Statistical methods in genetic research on smoking

Andrew C Heath, Pamela AF Madden and Nicholas G Martin

Stat Methods Med Res 1998 7: 165
DOI: 10.1177/096228029800700205

Statistical Methods in Medical Research 1998; **7**: 165–186

Statistical methods in genetic research on smoking

Andrew C Heath and **Pamela AF Madden** Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri, USA and **Nicholas G Martin** Division of Epidemiology and Population Health, Queensland Institute of Medical Research, Brisbane, Australia

A growing body of evidence suggests that genetic factors have an important influence on the onset and course of smoking. Here we review some of the statistical methods that have been used to test for genetic influences on smoking behaviour, with a particular focus on studies of large national twin samples. We show how many of the hypotheses that have been tested using a genetic model-fitting approach have also been reformulated using logistic regression models that will be more familiar to epidemiologists. Such an approach is more easily extended to allow for sociocultural, as well as genetic, influences on smoking behaviour. Using either approach, data are consistent in indicating that certainly in men, and possibly in women, genetic factors play an important role in predicting which individuals who become cigarette smokers progress to being long-term persistent smokers.

1 Introduction

In 1958, a controversial article published by RA Fisher[1] argued, on the basis of data from a small number of twin pairs, that propensity to smoke cigarettes was partially influenced by genetic factors and further postulated that smoking-disease associations might not reflect a simple cause-and-effect relationship, but rather a tendency for the same genetic factors that made some individuals more prone to be smokers also to make them more disease-prone. Fisher's work helped stimulate a series of large sample twin studies in Scandinavia and the USA[2–4] that were in part designed to falsify Fisher's hypothesis of an indirect association: if twins who were smokers had higher rates of disease than their twin siblings who were nonsmokers, even in the case of twin pairs who were monozygotic (MZ) (i.e. genetically identical), this would provide a very convincing refutation of Fisher's hypothesis. While few would now accept Fisher's second postulate, data gathered in the intervening 40 years provide remarkably consistent support for his first assertion. Paradoxically, however, because the possibility of genetic influences on smoking behaviour became associated with Fisher's argument against the harmful effects of smoking, that genetic differences between individuals might be important predictors of failure of smoking cessation efforts (and thus, when identified and understood, might lead to improved aids for smoking cessation) received little attention.

How can there be genes that influence smoking behaviour, when sociocultural influences on smoking are clearly so important? There are huge variations in rates of smoking between societies and within societies over time, as well as pronounced gender differences within some societies at some time points. Consideration of

Address for correspondence: AC Heath, 40 N. Kingshighway, Suite 1, Saint Louis, MO 63108, USA. E-mail: andrew@matlock.wuSTL.edu

 0962-2802(98)SM155RA

progress in genetic research on alcoholism, where some of the ways in which genetic differences can lead to differences in risk are better understood, supports the plausibility of genetic influences on smoking behaviour. Levels of alcohol consumption, like rates of smoking, show substantial variation over time (e.g. in the USA between the era of prohibition and the present) and between societies (e.g. between Western and Islamic societies). Nonetheless, several lines of evidence have led alcoholism researchers to the view that there are important genetic influences on alcoholism risk. In research on rodents, studies of inbred rodent strains that have been bred to be genetically identical have shown strain differences in the degree to which different strains will self-administer alcohol when given a choice between alcohol and water,[5] while selection experiments have shown that rats can be selectively bred for high versus low voluntary alcohol consumption.[6,7] Twin and adoption studies have demonstrated increased rates of alcoholism in the monozygotic cotwins of alcoholics, compared to dizygotic (DZ) cotwins of alcoholics, and increased rates of alcoholism in the adopted-away offspring of alcoholic biological parents, compared to control adoptees.[8] High-risk studies, comparing offspring of alcoholic parents and controls, have established differences in subjective ratings of alcohol's effects and objective measures of body-sway and hormonal changes, after administration of a standard body-weight adjusted dose of alcohol, between individuals at low versus high risk of alcoholism and have shown that individuals with low initial sensitivity to alcohol are more likely to develop subsequent problems with alcohol.[9] Finally, geneticists have identified in individuals of Asian ancestry polymorphisms that are associated with effects on the metabolism of alcohol, that lead to large differences in drinking patterns and alcoholism risk.[10,11] Early reports of positive genetic linkage findings in samples of European ancestry are beginning to emerge.[12] Advances in genetic research on alcoholism have in turn made important contributions to the development of pharmacotherapies for alcoholism.[13]

In the present paper, we review statistical methods that have been used to establish an important genetic influence on smoking. Wherever possible we use analyses of published data to illustrate data analytic approaches. While we begin with a discussion of genetic model-fitting techniques that may be unfamiliar to many readers, we then proceed to show how these same hypotheses have also been tested in a more familiar logistic regression framework.

2 Example data sets

By way of illustration, we will use three data sets based on mailed questionnaire surveys of large national twin panels conducted in Finland, Australia and the USA (see Table 1 for sample sizes and further details). In each survey, smoking was assessed as a risk-factor, so only limited data about whether an individual had ever smoked cigarettes, and was still smoking cigarettes, together with information about the age-of-onset of smoking, the age the respondent quit smoking and the respondent's current or previous typical number of cigarettes smoked per day, were ascertained. Table 2 summarizes the numbers of twin pairs concordant or discordant for smoking status (current, former or nonsmoker) from one of these three surveys, the survey of the

Table 1 Sample sizes for national surveys of the Swedish, Finnish, Australian and US Vietnam-era veterans twin panels

Sample	Sample sizes (pairs)					Reference
	Survey date	MZ male	DZ male	MZ female	DZ female	
Finnish twin panel	1975	1496	3440	1842	3703	Kaprio *et al.*[4]
Australian twin panel	1980–82	567	350	1232	747	Heath and Martin[14]
US Vietnam-era veterans (VETS) panel	1987	2204	1793	–	–	True *et al.*[15]

Table 2 Numbers of twin pairs concordant and discordant for smoking status in the Australian twin panel 1981 survey

		MZ female (N = 1232 pairs)			DZ female (N = 747 pairs)			MZ male (N = 567 pairs)			DZ male (N = 350 pairs)		
		I	II	III	I	II	III	I	II	III	I	II	III
I	Non-smoker	629			310			221			121		
II	Successful quitter	110	64		98	33		77	70		44	27	
III	Current smoker	124	115	190	146	61	99	31	61	77	61	53	44

Australian twin panel conducted in 1980–82.[14] Similar data summaries may be derived from the original publications on the Finnish[4] and US Vietnam-era veterans twin-panel surveys.[15]

2.1 Summary statistics

Traditionally, twin researchers have used as summary statistics estimates for each zygosity group of lifetime prevalence, which is the proportion of the sample that report that they have ever been smokers, and the so-called probandwise concordance rate,[16] which is the probability that the cotwin of a smoker will also be a smoker, estimated in the case of a survey of a general population sample of twin pairs as $2C/(2C + X)$, where C is the number of pairs concordant for smoking, and X is the number of smoking-discordant pairs. Evidence for a significantly higher concordance rate in MZ pairs than in DZ pairs has been used to infer genetic influences on a trait, under the supposition that the environments experienced by MZ pairs (e.g. parental smoking and smoking by peers) are no more highly correlated that the environments experienced by DZ pairs. 'Pairwise' concordance rates, estimated as $C/(C + X)$, have sometimes also been reported in the literature but are clearly redundant, since the number of twin pairs for each zygosity group, the lifetime prevalence, and the probandwise concordance rate, are sufficient to completely describe for any binary trait the observed numbers of concordant 'unaffected' (e.g. concordant never smokers), discordant and concordant affected (e.g. concordant for having smoked) pairs. If N is the observed number of twin pairs, P is the estimate of lifetime prevalence, and CR the probandwise concordance rate for a given zygosity group, then the number of concordant affected pairs is estimated as $N \times P \times CR$, the number of discordant pairs as $2N \times P \times (1 - CR)$ and the number of concordant unaffected pairs as

$N - N \times P \times CR - 2N \times P \times (1 - CR)$; and the pairwise concordance rate is simply $CR/(2 - CR)$. In the case of data from Table 2, we obtain estimates of the lifetime prevalence of smoking (± standard deviation of this estimate) of 39.5% ± 1.3% in MZ female twins, 42.2% ± 1.5% in DZ female twins, 46.2% ± 1.9% in MZ male twins, and 50.4% ± 2.2% in DZ male twins, with corresponding probandwise concordance rates of 75.9% ± 1.5%, 61.3% ± 2.4%, 79.4% ± 1.9% and 70.3% ± 2.8%, respectively. If we limit consideration to those pairs where both twins have been smokers, corresponding estimates of prevalence and probandwise concordance for smoking continuation or 'persistence' (i.e. whether or not a twin was still a smoker when surveyed) are 67.1% ± 1.9% and 76.8% ± 2.0% in MZ female pairs, 67.1% ± 2.8% and 76.4% ± 3.1% in DZ female pairs, 51.7% ± 2.9% and 71.6% ± 3.5% in MZ male pairs, and 56.9% ± 3.4% and 62.4% ± 5.0% in DZ male pairs.

Estimates of prevalence derived from data on twin pairs violate the assumption of statistical independence of observations, so that the usual formula for deriving the standard deviation of the estimate of a proportion cannot be applied, and one term ($2C$) appears in both the numerator and the denominator of the formula for the probandwise concordance rate, making estimation of the standard deviation of the estimate of this statistic somewhat complicated. Instead we have used the method of bootstrapping[17] to obtain estimates of the standard deviations of these statistics. This approach can be applied more generally to the analysis of family data or other data involving complex clustered sampling schemes, as well as to the development of empirical estimates of standard deviations for statistics whose sampling distribution is not known. It involves drawing some large number (e.g. 1000) of random samples, with replacement, from the observed data, using the twin pair as the unit for resampling. Thus, in the case of data on female MZ like-sex pairs, we drew 1000 samples of 1232 pairs. Because we sampled with replacement (i.e. the same observation could appear multiple times, or not at all, in a given sample), estimates of prevalence and probandwise concordance rate varied across samples; the standard deviations of these estimates were used as the desired empirical estimates of the standard deviations of our summary statistics. Some statistical packages (e.g. STATA, S-PLUS) already include a built-in option for bootstrapping, while in others (e.g. SAS) bootstrapping is easily programmed via a function for random number generation.

3 Model-fitting approaches

Despite its utility as a summary statistic, the probandwise concordance rate has no direct interpretation in terms of genetic and environmental effects. The exception to this is the case of a single-gene recessive trait where penetrance (i.e. the probability that an individual whose genotype puts him at risk of the disorder will express the disorder) is uncorrelated over family members, where the concordance rate in MZ pairs provides a direct estimate of penetrance. Smoking is clearly not a single gene recessive trait! If there is no familial resemblance for a binary trait, the probandwise concordance is expected to be equal to the prevalence of that trait, so it is not surprising that prevalence as well as probandwise concordance rate must be taken into account when drawing inferences about the genetic or environmental effects on smoking. In particular, this sensitivity to differences in prevalence makes the con-

cordance rate an inappropriate summary statistic for pooling results across gender, birth cohorts or across different societies in which rates of smoking may vary widely. The recurrence risk-ratio, the ratio of the concordance rate to the prevalence rate in the general population, has proved to be a very informative statistic in genetic research on complex traits of relatively low prevalence.[18–21]For the Australian data set of Table 2, corresponding recurrence risk-ratios for lifetime smoking and for smoking persistence among smokers, with bootstrapped standard errors, are: MZF: 1.92 ± 0.06, 1.14 ± 0.03; DZF: 1.45 ± 0.06, 1.14 ± 0.04; MZM: 1.63 ± 0.06, 1.39 ± 0.06 and DZM: 1.39 ± 0.06, 1.10 ± 0.07. However, for multifactorial models (i.e. allowing for the influences of multiple genetic and environmental effects) for traits which may differ widely in prevalence between population groups, this statistic is still very sensitive to differences in prevalence (Ridenour TA and Heath AC. Meta-analysis for behavioural genetic studies of dichotomous phenotypes, unpublished data).

An important advance in genetic research on smoking occurred through the work of Eaves, beginning in the 1970s,[22,23] using an insight by Pearson[24] that was rediscovered independently by Falconer.[25,26] If variation in a continuous trait is determined by the additive effects of even a quite small number of genes or environmental risk-factors that occur with relatively high probability, the distribution of that trait closely approximates a normal distribution.[27] It is not unreasonable to hypothesize that in the case of a binary trait such as smoking, propensity to start smoking, or propensity to continue in the smoking habit once smoking is started, are latent (i.e. not directly observable) variables (traditionally referred to as 'liability' variables) which are approximately normally distributed.

Human genetic research in the biometrical or quantitative genetic tradition[28–30] has shown how familial resemblance for quantitative traits could be modelled using genetic and environmental variance components.[30,31] Specifically, if $p0$, $p1$ and $p2$ denote the probabilities that a pair of relatives will have zero, one or two alleles at any autosomal genetic locus that are identical by descent, then their expected correlation for a quantitative trait is given by $1/VP\ \{RVA + p2\ VD + R^2\ VAA + p2^2\ VDD + Rp2\ VAD + VC\}$, where VP is the total phenotypic variance, $R = (0.5\ p1 + p2)$ is the coefficient of genetic relationship between the relatives, VA is the additive genetic variance, VD is the dominance genetic variance, VAA is the additive × additive epistatic variance, VDD is the dominance × dominance epistatic variance, VAD the additive × dominance epistatic variance (ignoring higher order epistatic terms) and VC is the variance due to environmental effects shared by the relatives.[32] Two alleles are said to be identical by descent 'if one of them has been derived by direct replication from the other or if both are copies of the same gene in a common ancestor'.[32] Thus, a child will share exactly one allele identical by descent with a biological parent, except in cases of inbreeding, but a pair of full siblings may share none, one or two alleles identical by descent. Here $VP = VA + VD + VAA + VDD + VAD + VC + VE$, where VE is the variance due to environmental effects that are not shared by relatives. Maximum-likelihood estimates of model parameters can be obtained by fitting models to summary covariance matrices in multigroup analyses using standard software for structural equation modelling.[33] In most practical applications, epistatic genetic variance components will be confounded with genetic dominance. While this basic decomposition of the observed phenotypic variance in a

trait into components due to additive and nonadditive genetic effects and shared and nonshared environmental effects was subsequently elaborated to allow for more complex models for genotype–environment correlation, genotype × environment interaction, environmental contributions to parent–offspring resemblance and assortative mating effects,[34–36] it remains central to subsequent quantitative genetic research on smoking.

In the case of a binary trait such as lifetime smoking, or a unidimensional categorical (i.e. ordinal) trait, while we clearly cannot treat the observed measure as though it were a continuous variable, it is possible, following Pearson and Falconer, to estimate genetic and environmental variance components for the hypothesized latent 'liability' variable, using the assumptions that (1) this liability variable is normally distributed, with the observed categorical distribution determined by abrupt thresholds on that underlying latent distribution; and (2) that the distribution of liability in pairs of relatives is bivariate normal.[22,23] Similar assumptions are used in the estimation of tetrachoric and polychoric correlations by software packages such as PRELIS.[37] For the general case of an n-category ordinal variable, data on relative pairs of a given type will be summarized as an $n \times n$ contingency table. Model parameters will be correlations between relatives (or genetic and environmental variance components from which predicted correlations between relatives are derived), and $n-1$ threshold values (assuming no differences in prevalence between relative types), scaled as normal deviates, such that individuals with liability scores $\infty < s_i \leq t_1$ are assumed to fall into response category one, those with liability scores $t_1 < s_i \leq t_2$ fall into response category two, and so on. For given parameter values, expected probabilities for the cells of each $n \times n$ contingency table may be derived by integrating the bivariate normal distribution, with predicted correlation between relatives for the latent-liability variable ('polychoric' correlation) ρi. The log-likelihood of the observed data (ignoring the constant term) is computed as

$$L = \sum_i \sum_j \sum_k f_{ijk} \ln p_{ijk}$$

where f_{ijk} is the observed frequency of relative pairs from the ith group (e.g. MZ pairs) in the j,kth cell of the ith observed contingency table, and p_{ijk} is the corresponding expected probability. Maximum-likelihood estimates are obtained by maximizing the log-likelihood of the observed data with respect to the model parameters, a task which can now be handled by some statistical software packages (e.g. MX).[38] The approximate sampling covariance matrix of the parameter estimates, from which standard errors may be obtained, can be derived as the inverse of the Fisher information matrix, whose m,nth element will be[39,40]

$$\sum_i \sum_j \sum_k \frac{Ni}{p_{ijk}} \frac{dp_{ijk}}{d\theta_m} \frac{dp_{ijk}}{d\theta_n}$$

where θ is the vector of parameter estimates (however, this does not appear to have been implemented in MX).

A likelihood-ratio chi-square test of the goodness-of-fit of a given model to the observed data is computed as $2\,(L0 - L)$, where $L0$ is the log-likelihood of the observed

data under a perfect fit model which equates the expected probabilities for each cell to the corresponding observed probabilities, and L the log-likelihood under the fitted model at the maximum-likelihood solution. The number of degrees of freedom for this chi-square statistic will be equal to the number of observed statistics ($2(n^2-1)$) in the case of MZ and DZ like-sex twin pair data on an n-category ordinal scale) minus the number of estimated parameters. Software packages such as MX typically print this likelihood-ratio 'goodness-of-fit' chi-square statistic without printing the estimated log-likelihoods.

Likelihood-ratio tests may be used to compare nested models (e.g. a model that estimates additive genetic and shared and nonshared environmental effects, compared to a model that fixes the shared environmental parameter to zero), by subtracting the goodness-of-fit chi-square under the more general model from that under the reduced model, with degrees of freedom for this chi-square statistic equal to the difference in degrees of freedom for the two nested models. However, it is clearly not very helpful to report only a point estimate for the proportion of the total phenotypic variance that is attributable to genetic effects (e.g. 57%) when that estimate, while significant, has 95% confidence limits of 2%–98%![41] It has, therefore, become standard to report also approximate likelihood-based 95% confidence intervals for estimates of genetic and environmental parameters, corresponding to those values of each parameter that produce a change in chi-square, compared to the maximum-likelihood solution, of 3.84. Neale and Miller[42] discuss technical aspects of the estimation of such confidence intervals, which is implemented in MX.[38]

In the case of Australian twin data on smoking, by way of illustration, we have, therefore, estimated *VA*, *VC* and *VE* variance components for measures of smoking behaviour, ignoring nonadditive genetic effects. In the absence of data on separated twin pairs, the effects of genetic nonadditivity (dominance or epistatis) and shared environment are, strictly speaking, confounded in twin data; the former will produce DZ correlations that are less than one-half the corresponding MZ correlation, and the latter DZ correlations that are greater than one-half the corresponding MZ correlation. Negative estimates of shared environmental variance components, therefore, imply the presence of genetic non-additivity, and vice versa. In practice, non-additive genetic variance components are generally small, allowing the detection of shared environmental influences.

Expected correlations between MZ and DZ pairs will be $VA + EC$ and $0.5\,VA + EC$, respectively, where the total phenotypic variance ($VA + VC + VE$) is standardized to unity. For the male like-sex pairs, the goodness-of-fit chi-square statistics were: (1) VE–VC model (degrees of freedom = 2): $\chi^2 = 11.28$, $p = 0.02$; (2) VE–VA model (degrees of freedom = 2): $\chi^2 = 7.35$, $p = 0.12$; and (3) VE–VC–VA model (degrees of freedom = 1): $\chi^2 = 0.43$, $p = 0.94$). (The degrees of freedom here, and in subsequent examples, have been adjusted to allow for the fact that we used a single summary statistic for numbers of smoking discordant pairs.) Although the VE–VC–VA model has one degree of freedom, this tests the assumption of equal thresholds in MZ and DZ pairs; in the case of a binary trait, we have no information with which to test the appropriateness of the assumption of an underlying normally distributed latent variable. Comparing the fit of models (1) and (3) confirmed significant evidence for apparent additive genetic effects on smoking onset (likelihood-ratio $\chi^2 = 10.86$, df =

1, $p < 0.001$); while comparing the fit of models (2) and (3) confirmed significant evidence for shared environmental effects on smoking onset (likelihood-ratio $\chi^2 = 6.92$, df = 1, $p < 0.001$). Estimated genetic and environmental variance components, and their 95% confidence interval, were *VA* 43% (17%–72%); *VC* 37% (10%–60%); *VE* 20% (14%–27%). A similar pattern of findings emerged from analyses of the female like-sex twin pair data, except that the hypothesis of no shared environment effects could not be rejected (likelihood-ratio $\chi^2 = 3.24$, df = 1, $p = 0.07$). Corresponding estimates of genetic and environmental variance components, and their 95% confidence interval, were: *VA* 63% (44–84%); *VC* 18% (0–36%); *VE* 19% (15–23%).

Such estimates must be interpreted with caution. First, they are valid only if the assumption of an underlying normal-liability distribution is at least approximately true.[43] Second, any tendency for MZ pairs to be more highly correlated in their environmental exposures than DZ pairs may cause overestimation of the magnitude of the genetic variance component. There is evidence that MZ pairs are more likely to share the same peers when growing up than DZ pairs, however, when similar analyses were computed separately for pairs who reported that they always or usually shared the same friends when growing up, and for pairs who rarely or never shared the same friends, significant evidence for genetic effects on onset of smoking was still obtained (Madden PAF *et al.* The genetics of smoking initiation and persistence: a cross-cultural trait study, unpublished manuscript). This approach of analysing a twin pair or other familial data which is conditional upon values of a dichotomous environmental variable (e.g. using separate groups of pairs concordant for exposure to a low risk environment, concordant for exposure to a high-risk environment, and where available pairs discordant for environmental exposure), with testing for heterogeneity of parameters between high-risk versus low-risk exposure conditions, provides one way in which genotype × environmental interaction effects (including the special cases of genotype×sex or genotype × cohort interaction) can be detected.[44]

3.1 Hierarchical/conditional genetic models

New challenges arise when we shift the focus from analyses of genetic influences on onset of smoking to the arguably more important question of whether there are genetic influences on probability of persistence of smoking in those who have started to smoke. Most simply, we could limit analyses to pairs where both twins are smokers, analysing persistence of smoking as a binary trait, as before.[45] In the case of data from the Australian twin study, this approach yields estimated variance components in men (with 95% confidence limits) of: *VA* 59% (15–73%), *VC* 0% (0–38%); *VE* 41% (27–58%); and, in women, *VA* 3% (0–54%); *VC* 43% (0–57%); *VE* 54% (41–67%). Thus, from the analyses in men, there is significant evidence for a genetic influence on smoking persistence, though we cannot exclude the possibility of a strong shared-environment influence as well. In women, the best-fitting model yields only a small nonsignificant estimate for the genetic parameter, but we also cannot reject the hypothesis of no shared-environmental effects, with a heritability estimate as high 54%! But is throwing away data from twin pairs, where only one twin is a smoker, the appropriate thing to do? If it is the case that at least some of the same familial factors that determine risk of smoking persistence also influence risk of becoming a smoker, then discarding nonsmokers would be expected to lead to biased estimates of genetic

and environmental parameters, since it will systematically eliminate the most discrepant pairs (which, if there are genetic influences on onset of smoking, will be disproportionately dizygotic pairs).

Eaves[23] had the important insight that it was necessary to test hypotheses about the relationship between genetic influences on onset of smoking, and genetic influences on persistence in the smoking habit. He considered two extreme cases: (1) a simple extension of the liability threshold model ('single-liability dimension' model), which assumes that continuing smokers, on average, have higher liability than successful quitters, while nonsmokers have the lowest liability; and (2) independent genetic determinants of onset and persistence of smoking, as well as independent environmental determinants. The first case represents a simple extension of the liability threshold model to the case of three response categories. However, when we applied this model to the Australian data, even when both additive genetic and shared environmental effects, as well as nonshared environmental effects, are included in the model, it gave a very poor fit to the data (men: $\chi^2 = 49.26$, df = 6, $p < 0.001$; women: $\chi^2 = 43.47$, df = 6, $p < 0.001$).

The second case extends the basic liability threshold model by hypothesizing that there are two orthogonal liability dimensions, the first of which determines the onset of smoking, the second the continuation of smoking in those who have started to smoke (described by Eaves as 'smoking persistence'). This implies that two sets of genetic and environmental parameters may be estimated, one for 'smoking initiation' and one for 'smoking persistence', as well as one threshold each for the initiation and persistence dimensions. Parameters for 'smoking initiation' determine expected probabilities, say x_{ijk}, for the 2×2 tables cross-classifying twin pairs for lifetime smoking (never smoked versus always smoked) from the ith zygosity group. Parameters for 'smoking persistence' determine expected conditional probabilities, say y_{ijk}, for the 2×2 tables cross-classifying twin pairs who are concordant lifetime smokers with respect to persistence of smoking (successful quitter versus continuing smoker). Let x_{i00} denote the probability that a twin pair from the ith group are concordant never smokers, and x_{i11}, x_{i01} and x_{i10} are the corresponding probabilities that pairs are concordant ever smokers, or discordant with either first twin or second twin a nonsmoker; and let y_{i11} denote the conditional probability that a twin pair who are concordant smokers are successful quitters, y_{i22} denote the conditional probability that a twin pair are continuing smokers, and y_{i12}, y_{i21} the corresponding probabilities for pairs who are concordant for lifetime smoking but discordant for continued smoking. Expected probabilities for the 3×3 table are then easily written in terms of these unconditional and conditional probabilities (see Table 3). The only derivation requiring some thought is that for pairs discordant for lifetime smoking, the expected probability of observing a pair where one twin is a never smoker and the cotwin is a continuing smoker is simply the product of the unconditional probability of observing a pair discordant for lifetime smoking, and the marginal conditional probability of being a continuing smoker. In other words, given the assumption of independent liability dimensions, the expected proportion of continuing smokers among smokers whose cotwin has never smoked does not differ from the proportion of continuing smokers among all smokers.

Table 3 Expected probabilities for cells of the two-way contingency table for smoking status for a given zygosity group under independent liability dimension and combined models (adapted from Eaves and Eysenck[23] and Heath and Martin[14])

Independent liability dimensions

	Twin B		
Twin A	Never smoked	Successful quitter	Continuing smoker
Never smoked	x_{00}	$x_{01}\ (y_{11} + y_{21})$	$x_{01}\ (y_{12} + y_{22})$
Successful quitter	$x_{10}\ (y_{11} + y_{12})$	$x_{11} y_{11}$	$x_{11}\ y_{12}$
Continuing smoker	$x_{10}\ (y_{21} + y_{22})$	$x_{11}\ y_{21}$	$x_{11}\ y_{22}$

Combined model

	Twin B		
Twin A	Never smoked	Successful quitter	Continued smoker
Never smoked	$x_{00} + x_{11}y_{00} + x_{01}y_{\bullet 0} + x_{10}y_{0\bullet}$	$x_{01}y_{\bullet 1} + x_{11}y_{01}$	$x_{01}y_{\bullet 2} + x_{11}y_{02}$
Successful quitter	$x_{10}y_{1\bullet} + x_{11}y_{10}$	$x_{11}y_{11}$	$x_{11}y_{12}$
Continuing smoker	$x_{10}y_{2\bullet} + x_{11}y_{20}$	$x_{11}y_{21}$	$x_{11}y_{22}$

Note: $y_{0\bullet} = (y_{00} + y_{01} + y_{02})$; $y_{1\bullet} = (y_{10} + y_{11} + y_{12})$; $y_{2\bullet} = (y_{20} + y_{21} + y_{22})$.

As in the case where a single underlying liability dimension was assumed, we may compute the log-likelihood of the observed data, summed over zygosity groups, for given values of genetic and environmental parameters and thresholds for the initiation and persistence dimensions, and hence obtain maximum-likelihood estimates of these parameters by maximizing the log-likelihood with respect to these parameters using programs such as MX.[38] Eaves[23] also illustrated how this approach could be adapted to such problems as the analysis of genetic effects on quantity smoked (cigarettes per day). It does, however, rely upon the strong assumption that there are no genetic or environmental correlations between the smoking initiation and persistence liability dimensions, i.e. that they are uncorrelated. When we fitted this model to smoking initiation–persistence data from the Australian twin panel, it gave a very poor fit to the female like-sex data ($\chi^2 = 19.03$, df $= 4$, $p < 0.001$), while giving an acceptable fit to the male like-sex data ($\chi^2 = 7.17$, df $= 4$, $p = 0.13$).

Can we improve our ability to predict the observed data? One obvious approach would be to relax the assumption of orthogonal liability dimensions for smoking initiation and persistence. We might expect that it would be possible to relax this assumption by estimating genetic and shared- and nonshared-environmental correlations between the two dimensions, as in a standard bivariate problem in genetic analysis.[33,46] However, since we cannot assess smoking persistence in an individual who has never smoked, we can never estimate a within-family environmental correlation between smoking initiation and smoking persistence. One approach would be to test a submodel of the general bivariate genetic model[47] where, if our model for the mean liability score for the smoking initiation dimension is $I = A + C + E$, our corresponding model for the mean liability score for the smoking persistence dimension is $P = b\ I + A' + C' + E'$, where A, C, and E denote additive genetic, shared-environmental and nonshared-environmental effects on smoking initiation, b denotes the partial regression of persistence-liability on initiation-liability within individuals,

and A', C' and E' denote additive genetic, shared-environmental and nonshared-environmental effects that are specific to smoking persistence.[48] In other words, this model allows for the possibility that liability to smoking initiation does have an effect on probability of continued smoking, but that there are other factors, genetic or environmental, which come into play when an individual becomes a regular smoker that also influence this outcome. Introducing a single parameter still leaves a single degree of freedom with which to test our assumptions about the causes of variation in liability to smoking initiation and persistence (the remaining degrees of freedom merely test equality of marginal probabilities across zygosity groups). For given values of genetic and environmental parameters of this bivariate 'mediational' model (so-called because some of the genetic and environmental influences on smoking persistence are hypothesized to be mediated via effects on smoking initiation), and given threshold values for the initiation and persistence dimensions, we can derive (by integrating the quadrivariate normal distribution with expected covariance matrix derived from the values of genetic and environmental parameters) expected cell frequencies for a four-way $2 \times 2 \times 2 \times 2$ contingency table, i.e. cross-classifying 'initiation' and 'persistence' binary traits in first and second twins, for each zygosity group. Using the assumption that an individual who is negative for 'initiation' will always be a never smoker, we can then derive expected probabilities for the 3×3 contingency tables as before. Hence, maximum-likelihood estimates of model parameters can be obtained by maximizing the log-likelihood with respect to the model parameters in the usual manner.

An alternative approach to relaxing the assumptions of the independent-liability dimensions model, which has proved informative in practical applications[14,15] is represented, in the form of a probability tree (which illustrates only marginal, i.e. within-person probabilities), in Figure 1. In this 'combined' model, we retain the assumption of orthogonal liability dimensions but introduce a new parameter y_0 to allow for the possibility that individuals who became smokers (i.e. who were above the threshold on the smoking initiation dimension) but who almost immediately quit the habit (i.e. who had very low values on the smoking persistence dimension) will classify themselves as 'never smokers' when responding to general health surveys. The conditional probability that a twin pair who are concordant lifetime smokers will report themselves as concordant never smokers under the model is thus y_{00}, y_{01} is the probability that the first twin will report herself as a lifetime smoker and the cotwin as an ex-smoker, and so on. Expected cell frequencies, in terms of predicted probabilities x_{ijk} and y_{ijk}, are summarized in Table 3. Compared to the independent-liability dimensions model, this 'combined' model introduces one new parameter, the additional threshold value from which y_0 is estimated. In the case where $y_0 = 0$, it reduces to the independent-liability dimensions model; in the case where $x_0 = 0$, it reduces to the single-liability dimension case. These two alternate models are thus nested within it, allowing likelihood-ratio chi-square comparisons to the more general model. In the case of the Australian male like-sex pairs, the fit of a combined model, allowing for additive genetic and shared environmental effects on both initiation and persistence dimensions, represented only a slight nonsignificant improvement compared to the independent-liability dimensions model (goodness-of-fit: $\chi^2 = 5.47$, df = 3, $p = 0.14$; likelihood-ratio chi-square versus independent-liability dimensions model:

$\chi^2 = 1.70$, df = 1, $p = 0.19$). In the case of the female like-sex pairs, however, the combined model gave an excellent fit to the data ($\chi^2 = 2.12$, df = 3, $p = 0.99$), whereas the independent-liability dimensions model had been rejected ($p < 0.001$). Since the combined model, and the partial regression model described in the previous paragraph are not nested, they cannot be directly compared by likelihood-ratio chi-square. They can, however, be compared using Akaike's information criterion (AIC),[49] estimated as (χ^2-2df), with the model with the lowest AIC being preferred as the most parsimonious. In this case, since the two models have the same number of degrees of freedom, this reduces to selecting the model with the lowest chi-square. However, we shall not attempt this comparison here.

Table 4 summarizes, by gender, estimates under the combined model of additive genetic, shared- and nonshared-environmental variance components for smoking-initiation and smoking-persistence dimensions from the Australian twin study. Also shown are corresponding estimates for the Finnish twin panel and Vietnam-era twin panel. Results in women are quite disparate, with high heritability of smoking initiation, but effectively zero heritability of the smoking-persistence dimension, observed in Australian women, but with moderate to high heritabilities for both initiation and persistence in the Finnish women. Results in males are more strikingly consistent, with moderate heritability for smoking initiation (31–40%) and high heritability of smoking persistence (50–71%). In contrast to the Australian data, estimates from the Finnish sample are strikingly consistent in men and women.

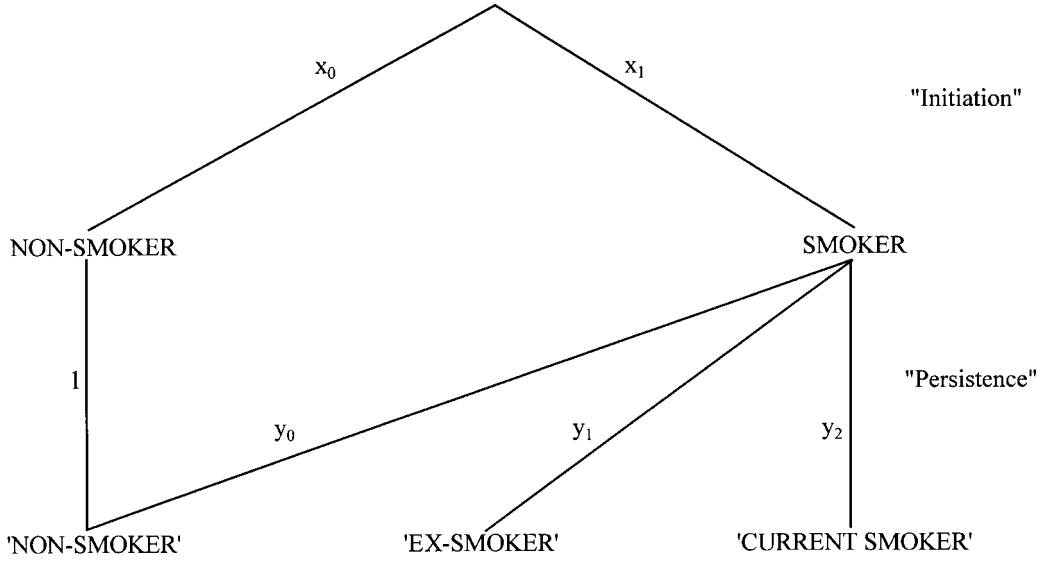

Figure 1 Probability tree representation of a combined model for smoking initiation and smoking persistence

Table 4 Genetic and environmental variance components for smoking initiation and continuation, and 95% confidence intervals, estimated under a combined model (see Figure 1 and text for details of model)

	Women		Men		
	Finnish	Australian	Finnish	Australian	US veterans
Initiation					
Additive genetic variance (%)	32 (21–42)	70 (46–92)	31 (19–43)	40 (4–76)	39 (23–56)
Shared environmental variance (%)	59 (50–69)	18 (0–41)	58 (47–69)	51 (15–85)	49 (32–64)
Nonshared environmental variance (%)	9 (6–12)	12 (0–17)	11 (8–15)	9 (3–17)	12 (9–16)
Continuation					
Additive genetic variance (%)	49 (16–80)	4 (0–58)	50 (27–71)	71 (31–84)	68 (45–74)
Shared environmental variance (%)	23 (0–47)	57 (7–72)	18 (1–35)	0 (0–36)	1 (0–21)
Nonshared environmental variance (%)	28 (18–42)	39 (26–53)	33 (25–42)	29 (16–45)	31 (26–38)

4 Logistic regression analysis

An alternative approach with its roots in epidemiology is to use multiple logistic regression analysis to predict the respondents' smoking status as a function of his or her cotwin's status and the twin pair zygosity (i.e. whether they are monozygotic twins who are genetically identical, or fraternal twins who on average share 50% of their genes in common and are no more alike than ordinary full siblings).[50,51] Here a double-entry procedure has been used, with each individual entered into the data set once as a respondent (i.e. entering into the left-hand side of the regression equation), and once as the cotwin of a respondent, with each observation assigned a sampling weight of 0.5 to correct for this double entry. Two sets of binary dummy variables, for MZ and for DZ pairs, have been used to code the smoking status of the respondent's cotwin (Table 5).

Table 5 summarizes results for initiation of smoking, i.e. whether or not the respondent reports having been a smoker. Odds ratios and 95% confidence limits are reported. MZ twins whose cotwin has never smoked have been used as a comparison group, so that effects of five dummy variables have been estimated in each multiple logistic regression analysis. For each sample, in both women and men, and assuming no overall differences in the prevalence of smoking behaviour in MZ versus DZ twins, results are in every case consistent with a significant genetic influence (or alternatively, environmental influences that are shared more often by MZ than by DZ pairs) on smoking initiation: DZ cotwins of nonsmokers have significantly higher rates of smoking than do the MZ cotwins of nonsmokers (i.e. exhibit significantly greater discordance, shown by odds ratios significantly greater than unity); and MZ cotwins of smokers have significantly higher rates of smoking (higher odds ratios) than do DZ cotwins of smokers. (If genetic factors are important, MZ cotwins of nonsmokers should be at the lowest risk, and MZ cotwins of smokers should be at the highest risk.) In addition, odds ratios are higher for cotwins of continuing smokers than for cotwins of successful quitters. This latter finding is consistent with the interpretation that some of the same genetic factors that influence initiation of smoking also influence persistence in the smoking habit by smokers (although it is also consistent with the 'combined' model interpretation that some transient smokers classify themselves as never smokers). In contrast, if genetic influences on risk of continuing in the smoking

Table 5 Associations between initiation of smoking and cotwin's smoking history

	Women				Men					
	Finnish		Australian		Finnish		Australian		US veterans	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
MZ cotwin continuing smoker	30.9	22.8–41.7	20.9	16.4–26.6	25.3	18.5–34.8	11.5	8.1–16.3	23.9	18.1–31.5
DZ cotwin continuing smoker	18.1	14.4–22.7	9.3	7.3–11.9	13.1	10.4–16.6	7.6	5.3–10.9	13.9	10.8–18.0
MZ cotwin successful quitter	22.4	15.6–32.1	11.9	9.1–15.5	13.9	9.9–19.5	7.2	5.2-10.1	12.0	9.3–15.6
DZ cotwin successful quitter	10.8	8.3–14.2	6.9	5.1–9.3	9.9	7.7–12.8	7.1	4.8–10.7	8.6	6.6–11.2
MZ cotwin never smoked	1.0	–	1.0	–	1.0	–	1.0	–	1.0	–
DZ cotwin never smoked	1.8	1.5–2.2	2.1	1.7–2.6	2.1	1.7–2.7	1.5	1.1–2.0	2.3	1.8–2.9

Table 6 Associations between continuing smoking[a] and cotwin's smoking history

	Women				Men					
	Finnish		Australian[b]		Finnish		Australian[b]		US veterans	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
MZ cotwin continuing smoker	5.7	3.7–8.6	3.4	2.4–4.7	5.5	4.0–7.6	5.3	3.4–8.3	5.9	4.6–7.5
DZ cotwin continuing smoker	5.0	3.4–7.3	3.1	2.1–4.6	3.7	2.8–4.9	3.5	2.2–5.7	3.7	2.9–4.7
MZ cotwin successful quitter	1.0	–	1.0	–	1.0	–	1.0	–	1.0	–
DZ cotwin successful quitter	2.3	1.5–3.5	1.0 (NS)	0.7–1.6	1.6	1.2–2.1	2.4	1.4–4.0	1.9	1.5–2.5
MZ cotwin never smoked	2.1	1.3–3.4	1.2 (NS)	0.8–1.8	1.6	1.0–2.4	1.7	1.0–2.7	1.4	1.0–1.9
DZ cotwin never smoked	2.4	1.6–3.5	1.6	1.1–2.3	2.0	1.5–2.8	2.5	1.5–4.2	1.7	1.3–2.4

[a]Excluding nonsmokers.
[b]Controlling for birth cohort.
NS = not significant.

habit were statistically independent of genetic influences on risk of initiation of smoking, odds ratios would be expected to be no higher for cotwins of successful quitters than for cotwins of continuing smokers.

Table 6 summarizes the results for the multiple logistic regression analyses predicting persistence in the smoking habit among smokers, with nonsmokers excluded from the analysis. Here, to facilitate interpretation, we have switched to using MZ cotwins of successful quitters as a comparison group. As before, the data from Australian women do not indicate a significant genetic influence on smoking persistence. For all other groups, however, the odds ratios for DZ cotwins of successful quitters are significantly greater than unity, consistent with a genetic influence; and in the Finnish and US veteran males, the odds ratios are significantly higher for MZ than for DZ cotwins of continuing smokers.

4.1 Likelihood-ratio tests for genetic effects

Inferences about the importance of genetic effects from results summarized in Tables 5 and 6 are complicated by two factors. First, these analyses do not take into account possible zygosity differences in the prevalence of smoking. Second, evidence for genetic effects is derived from multiple comparisons and, as in the case of the analyses of smoking persistence, not all of these may be significant. A number of alternative parameterizations of the logistic regression model can be used to provide a test for genetic effects based on a single degree of freedom. Considering first the case of smoking initiation, collapsing data from continuing smokers and successful quitters, we fitted a logistic regression model which included dummy variables for (1) twin pair zygosity status; (2) having an MZ cotwin who became a smoker; and (3) having a DZ cotwin who became a smoker, and compared this to a model which included only twin pair zygosity status, and an effect of having either an MZ or DZ cotwin who became a smoker (assumed to be the same regardless of zygosity). A similar analysis, limited to data from pairs that were concordant ever smokers, was conducted using smoking persistence as the outcome measure. In each case, a likelihood-ratio chi-square, estimated as twice the difference in log-likelihoods, was used to compare the full 3-parameter (plus intercept) and reduced 2-parameter logistic regression models. Results are shown in Table 7. Except in the case of smoking persistence in Finnish women, and smoking initiation in Australian men, the test for genetic influences was in every case significant.

Also shown in Table 7, are the likelihood-ratio chi-squares obtained by fitting genetic models to the 2×2 contingency tables for smoking initiation, and for smoking persistence (excluding pairs where either twin had never smoked), as described previously. In every case, the chi-square statistic to test the hypothesis of no genetic influence obtained by structural equation model-fitting and by logistic regression analysis were very similar. Indeed, if the prevalence estimates were identical in MZ and DZ pairs, the likelihood-ratio chi-square statistic would be identical regardless of whether logistic regression or model-fitting approaches were used. This should not be surprising, since in this latter case we are in either case testing the equality of proportions in two contingency tables.

In the case of linear regression using continuous outcome measures, Fulker[52] has pointed out that when zygosity is coded as a dummy variable, set to 1.0 for MZ pairs

Table 7 Likelihood-ratio tests of hypothesis of no genetic effects on smoking behaviour. (All tests are based on one degree of freedom, and are significant at the 0.001 level unless otherwise noted)

	Women		Men	
	Logistic regression	Model-fitting	Logistic regression	Model-fitting
Smoking initiation	χ^2	χ^2	χ^2	χ^2
Finland	60.75	59.02	70.91	71.99
Australia	45.92	45.34	2.56 (NS)	2.55 (NS)
USA	–	–	61.21	57.48
Continued smoking				
Finland	11.91	12.50	18.32	18.18
Australia	0.03 (NS)	0.03 (NS)	7.77**	7.83**
USA	–	–	38.49	38.59

NS = not significant.
**$p < 0.01$.

and 0.5 for DZ pairs, or more generally set equal to the coefficient of genetic relationship (see above),[32] fitting the regression model $P = b0 + b1\ \text{zyg} + b2\ \text{cotwin} + b3\ (\text{zyg} \times \text{cotwin})$ provides direct estimates of genetic and shared environmental parameters (regression coefficients $b2$ and $b3$, respectively) that are the same as those obtained by model-fitting methods. The rationale here is that the main effect of cotwin's score will detect shared environmental influences, while the interaction of cotwin's score with zygosity will detect genetic effects. A similar parameterization could be used in the case of logistic regression analysis. However, without considerable contortions,[50] there will be no direct correspondence between the estimated parameters of the logistic regression model, and the estimates of genetic and environmental parameters that would be recovered by model-fitting methods. Furthermore, there is no test for shared environmental effects in the logistic regression model that is equivalent to the likelihood-ratio chi-square test obtained by model-fitting. Thus, dropping a shared-environmental parameter from a model allowing for additive genetic and nonshared-environmental effects, in model-fitting analyses of the Finnish smoking initiation data, produces substantial changes in chi-square (137.43 in women, 47.36 in men); whereas dropping the main effect of cotwin's smoking status from a logistic regression model produces much more minor changes ($\chi^2 = 21.68$ in women, 4.20 in men)

4.2 Advantages of a regression approach

Use of a regression approach to the analysis of twin and other family data on smoking has several attractions. A regression approach is readily extended to test for possible mediators of genetic influences on smoking behaviours. While genetic model-fitting approaches can be used for the same purpose,[33,47] these are difficult to implement for binary variables, such as continued smoking, that are only assessed in a subset of individuals (i.e. those who become smokers), and especially so when some of the same genetic or family environmental factors that determine risk of continued smoking also determine whether or not an individual becomes a smoker. Many behavioural genetic studies have shown greater resemblance of MZ than DZ twin pairs

for personality traits,[53,54] educational attainments,[55,56] and a variety of other psychological and sociodemographic traits. Thus, it is important to progress beyond the question of whether there are genetic influences on smoking behaviour, to the question of how such genetic influences may arise. To the extent that certain variables are important mediators of genetic influence, including the respondent's scores on these variables in a multiple logistic regression equation should reduce the estimated residual association with cotwin's smoking status.

By way of illustration, Table 8 summarizes pertinent personality and sociodemographic correlates of smoking initiation and persistence in the 1981 survey of the Australian twin panel, controlling for birth cohort, estimated from a multiple logistic regression analysis and Table 9 summarizes the partial odds ratios for the association with cotwin's smoking status when these sociodemographic and personality variables, as well as birth cohort, are controlled for. In Table 8, the odds ratios for continuous personality measures are computed for a change in score equal to the inter-quartile range for each measure. In terms of personality, those who become smokers are more likely to be extroverted, neurotic, socially nonconforming, and (if women) toughminded, as assessed by the Eysenck Personality Questionnaire.[57] They are less likely to report a Protestant (e.g. Methodist) or (if women) Jewish religious affiliation, and more likely to report a religious affiliation of Roman Catholic (if men), or to report no religion (if women). They are more likely to be unmarried, or separated or divorced. For all of these measures, substantial twin pair resemblance has previously been reported.[53,56,58]

Table 9, however, shows that the familial transmission of personality and sociodemographic factors only partially explains twin pair concordance for the initiation and persistence of smoking. While odds ratios are certainly reduced, compared to those in Tables 5 and 6 that were not adjusted for personality and sociodemographic variables, the findings show significant evidence for genetic effects on smoking

Table 8 Sociodemographic and personality correlates of initiation and continuation of smoking in the Australian twin panel 1981 survey, estimated by multiple logistic regression

	Women				Men			
	Initiation		Continuation		Initiation		Continuation	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Never married	0.73	0.62–0.86	1.65	1.29–2.12	0.75	–	–	
Separated/divorced	1.85	1.43–2.40	–	–	1.52	1.18–1.95	–	–
0–10 years education	1.58	1.29–1.94	2.15	1.58–2.91	2.55	1.91–3.41	1.95	1.34–2.85
11–12 years education	1.59	1.32–1.91	1.48	1.12–1.95	1.78	1.43–2.22	1.96	1.44–2.96
Other Protestant religion	0.59	0.51–0.68	–	–	0.58	0.47–0.73	–	–
Roman Catholic religion	–	–	–	–	1.52	1.18–1.95	–	–
No religion	1.37	1.07–1.75	–	–	–	–	–	–
Jewish/other minority religion	0.57	0.41–0.81	–	–	–	–	–	–
Extraversion (E)	1.80	1.60–2.01	1.24	1.05–1.48	1.32	1.12–1.55	–	–
Neuroticism (N)	1.40	1.24–1.58	–	–	1.42	1.19–1.69	–	–
Social nonconformity (L)	1.43	1.27–1.60	–	–	1.68	1.41–1.91	–	–
Toughmindedness (P)	1.38	1.24–1.53	1.20	1.03–1.48	–	–	1.31	1.11–1.53

Note all analyses also control for birth cohort.

Table 9 Association between respondent's and cotwin's smoking status in the Australian twin panel 1981 survey, controlling for birth cohort and for respondent's sociodemographic and personality variables

	Women				Men			
	Initiation		Continuation		Initiation		Continuation	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
MZ cotwin continuing smoker	17.24	12.14–24.48	3.26	2.03–5.23	9.79	5.91–16.22	4.84	2.57–9.13
DZ cotwin continuing smoker	8.25	5.74–11.86	3.02	1.73–5.26	6.64	3.93–11.25	3.19	1.61–6.32
MZ cotwin successful quitter	10.25	6.96–15.10	1.00	–	6.93	4.28–11.21	1.00	–
DZ cotwin successful quitter	6.76	4.36–10.48	1.10	0.59–2.04	6.49	3.58–11.78	2.32	1.12–4.80
MZ cotwin never smoked	1.00	–	1.28	0.76–2.16	1.00	–	1.67	0.85–3.29
DZ cotwin never smoked	2.08	1.55–2.80	1.60	0.95–2.69	1.46	0.93–2.28	2.45	1.17–5.15

initiation in women, and on persistence of smoking in men. As before, we find significant evidence for familial influences on persistence of smoking in women, but there is not significant evidence for a genetic influence. However, there is also no longer significant evidence for genetic effects on smoking initiation in men, although there is still strong evidence for strong familial (but possibly shared environmental) influences. Constraining MZ male odds ratios to be the same as DZ male odds ratios, produces a clearly nonsignificant likelihood-ratio chi-square ($\chi^2 = 4.02$, df = 3, $p = 0.26$, without adjustment for the nonindependence of observations on twin pairs).

A regression approach also extends readily to a survival analysis framework,[59] using proportional hazards or accelerated failure-time models to take into account the fact that data on many current smokers are 'censored' in the sense that they have not been followed for long enough to have quit smoking. Use of a simple binary classification of current versus ex-smoker loses important information, for example, twins who are current smokers but have only smoked for a few years will surely include many more future successful quitters than twins who have smoked for several decades. It is indeed remarkable that such robust evidence for genetic influences on the persistence of smoking has emerged despite the crudeness of the summary measure of smoking status that has most often been used for analysis. While attempts have been made to combine the elements of genetic and accelerated failure-time models,[60] fitting such models has proved computationally intensive, and their application to multivariate problems a daunting prospect. Fitting survival models using dummy variables to represent the cotwin's smoking history circumvents these problems.

5 Discussion

We have given a brief and selective review of methods that have been applied in genetic analyses of smoking data. We have focused on twin data because adoption data on smoking persistence are rare, and because interpretation of intergenerational data is made more complicated by changes in the regulation and marketing of cigarettes (e.g. changing nicotine yields). Although a variety of different approaches have been advocated for the summary of twin pair concordance,[43,45] we have focused on two complimentary approaches – genetic model-fitting under a normal-liability threshold model, and logistic regression analysis – which yield quantitatively similar likelihood-ratio chi-square tests for the significance of genetic effects on smoking. Our illustrative analyses confirm previous reports of a significant and substantial genetic influence on smoking persistence,[14,15] at least in men, and in the case of the Australian data, extend these by showing that genetic effects cannot be accounted for by the inheritance of sociodemographic or personality variables. As the evidence for a strong genetic involvement in the course of cigarette smoking continues to grow, statistical approaches focused on gene-mapping[21,61,62] are likely to assume increased importance in smoking research as in most other areas of genetic research.

Acknowledgements

Supported, in varying percentages, by NIH grants AA07535, AA07728, AA09022 and AA10249 (to ACH), DA00272 (to PAFM), CA75581 and grants from the Australian NH & MRC (to NGM). We are grateful to Mike Neale for his advice about the development of an appropriate MX script for fitting the combined model.

References

1 Fisher RA. Cancer and smoking. *Nature* 1958; **182**: 596.
2 Cederlof R, Epstein FH, Friberg LT, Hrubec Z, Radford EP. Twin registries in the study of chronic disease (with particular reference to the relation of smoking to cardiovascular and pulmonary diseases). *Acta Medica Scandinavica* 1971; **523**: 1–40.
3 Hrubec Z, Neel JV. The National Academy of Sciences – National research council twin registry: ten years of operation. In: Nance WE ed. *Twin research, part B, biology and epidemiology*. New York: Alan R Liss, 1978: 154–72.
4 Kaprio J, Sarna S, Koskenvuo M, Rantasaio I. The Finnish twin registry: baseline characteristics. Section II – History of symptoms and illnesses, use of drugs, physical characteristics, smoking, alcohol and physical activity. *Kansanterveystieteen julkaisuja M* 1978; **37**: 71–96.
5 McClearn GE, Rodgers DA. Differences in alcohol preference among inbred strains of mice. *Quarterly Journal of Studies on Alcohol* 1959; **20**: 691–95.
6 Eriksson K. Genetic selection for voluntary alcohol consumption in the albino rat. *Science* 1968; **159**: 739–41.
7 Li TK, Lumeng L, Doolittle DP. Selective breeding for alcohol preference and associated responses. *Behavior Genetics* 1993; **23**: 163–70.
8 Heath AC, Slutske WS, Madden PAF. Gender differences in the genetic contribution to alcoholism risk and to alcohol consumption patterns. In: Wilsnack RW, Wilsnack SC eds. *Gender and alcohol: individual and social perspectives*. Rutgers, NJ: Rutgers University Press, 1997; **5**: 114–49.
9 Schuckit MA, Smith T. An 8-year follow-up of 450 sons of alcoholic and control subjects. *Archives of General Psychiatry* 1996; **53**: 202–10.
10 Higuchi S, Matsushita S, Imazeki H, Kinoshita T, Takagi S, Kono H. Aldehyde dehydrogenase genotypes in Japanese alcoholics. *Lancet* 1994; **343**: 741–42.
11 Muramatsu T, Zu-Cheng W, Yi-Ru F, Kou-Bao H, Heqin Y, Yamada K *et al*. Alcohol and aldehyde dehydrogenase genotypes and drinking behaviour of Chinese living in Shanghai. *Human Genetics* 1995; **96**: 151–54.
12 Reich T, Edenberg H, Goate A, Williams JT, Rice J, van Eerdewegh P *et al*. A genome-wide search for genes affecting the risk for alcohol dependence. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* 1998 (in press).
13 Litten RZ, Allen J, Fertig J. Pharmacotherapies for alcohol problems: a review of research with focus on developments since 1991. *Alcoholism: clinical and experimental research* 1996; **20**: 859–76.
14 Heath AC, Martin NG. Genetic models for the natural history of smoking: evidence for a genetic influence on smoking persistence. *Addictive Behaviors* 1993; **18**: 19–34.
15 True WR, Heath AC, Scherrer JF, Goldberg J, Lin N, Eisen SA *et al*. Genetic and environmental contributions to cigarette smoking. *Addiction* 1997; **92**: 1277–87.
16 Fisher RA. The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* 1934; **6**: 13–25.
17 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986; **1**: 54–77.
18 Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* 1990; **46**: 222–28.
19 Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics* 1990; **46**: 229–41.
20 Risch N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics* 1990; **46**: 219–21.

21 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–17.
22 Eaves LJ, Last K, Young PA, Martin NG. Model-fitting approaches to the analysis of human behavior. *Heredity* 1978; **41**: 249–320.
23 Eaves LJ, Eysenck HJ. The genetics of smoking. In: Eysenck HJ eds. *The causes and effects of smoking.* London: Maurice Temple Smith, 1980: 140–314.
24 Pearson K. Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A* 1900; **195**: 1–47.
25 Falconer DS. The inheritance of liability to certain diseases estimated from the incidence among relatives. *Annals of Human Genetics* 1965; **29**: 51–76.
26 Falconer DS. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of Human Genetics* 1967; **31**: 1–20.
27 Kendler KS, Kidd KK. Recurrence risks in an oligogenic threshold model: the effect of alterations in allele frequency. *Annals of Human Genetics* 1986; **50**: 83–91.
28 Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 1918; **52**: 399–433.
29 Falconer DS. *Introduction to quantitative genetics.* Edinburgh: Oliver and Boyd, 1960.
30 Jinks JL, Fulker DW. A comparison of the biometrical genetical, MAVA and classical approaches to the analysis of human behavior. *Psychological Bulletin* 1970; **73**: 311–49.
31 Eaves LJ. A model for sibling effects in Man. *Heredity* 1976; **36**: 205–14.
32 Bulmer MG. *The mathematical theory of quantitative genetics.* Oxford: Clarendon Press, 1980.
33 Neale MC, Cardon LR. *Methodology for genetic studies of twins and families, NATO ASI Series.* Dordrecht: Kluwer Academic, 1992.
34 Rao DC, Morton NE, Yee S. Resolution of cultural and biological inheritance by path analysis. *American Journal of Human Genetics* 1976; **28**: 228–42.
35 Cloninger CR, Reich T. Multifactorial inheritance with cultural transmission and assortative mating I. Description and basic properties of the unitary models. *American Journal of Human Genetics* 1978; **30**: 618–43.
36 Cloninger CR, Rice J, Reich T. Multifactorial inheritance with cultural transmission and assortative mating. II. A general model of combined polygenic and cultural inheritance. *American Journal of Human Genetics* 1979; **31**: 176–98.
37 Joreskog KG; Sorbom D. *PRELIS 2 user's reference guide.* Chicago, IL: Scientific Software International, 1993.
38 Neale MC. *Mx: statistical modeling.* Richmond, VA: Virginia Commonwealth University, 1997.
39 Tallis GM. The maximum-likelihood estimation of correlation from contingency tables. *Biometrics* 1962; **18**: 342–53.
40 Olsson U. Maximum-likelihood estimation of the polychoric coefficient. *Psychometrika* 1979; **44**: 443–60.
41 Heath AC, Madden PAF. Genetic influences on smoking behavior. In: Turner JR, Cardon LR, Hewitt JK eds. *Behavior genetic applications in behavioral medicine research.* New York: Plenum, 1995.
42 Neale MC, Miller MB. The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics* 1997; **27**: 113–20.
43 Kraemer HC. What is the 'right' statistical measure of twin concordance (or diagnostic reliability and validity)? *Archives of General Psychiatry* 1997; **54**: 1121–5.
44 Heath AC, Jardine R, Martin NG. Interactive effects of genotype and social environment on alcohol consumption in female twins. *Journal of Studies on Alcohol* **50**: 38–48.
45 Hannah MC, Hopper JL, Mathews JD. Twin concordance for a binary trait. II. Nested analysis of ever-smoking and ex-smoking traits and unnested analysis of a 'committed-smoking' trait. *American Journal of Human Genetics* 1985; **37**: 153–65.
46 Eaves LJ, Gale JS. A method for analysing the genetic basis of covariation. *Behavior Genetics* 1974; **4**: 253–67.
47 Heath AC, Kessler RC, Neale MC, Hewitt JK, Eaves LJ, Kendler KS. Testing hypotheses about direction of causation using cross-sectional family data. *Behavior Genetics* 1993; **23**: 29–50.
48 Heath AC, Todorov AA, Madden PAF, Bucholz KK, Dinwiddie SH. Modelling the role of genetic factors in the natural history of substance use disorders. Paper presented at the World Congress on Psychiatric Genetics, New Orleans, Louisiana, October 3–5, 1992.
49 Sham PC, Walters EE, Neale MC, Heath AC, MacLean CJ, Kendler KS. Logistic regression analysis of twin data. *Behavior Genetics* 1994; **24**: 229–38.

50 Heath AC, Bucholz KK, Madden PAF, Dinwiddie SH, Slutske WS, Statham DJ *et al*. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in men and women. *Psychological Medicine* 1997; **27**: 1381–96.
51 DeFries JC, Fulker DW. Multiple regression analysis of twin data. *Behavior Genetics* 1985; **15**: 467–73.
52 Eaves LJ, Eysenck HJ, Martin NG. *Genes, culture, and personality: an empirical approach*. London: Academic Press, 1989.
53 Loehlin JC. *Genes and environment in personality development: individual differences and development series, volume 2*. Newbury Park, CA: Sage Publications, 1992.
54 Vogler GP, Fulker DW. Familial resemblance for educational attainment. *Behavior Genetics* 1983; **13**: 341–54.
55 Heath AC, Berg K, Eaves LJ, Solaas MH, Corey LA, Sundet HM *et al*. Education policy and the heritability of educational attainment. *Nature* 1985; **314**: 734–36.
56 Baker LA, Treloar SA, Heath AC, Martin NG. Genetics of educational attainment in Australian twins: sex differences and secular changes. *Behavior Genetics* 1996; **26**: 89–102.
57 Eysenck HJ, Eysenck SBG. *Manual of the Eysenck personality questionnaire*. London: Hodder & Stoughton, 1975.
58 Eaves LJ, Martin NG, Heath AC. Religious affiliation in twins and their parents: testing a model of cultural inheritance. *Behavior Genetics* 1990; **20**: 1–22.
59 Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. New York: John Wiley & Sons, 1980.
60 Meyer JM, Eaves LJ, Heath AC, Martin NG. Estimating genetic influences on the age-at-menarche: a survival analysis approach. *American Journal of Medical Genetics* 1991; **39**: 148–54.
61 Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994; **265**: 2037–48.
62 Risch N, Zhang H. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* 1995; **268**: 1584–89.