast cancer risk in a Historical Cohort Study of 426 families (USA). *Cancer Causes Control* 2004; **15:** 1–9.

45 Grabrick DM, Hartmann LC, Cerhan JR, et al. Risk of breast cancer with oral contraceptive use in women with a family history of breast cancer. *JAMA* 2000; **284:** 1791–98.

46 Hopper JL, Carlin JB. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. *Am J Epidemiol* 1992; **136:** 1138–47.

47 Peto J. Genetic predisposition to cancer. In: Cairns JL, Lyon LJ, Skolnick MH, eds. Banbury Report 4; Cancer Incidence in Defined Populations. New York: Cold Spring Harbor, 1980: 203–13.

48 Milne RL, Knight JA, John EM, et al. Oral contraceptive use and risk of early-onset breast cancer in carriers and noncarriers of BRCA1 and BRCA2 mutations. *Cancer Epidemiol Biomarkers Prev* 2005; **14:** 350–56.

49 Whittemore AS, Balise RR, Pharoah PD, et al. Oral contraceptive use and ovarian cancer risk among carriers of BRCA1 or BRCA2 mutations. *Br J Cancer* 2004; **91:** 1911–15.

50 Seybolt LM, Vachon C, Potter K, et al. Evaluation of potential sources of bias in a genetic epidemiologic study of breast cancer. *Genet Epidemiol* 1997; **14:** 85–95.

51 King MC, Marks JH, Mandell JB. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* 2003; **302:** 643–46.

52 Easton DF, Hopper JL, Thomas DC, et al. Breast cancer risks for BRCA1/2 carriers. *Science* 2004; **306:** 2187–91.

53 Scott CL, Jenkins MA, Southey MC, et al. Average age-specific cumulative risk of breast cancer according to type and site of germline mutations in BRCA1 and BRCA2 estimated from multiple-case breast cancer families attending Australian family cancer clinics. *Hum Genet* 2003; **112:** 542–51.

54 Begg CB. On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst* 2002; **94:** 1221–26.

55 Pharoah PD, Antoniou A, Hopper J, Easton D. Re: On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst* 2003; **95:** 75–76 and 77–78.

56 Lakhani SR, Jacquemier J, Sloane JP, et al. Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations. *J Natl Cancer Inst* 1998; **90:** 1138–45.

57 Lakhani SR, Van De Vijver MJ, Jacquemier J, et al. The pathology of familial breast cancer: predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, HER-2, and p53 in patients with mutations in BRCA1 and BRCA2. *J Clin Oncol* 2002; **20:** 2310–18.

58 Armes JE, Egan AJ, Southey MC, et al. The histologic phenotypes of breast carcinoma occurring before age 40 years in women with and without BRCA1 or BRCA2 germline mutations: a population-based study. *Cancer* 1998; **83:** 2335–45.

59 Armes JE, Trute L, White D, et al. Distinct molecular pathogeneses of early-onset breast cancers in BRCA1 and BRCA2 mutation carriers: a population-based study. *Cancer Res* 1999; **59:** 2011–17.

60 Southey MC, Jenkins MA, Mead L, et al. Use of molecular tumour characteristics to prioritize mismatch repair gene testing in early-onset colorectal cancer. *J Clin Oncol* 2005: **23:** 6524–32.

61 Whittemore AS, Halpern J. Logistic regression of family data from retrospective study designs. *Genet Epidemiol* 2003; **25:** 177–89.

62 Hopper JL. The epidemiology of genetic epidemiology. *Acta Genet Med Gemellol (Roma)* 1992; **41:** 261–73.

63 Hicks BM, Krueger RF, Iacono WG, McGue M, Patrick CJ. Family transmission and heritability of externalizing disorders: a twin-family study. *Arch Gen Psychiatry* 2004; **61:** 922–28.

64 D'Onofrio BM, Turkheimer EN, Eaves LJ, et al. The role of the children of twins design in elucidating causal relations between parent characteristics and child outcomes. *J Child Psychol Psychiatry* 2003; **44:** 1130–44.

65 Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 1918; **52:** 399–433.

66 Willer CJ, Dyment DA, Risch NJ, Sadovnick AD, Ebers GC. Twin concordance and sibling recurrence rates in multiple sclerosis. *Proc Natl Acad Sci USA* 2003; **100:** 12877–82.

67 Harrap SB, Stebbing M, Hopper JL, Hoang HN, Giles GG. Familial patterns of covariation for cardiovascular risk factors in adults: the Victorian Family Heart Study. *Am J Epidemiol* 2000; **152:** 704–15.

68 Evans A, Van Baal GC, McCarron P, et al. The genetics of coronary heart disease: the contribution of twin studies. *Twin Res* 2003; **6:** 432–41.

69 Scurrah KJ, Palmer LJ, Burton PR. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genet Epidemiol* 2000; **19:** 127–48.

70 Hamilton AS, Mack TM. Puberty and genetic susceptibility to breast cancer in a case-control study in twins. *N Engl J Med* 2003; **348:** 2313–22.

71 Skytthe A, Kyvik K, Holm NV, Vaupel JW, Christensen K. The Danish Twin Registry: 127 birth cohorts of twins. *Twin Res* 2002; **5:** 352–57.

72 Kaprio J, Koskenvuo M. Genetic and environmental factors in complex diseases: the older Finnish Twin Cohort. *Twin Res* 2002; **5:** 358–65.

73 Bergem AL. Norwegian Twin Registers and Norwegian twin studies: an overview. *Twin Res* 2002; **5:** 407–14.

74 Pedersen NL, Lichtenstein P, Svedberg P. The Swedish Twin Registry in the third millennium. *Twin Res* 2002; **5:** 427–32.

75 Rasmussen F, Johansson-Kark M. The Swedish Young Male Twins Register: a resource for studying risk factors for cardiovascular disease and insulin resistance. *Twin Res* 2002; **5:** 433–35.

76 Kendler KS, Prescott CA. A population-based twin study of lifetime major depression in men and women. *Arch Gen Psychiatry* 1999; **56:** 39–44.

77 Goldberg J, Curran B, Vitek ME, Henderson WG, Boyko EJ. The Vietnam Era Twin Registry. *Twin Res* 2002; **5:** 476–81.

78 Hansen J, Alessandri PT, Croft ML, Burton PR, de Klerk NH. The Western Australian Register of Childhood Multiples: effects of questionnaire design and follow-up protocol on response rates and representativeness. *Twin Res* 2004; **7:** 149–61.

79 Golding J, Pembrey M, Jones R. ALSPAC: the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol* 2001; **15:** 74–87.

80 Croft ML, Read AW, de Klerk N, Hansen J, Kurinczuk JJ. Population based ascertainment of twins and their siblings, born in Western Australia 1980 to 1992, through the construction and validation of a maternally linked database of siblings. *Twin Res* 2002; **5:** 317–23.

81 Elbaz A, McDonnell SK, Maraganore DM, et al. Validity of family history data on PD: evidence for a family information bias. *Neurology* 2003; **61:** 11–17.

82 Keogh LA, Southey MC, Maskiell J, et al. Uptake of offer to receive genetic information about BRCA1 and BRCA2 mutations in an Australian population-based study. *Cancer Epidemiol Biomarkers Prev* 2004; **13:** 2258–63.

83 Woolf CM. Investigations on genetic aspects of carcinoma of the stomach and breast. *Publ Public Health Univ Calif* 1955; **2:** 265–349.

84 Clemmesen J. Statistical studies in the aetiology of malignant neoplasms. I. Review and results. Supplement 174. I. *Acta Pathol Microbiol Scand* 1965; **54:** 1–543.

85 Rogers A, Murtaugh MA, Edwards S, Slattery ML. Contacting controls: are we working harder for similar response rates, and does it make a difference? *Am J Epidemiol* 2004; **160:** 85–90.