Epidemiological approaches to the immunogenetics of autoimmune rheumatic disease

Alexander J MacGregor, Alan J Silman

The familial clustering which occurs in certain autoimmune rheumatic diseases indicates an aetiological role for either shared genetic or shared environmental factors. Evidence implicating genetic factors in the aetiology of these diseases has come from the associations which have been described with specific antigens of the HLA system. These associations have a clear biological basis, given the central importance of HLA molecules in the immune system.

The application of genetic epidemiology attempts to quantify the relative importance of genetic factors in disease and is based on data derived from the study of disease associations in populations and in related individuals. This review considers the limitations and strengths of the methods used to collect and analyse data derived from these two approaches.

Studies of immunogenetic markers in diseased populations

These studies adopt the case control approach to compare the frequency of the exposure variable (the genetic marker) in those with disease (the cases) with that in a suitable control group. As with case control studies in general, population studies are susceptible to the effects of bias and confounding. These effects may arise (a) in the methods of selection of cases and controls, (b) in the methods of disease definition, and (c) from the methods used to analyse data.

Selection of study subjects

The origin of the study sample has an important influence on the measured strength of a genetic association. This has been illustrated in studies of rheumatoid arthritis (RA) in which a strong association has been described with DR4 in patients with seropositive erosive disease in samples selected from hospital inpatient clinics, whereas studies of patients in the community have not shown the same association.

A cross sectional sample of hospital clinic attenders may not be representative of all subjects affected by the disease. Patients with mild disease or disease which has been in remission for long periods and subjects with severe disease which has shortened survival will be underrepresented. A genetic association which is found in a hospital based sample may not therefore be generalisable to all subjects affected by the disease.

In addition to severity itself, it is important to consider whether specific manifestations of the disease under study rather than the occurrence of the disease itself have led to the inclusion of the patient in the sample. For example, it is more likely that patients with systemic disease in the autoimmune rheumatic diseases will be treated in a hospital setting. In RA a number of specific disease features have been associated with genetic markers. Examples include the association between erosiveness and HLA-DR4 and between extra-articular manifestations of disease and HLA-DR3. In a hospital sample a genetic association may therefore be more a reflection of specific aspects of a disease than of disease occurrence.

The selection of controls is equally important. Ideally controls should be selected from the same population from which the cases were identified. This is rarely practical. Typical sources of controls are cadaver kidney donors, blood donors, hospital staff, and clinic attenders unaffected by the disease under study. In most circumstances these form an adequate comparison group as it is unlikely that subjects would have been selected for a particular genotype. It is essential to ensure, however, that the genetic backgrounds of the case and control groups are similar. Controls must be of a similar ethnic origin to cases. There are significant ethnic differences in HLA type and patterns of linkage disequilibrium. The geographical region from which cases and controls are derived is also important. HLA antigen frequency varies within ethnic groups. Matching cases and controls by geographical region will not only take into account this regional variation in HLA within a particular group but will also have the effect of controlling for potential confounding factors in the environment.

The age of the control group should also be considered. The frequency of misclassification of disease status will be higher in control groups with a younger average age as a proportion of these will later develop disease. This problem is likely to be only a minor consideration in studies of the autoimmune rheumatic diseases which occur most commonly in the elderly.
particular age range. In contrast, in studies of the more common rheumatic diseases such as osteoarthritis, a significant proportion of younger subjects in a control group is likely to develop the disease later.

CASE DEFINITION
In the rheumatic diseases methods of case definition directly determine the strength of a measured genetic association. Disease status cannot be precisely defined. Most studies use classification criteria to assign disease status. Examples include the American College of Rheumatology classification schemes for systemic lupus erythematosus and RA.7 8 All such schemes result in the misclassification of disease status in a proportion of apparent cases and apparent controls. If disease is misclassified irrespective of genetic status (non-differential misclassification) the overall effect is to reduce the measured strength of any genetic association that is present. If the immunogenetic background influences the disease type, misclassification may introduce bias. The degree to which a genetic effect is underestimated through misclassification is directly dependent on the specificity and sensitivity of criteria used for disease definition.9

Most classification schemes for rheumatic diseases have been developed to maximise their efficiency for distinguishing disease in an outpatient clinic setting. A community sample is expected to contain fewer subjects with severe joint disease. In these circumstances identical criteria misclassify a greater proportion of true positive cases and correctly classify a greater proportion of true negative cases than they would in a hospital clinic setting. The criteria are less sensitive but more specific. For diseases with an incidence similar to that of RA it is possible to show that the underestimate of relative risk through misclassification depends more on the specificity than the sensitivity of criteria that are used for disease definition.10 This suggests that the same criteria are more likely to underestimate a genetic effect when used for disease definition in a hospital based sample than if used in a community sample, as misclassification of controls is likely to occur with a greater frequency in hospital based samples.

The degree to which relative risk is underestimated can be reduced if the specificity of disease classification criteria is increased. This would, however, restrict the representativeness of the study sample. In the absence of a gold standard for diagnosis the only practical solution to the problems presented by case definition is to confirm that associations found with genetic markers persist when disease classification criteria with different specificities and sensitivities are used in the same population.

ANALYSIS
As population studies of association are based on a case control design, results should be presented as odds ratios with appropriate confidence intervals. It is insufficient only to quote statistical significance as this is not in itself an estimate of the strength or biological relevance of an association.

Studies of genetic association often use multiple comparisons to test for associations with a large number of polymorphic alleles. Increasing the number of statistical tests performed increases the likelihood that one will be found to be significant at conventional levels of probability by chance alone. Furthermore, the prior probability of there being an association for markers selected at random is small. Rejecting the null hypothesis in favour of accepting that an association is present should therefore require more stringent critical probability values than are conventionally quoted.11 Confidence interval boundaries should be adjusted appropriately.

A further important consideration when investigating associations with several markers is that a strong association with one marker will create misleading deviations in the frequency of other markers. Strong associations with one particular allele may mask a second important positive association or create the misleading impression of a negative association with other alleles. Estimating the importance of these relative predisposition effects cannot readily be achieved using a simple odds ratio based approach and requires alternative methods of analysis.12 The absence of a genetic association should also be interpreted with caution. It is important to take the statistical power of a study into account before dismissing the possibility that an association is present. Again, this is particularly important in the case of HLA associations where rare alleles may be present in small numbers despite a large population sample. Furthermore, the region of the chromosome tracked by a genetic marker in a population study is small. The absence of an association in a population should not be used as a basis for dismissing the possibility of genetic linkage.

Finally, associations with particular markers need to be interpreted within the context of known patterns of association of genes. Many genes of the major histocompatibility system are held in linkage disequilibrium. Knowledge of the patterns of linkage disequilibrium need to be considered when assessing the possible contributions of individual genetic markers to disease susceptibility.

Family studies
Disease clustering within families allows the investigation of a number of aspects of genetic association. From the epidemiological viewpoint it is possible to assess the relative importance of genetic over environmental factors in disease occurrence in subjects where the two factors are shared to a different extent. From the genetic viewpoint the distribution of disease within families provides the opportunity to track markers of disease susceptibility. In particular, it is possible through family
studies to determine the composition of the genotype of affected probands and hence recognise groups of genes with which disease susceptibility is associated. Understanding the distribution of genotypes among affected and unaffected family members allows formal genetic analysis of the mode of inheritance of the disease. Family studies can be used to quantify physical linkage between markers and disease susceptibility genes. This forms a basis for more detailed characterisation of these genes at the molecular level.

**STUDY DESIGNS**

The study of disease patterns in monozygotic and dizygotic twins is the classical approach to studying the relative importance of genetic and environmental influences in disease. The magnitude of disease concordance in monozygotic twins indicates the maximum extent of the genetic influence. Excess disease concordance rates in monozygotic over dizygotic twins indicate the relative contribution of genetic over environmental factors. In this context families can be used as the basis of case control studies in which cases and controls may be selectively matched by the extent to which they share a common genetic or environmental background. This approach can be extended to siblings and more distant family members when the extent of genetic sharing is known. In addition, disease occurrence can be studied in families followed as cohorts. For example, showing similarities in the timing of disease onset by calendar year indicates the predominance of environmental influences whereas similarities in disease onset by age indicate predominating genetic influences.

Several strategies have been developed for tracking genes within families to assess the relationship with disease. Models for the mode of inheritance of disease (autosomal or X linked, dominant or recessive) may be assessed by the study of family phenotype data in cosegregation analysis. Linkage between disease and genetic markers is most readily estimated by measuring haplotype sharing in affected sibling pairs. In this approach, linkage is quantified by the distortion of the distribution of markers within siblings from the distribution that would be expected under random segregation. More complex studies of genetic linkage include lod score analysis and location score analysis. In lod score analysis information from markers within genotypes of affected and unaffected subjects in extended pedigrees is used to assess the odds in favour of their being a genetic linkage at a particular recombination distance. Location score analysis combines data from multiple markers to estimate the likelihood of a series of possible positions for a disease locus within a framework of fixed marker loci.

Despite their potential, family studies have a number of important shortcomings when applied to the study of autoimmune rheumatic disease. There is a debate as to whether familial disease is similar to the disease that occurs sporadically in unrelated subjects. The heterogeneity of disease patterns found within families has been given as evidence to indicate that this is unlikely to be so. There are further specific problems in case selection and ascertainment and in disease definition and analysis.

**SELECTION AND ASCERTAINMENT OF STUDY SUBJECTS**

The effect of disease severity bias is as important in family studies as it is in studies based in populations. Family members recruited through hospitals are more likely to be affected by severe disease than those found in community surveys.

Families with several affected members are more likely to be recruited into a study. Moreover, the numerical contribution from subjects from these families may be disproportionately large. Methods are available in analysis to assign appropriate weights to data according to the method by which families were ascertained. These methods are essential in segregation analysis to ensure that information from multicause families is not overrepresented.

**CASE DEFINITION**

Within individual families the unaffected members form the group with which cases are compared. Misclassification of a case necessarily results in misclassification of a control and vice versa. Misclassification is likely to be non-differential with respect to a subject's genetic status. The overall effect of this process is to reduce the strength of any genetic association that is measured.

Family studies are more prone to the effects of misclassification errors than are population studies where the case and control groups are selected independently. Existing criteria are not designed to define the absence of disease and their use may be inappropriate in family studies where minor or transient symptoms in subjects may represent important evidence of disease.

**ANALYSIS**

Twins with disease, affected siblings, and families with several affected members are rare. Large numbers of these groups are needed to provide precision for adequate statistical analysis. Lod score analysis, for example, is critically dependent on the number of cases included in a study for diseases where gene penetrance is low. It has not yet been possible to use lod score analysis to show significant linkage of any marker to a putative disease susceptibility gene in RA.

**Conclusions**

Advances in the understanding of the relationship between the HLA system and disease and the description of polymorphisms spanning the genome have greatly increased the potential for understanding the genetic
Epidemiological approaches to the immunogenetics of autoimmune rheumatic disease

Epidemiological approaches to the immunogenetics of autoimmune rheumatic diseases. Powerful analytical methods are available to translate information from the study of the distribution of disease and markers into physical maps of disease susceptibility genes. This holds the future promise of isolating these genes and their products. Important issues such as the methods of selection of patients and controls, methods of ascertainment of families, and the classification of disease make the search for genetic association particularly complex when applied to the autoimmune rheumatic diseases, however. Furthermore, the rarity of familial material and the low penetrance of disease susceptibility genes mean that future advances may only be made through coordinated research programmes on a large scale.