Analysis of Melanoma Onset: Assessing Familial Aggregation by Using Estimating Equations and Fitting Variance Components via Bayesian Random Effects Models

Kim-Anh Do[1], Joanne F. Aitken[2], Adèle C. Green[3], and Nicholas G. Martin[3]

[1] *Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA*
[2] *Queensland Cancer Fund Epidemiology Unit, Brisbane, Australia*
[3] *Epidemiology and Population Health Unit, Queensland Institute of Medical Research, Brisbane, Australia*

We investigate whether relative contributions of genetic and shared environmental factors are associated with an increased risk in melanoma. Data from the Queensland Familial Melanoma Project comprising 15,907 subjects arising from 1912 families were analyzed to estimate the additive genetic, common and unique environmental contributions to variation in the age at onset of melanoma. Two complementary approaches for analyzing correlated time-to-onset family data were considered: the generalized estimating equations (GEE) method in which one can estimate relationship-specific dependence simultaneously with regression coefficients that describe the average population response to changing covariates; and a subject-specific Bayesian mixed model in which heterogeneity in regression parameters is explicitly modeled and the different components of variation may be estimated directly. The proportional hazards and Weibull models were utilized, as both produce natural frameworks for estimating relative risks while adjusting for simultaneous effects of other covariates. A simple Markov Chain Monte Carlo method for covariate imputation of missing data was used and the actual implementation of the Bayesian model was based on Gibbs sampling using the free ware package BUGS. In addition, we also used a Bayesian model to investigate the relative contribution of genetic and environmental effects on the expression of naevi and freckles, which are known risk factors for melanoma.

Melanoma is a complex chronic disease, the incidence of which has more than doubled over the past 20 years (MacLennan et al., 1992). The disease's complexity can be, in part, explained in terms of the joint effects of genotype and environment, and censorship due to the late onset of the disease in most patients.

Studies have revealed a number of possible underlying factors that may contribute to the risk of melanoma (English et al, 1997; Siskind et al, 2002; Swerdlow & Green, 1987). Such factors have included sun exposure, skin, hair and eye colour, degree of freckling, number of naevi, place of birth, and ethnic origin. It is also thought that certain genes are responsible for a person's susceptibility to the disease (Aitken et al., 1998) and that a number of the above mentioned risk factors may themselves be genetically influenced.

A preliminary segregation analysis conducted by Aitken et al. (1998) investigated whether the familial clustering of cutaneous melanoma was consistent with Mendelian inheritance of a major autosomal gene. Analyses were performed with the SAGE statistical package using the maximum likelihood REGTL program for a binary trait. The hypothesis of co-dominant Mendelian inheritance gave a significantly better fit to the data than either dominant or recessive Mendelian inheritance. Overall, both Mendelian inheritance of a single major gene and purely environmental transmission were rejected. However, there was strong evidence of familial dependence in melanoma occurrence.

Despite these findings, the etiology of melanoma is still not well understood. For example, it is unclear why some people are more susceptible to the disease than others and for susceptible cases, it is unclear whether certain risk factors play a role in the progression of the disease or if genetic factors are a major source of influence.

Address for correspondence: Dr. K-A Do, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, Texas 77030, USA. Email: kim@mdanderson.org

Research into chronic diseases such as melanoma has largely been achieved through family studies, where the age of onset of the disease is modelled within a survival analysis framework. An example of this type of analysis is the work by Abel and Bonney (1990), who developed a model accounting for age of onset, where the hazard function was expressed in terms of a major gene effect and residual family dependence using the regressive approach described in Bonney (1986).

Clayton (1978) and Vaupel (1979) introduced the so-called "frailty" model which has been extended by many others (Andersen et al., 1993; Clayton & Cuzick, 1985; Gauderman & Thomas, 1994; Hougaard, 1986; Li & Thompson, 1997; Li & Wijsman, 1998; Nielsen et al., 1992; Siegmund & McKnight, 1998). A variant of these approaches uses estimating equations, which accommodates correlated age of onset outcomes and is known to be robust and computationally efficient (Hsu & Prentice, 1996; Hsu & Zhao, 1996). A common theme is the investigation of possible underlying genetic and environmental factors that may influence the age of onset of the disease of interest.

Recently, the focus of familial studies involving censored data has moved towards the development of Bayesian methods using packages such as BUGS and WinBUGS (Gilks et al., 1994; Spiegelhalter et al., 1996a, 1996b, 2003), the Genetic Analysis Package (GAP, 1996) and MIXD (Olshen & Wijsman, 1996; Thompson, 1994). Examples of this include the implementation of Markov Chain Monte Carlo (MCMC) methods for linkage analysis (Heath, 1997; Kong et al., 1992; Lange & Sobel, 1991); the estimation of parameters in a mixed model, with and without covariates (Guo & Thompson, 1991; Thomas, 1992); estimation of a gene-smoking interaction and covariate imputation (Gauderman et al., 1997); combined linkage and segregation analysis (Guo & Thompson, 1992; Faucett et al., 1993); and development of mixed models for large complex pedigrees (Guo & Thompson, 1994). Recent work by Do et al. (2000) demonstrates an application of Bayesian methodology to menopausal age in twins using a generalised linear mixed model (GLMM). In this application, they investigated the contribution of covariates and any underlying genetic and environmental factors to explain variation in menopausal age. A similar investigation by Scurrah et al. (2000) fitted a Bayesian model to survival data. In this application, the authors used a Bayesian piece-wise exponential model to explore the time to onset of respiratory disease, given known risk factors and possible familial effects. Both papers highlight the flexible nature of the Bayesian approach, which can be seen through the inclusion of priors and the integration of fixed and random effects.

The Bayesian approach to fitting a genetic model in a GLMM framework was examined in detail in Kuhnert and Do (2003). In this paper, a simulation study investigated the flexible nature of the Bayesian model and its ability to incorporate genetic components through random effects. This was compared with standard maximum likelihood methods for estimating genetic components in the model. Results from a simulation study indicated a consistent advantage in using the Bayesian method to detect a correct model under particular scenarios of additive genetics and common environmental effects. Moreover, for binary data, there was difficulty in detecting the correct model under low and moderate levels of heritability. Results, however, were improved for ordinal data under similar scenarios.

We present an alternative model using Bayesian methodology, which takes into account the complex features inherent with melanoma, using a large dataset comprising 1449 families. The model does not assume proportional hazards, but a multiplicative model, where the Weibull distribution is used to model the age at onset of melanoma. This is fit in a Bayesian framework, which incorporates fixed and random effects to estimate possible risk factors, covariates and any underlying genetic factors. Another approach to accommodate the correlated age-at-onset outcomes rigorously is to use the estimating equations for assessing familial aggregation of age-at-onset (Hsu & Prentice, 1996; Hsu & Zhao, 1996). It has two desirable features: (i) robustness — no higher-order distributional assumptions are required beyond pairwise ones; and (ii) computational efficiency.

In this article, our interest lies primarily with investigating whether certain risk factors are able to explain a considerable proportion of the familial dependence and if including them into the model reduces the residual variation and results in increased power for detecting a major gene effect. To our knowledge, this study will be the first in melanoma research that extends current methodology and incorporates covariate and genetic effects simultaneously, with age of onset, using Bayesian methodology on family data.

Material and Methods

The Data on the Age at Onset of Melanoma and Potential Risk Factors

We analysed data from the Queensland Familial Melanoma Project. Family ascertainment and data collection have been described in detail (Aitken et al., 1996). Assessing standard melanoma risk factors include counts of naevi on the arms and back, demographic and medical details, lifetime residence and family history of melanoma and other cancers. Briefly, we ascertained all 12,016 first incident cases of cutaneous melanoma (invasive and in situ) diagnosed in Queensland residents between 1982 and 1990 and

reported to the Queensland Cancer Registry, or found by comparing cancer registrations for 1984 and 1987 with records of pathology laboratories throughout Queensland. It is estimated that registry records are approximately 95% complete for the study period. Doctors' permission was obtained to approach 10,407 cases of whom 7784 (75%) returned a brief family history questionnaire, stating whether any of their first-degree relatives (parents, siblings, children) had had a diagnosis of melanoma. A total of 2920 probands was sampled from these respondents, including all who had claimed a positive family history (n = 1529) and an approximate 20% random sample of the remainder (n = 1391). Probands were sent a detailed family history questionnaire, asking for the names and addresses of all first-degree relatives, relatives' vital status, dates of birth, and ages, and whether any of these relatives had had a melanoma diagnosed by a doctor. To avoid bias in determining the mode of inheritance, second and lower degree relatives were enrolled in the study according to a sequential sampling scheme (Cannings & Thompson, 1977). First degree relatives of all relatives with confirmed melanoma were ascertained through the detailed family history questionnaire, described above, which was mailed to all confirmed positive relatives. In total, 15,989 relatives belonging to 1912 separate families were reported by 2118 (73%) probands or other positive relatives. A total of 1044 relatives for whom date of birth was unknown were excluded, leaving 14,945 relatives for analysis. There were 188 families independently ascertained through two or more probands. To avoid ascertainment bias, these families were included in the dataset separately for each proband in the family.

Medical confirmation and dates of diagnosis were sought for the relatives reported by probands or other relatives to have had melanoma. After eliminating 18.7% of subjects who refused access to their medical records, or those with lost records, or those with false positive reports (basal or squamous cell carcinoma, solar keratoses, or benign naevi), medical confirmation of melanoma as the diagnosis was obtained for 48.2% of the original number of relatives. Only the medically verified cases among relatives were classified as true events; all other relatives were treated as unaffected (censored at last date of contact).

Risk factor questionnaires were subsequently mailed to all living relatives aged between 18 and 75 years ascertained through the sequential sampling procedure. Other relatives provided proxy reports. The combined number of proxy-reports and self-reports was 9746 relatives for whom standard risk factor information was available.

For the Bayesian analysis, we focused on families that included at least one parent and at least one child, where each member in the family should have information on age at diagnosis or age at last follow-up and with maximum one missing covariate.

The demographic covariates and hypothesized melanoma risk factors included gender, birth year, place of birth, ability to tan (very brown, moderate tan, slight tan, no tan), propensity to burn (never burn and always tan, sometimes burn and usually tan, usually burn and sometimes tan, always burn and never tan), number of sunburns (0, 1, 2–5, ≥ 6), skin color (olive/dark, medium, fair/pale), hair color (black, light/dark brown, fair/blonde, light/dark red), eye color (brown, green/hazel, blue/grey), total freckling in summer (0, 1–100, > 100), number of naevi (none, few, moderate number, very many), and numerous measures of cumulative lifetime exposures to sun and ultraviolet rays.

Preliminary Exploratory Analysis

As a preliminary analysis, we ignored correlations within families and applied a combination of parametric and non-parametric survival analysis techniques as exploratory tools to identify possible risk factors for melanoma. Once these fixed effects were identified, we considered incorporating these into a subsequent generalized estimating equation model or a Bayesian model with random effects that could account for within-family correlations. The aim was to quantify the genetic and familial associations in the presence of observed covariate effects.

Manipulation of the entire melanoma dataset resulted in a subset of 9669 observations with a range of explanatory variables that described phenotypic characteristics for each individual, along with some demographic details such as birth year and gender. The response variable was the time to diagnosis (or age at the last follow-up), with the proportion of censored cases being approximately 76%. The median age at onset of melanoma was 43. The correlation estimates for age of onset for different relationship pairs with both affected members were: 0.67, 0.55, and 0.39 for sib–sib, parent–child, and second/lower order pairs respectively.

The first stage of modelling involved fitting univariable proportional hazard models to assess each variable's individual effect on the time of onset of melanoma. The SAS (Allison, 1995) package was used to fit proportional hazards models of the following form

$$h(t, x) = \Psi\,(x; \beta)h_0(t)$$

where Ψ represents a log-linear function $e^{\beta' x}$ of the explanatory variables x and corresponding coefficients β and $h_0(t)$ represents the baseline hazard at time t.

Significant variables associated with the age at onset of melanoma consisted of eye, hair and skin colour; freckling; number of moles; skin type; ability to burn; ability to tan; previous skin cancers; ultraviolet exposure between the ages of 5 and 12 years;

cumulative sun exposure up to the age of 19 years; and birth year.

The second stage fitted multivariable proportional hazard models to those explanatory variables that were significant at the univariable stage. Table 1 displays the results from the final model, which only included significant variables (p-value < .05). It is worth noting that a similar result could be obtained using an automated stepwise procedure.

Results from this analysis highlight some interesting but quite obvious risk factors noted in previous analyses (Aitken et al., 1996, 1998). For example:

- An increase of 1 year in birth year induces 17% increase in risk of earlier melanoma onset.
- People with neither freckles nor naevi have the lowest risk of melanoma onset.
- The risk of earlier melanoma onset is increased by up to 37% for blue eyed people and even further (46%) for green eyed people, when compared to individuals with brown eyes.
- "Red Heads" have an increased risk of earlier melanoma onset (46%) when compared to individuals with black hair. However, no significant increase was noted for individuals with fair or light red hair.
- A person's tendency to burn easily increases the risk of earlier melanoma onset, in some cases by up to 100% compared to those that never burn.

However striking this last statement is, issues of confounding, must also be considered. The most obvious illustration of this is the confounding that occurred between mole count and freckling. This is seen through close inspection of the parameter estimates which changed in magnitude when mole count was added to the model after adjusting for freckling. (Results not displayed here.)

To reduce the dimension of the problem further and avoid some of these confounding issues, a survival tree was constructed using RPART (Recursive Partitioning and Regression Trees), see Therneau and Atkinson (1997). Survival trees are a special case of

Table 1

Results from Fitting a Multivariable Proportional Hazards Model to the Melanoma Data Based on Univariable Results. The Results Reported in this Table Are the Parameter Estimates β, Their Standard Errors se(β), the Relative Risk e^{β} and the p-value for Each Estimate

Variable	β	se(β)	e^{β}	p-value
Birth Year	0.16	< 0.01	1.17	< .05
Eye Colour (Baseline: Brown)				
Blue/Grey	0.31	0.07	1.36	< .05
Green/Hazel	0.38	0.07	1.46	<.05
Hair Colour (Baseline: Black)				
Light Red/Ginger	0.17	0.15	1.19	.27
Dark Red/Auburn	0.38	0.15	1.46	< .05
Fair/Blonde	0.06	0.12	1.06	.62
Light Brown	0.14	0.12	1.15	.22
Dark Brown	0.02	0.12	1.02	.87
Skin Type (Baseline: never burn)				
Always burn	0.69	0.16	1.968	< .05
Usually burn	0.45	0.15	1.57	< .05
Sometimes burn	0.31	0.15	1.36	< .05
Freckling (Baseline: none)				
1 to 100	0.17	0.06	1.18	< .05
> 100	0.09	0.08	1.10	.23
Mole Count (Baseline: none)				
Few	0.29	0.07	1.34	< .05
Moderate	0.79	0.08	2.20	< .05
Many	1.12	0.10	3.08	< .05
Number of Sunburns (Baseline: none)				
One	–0.07	0.11	0.93	.49
2 to 5	–0.06	0.09	0.94	.50
> 6	0.17	0.09	1.19	.07
Cumulative Sun Exposure (< 5 yrs)	0.04	0.01	1.04	< .05
UV Exposure (5–12 yrs)	0.0003	< 0.0001	1	< .05

decision trees that were incorporated into RPART by Therneau (1997) using the ideas put forth by LeBlanc and Crowley (1992) for survival data with censoring. These authors showed through simulation studies that survival trees could outperform standard parametric methods such as proportional hazards modelling, particularly in situations where the underlying distribution was not exponential. Along with Breiman et al. (1984), they showed decision trees to be useful exploratory tools for identifying important variables, interactions and outliers.

The methodology for survival trees begins by identifying optimal splits, using the log rank statistic to separate the data into homogeneous groups. Each split is comprised of a parent node and two daughter nodes which are linked to the parent by branches. Figure 1 is an example of a survival tree produced on the melanoma data. Once a large tree is grown with many terminal nodes that contain very few observations, a pruning procedure is introduced to identify a sequence of sub-trees. The technique of pruning involves snipping back splits of the tree, one at a time until only the root node remains. Cross-validation is then introduced to aid in the selection of the optimal model. One nice feature of survival trees is the use of Kaplan-Meier curves to provide information about survival rates at each terminal node of the tree. This approach may allow for better interpretation of the terminal nodes and highlight different scenarios which yield similar survival rates as can be seen in Figure 1.

The variables identified from the multivariable model were used as input into RPART. Variables such as the sun and UV exposures were omitted from the modelling stage, since nearly half of the data for each of these variables were missing (cumulative sun exposure: 43.8%; UV exposure: 40.4%). A large survival tree was produced in SPLUS and using cross-validation, a model splitting solely on birth year was identified. This survival tree yielded the minimum cross-validated error rate (0.93) suggesting that birth year was an important indicator for melanoma. Selecting a slightly larger but more informative model (error = 0.94) resulted in a survival tree consisting of nine terminal nodes. This revealed splits on birth year, mole count and freckling. All other variables appeared to be either competing at each node or acting as a surrogate variable for one of these primary splits. Skin type, in particular, arose as an important surrogate for freckling.

In Figure 1, we aim to provide a more visual interpretation to the results from the survival analysis. Each split is shown at the top of each node and can be assumed to be the split that directs observations towards the left side of the tree. For example, after splitting on birth year at 1933, a split to the left corresponds to an individual with few moles. A split to the right indicates an individual with many moles and so on. The terminal nodes in this figure are displayed as Kaplan-Meier plots, showing the survival curves along with the number of observations *n*, and the risk ratio, calculated with reference to the baseline group of 9669 individuals. The reference survival curve is shown at the root node along with a bar which shows the percentage of events in the entire dataset. Grids are placed on each survival plot at equal intervals of 20 years of age. These grids have been plotted to simplify the interpretation. An indicator to the left of the plot displays the risk ratio. Ratios above one are shown by an "up arrow", while a decrease in risk is illustrated by a "down arrow". From this model we can see that there are a few scenarios that indicate high risk for earlier onset of melanoma. These scenarios may be described as:

- individuals born after 1966 with many moles (RR = 11.3)
- individuals born between 1947 and 1966 with many moles (RR = 4.5)
- individuals born between 1933 and 1953 with few moles, but many freckles (RR = 2.4)
- individuals born between 1923 and 1933 with many moles (RR = 1.41)

It is obvious from these results that birth year has a substantial impact on the age-at-onset of melanoma. Once this is taken into account, mole count and freckling only provide a small contribution to the risk.

Family History of Melanoma

As described earlier, in the Data section, family history was collected regarding first-degree (siblings and parents) and second-degree relatives. From this pedigree structure, other higher-order types of relative pairs could also be formed. Some of these relatives were diseased with melanoma, resulting in pairs of relatives who both may be diseased (++), both not diseased (—), and one diseased while the other was not diseased (+–). Table 2 lists the concordant and discordant pairs of specific relationships: sib–sib, parent–child, and second-degree/lower order relative pairs. The second-degree relative pairs include grandparent–grandchild, and aunt–niece, while the lower order relative pairs include the in-law pairs. From Table 2, the percentages of both diseased pairs are 0.6%, 0.9% and 0.4% among sib–sib, parent–child, and second/lower order pairs, respectively. Crude estimates of correlation coefficients can be calculated from these percentages, without accounting for ages at onset among these relatives. However, the risk of developing melanoma may depend on the subject's age. Hence, adjusting for the age at onset is essential in quantifying the correlation of age at onset between pairs of relatives. In a subsequent section, we describe how this can be done rigorously via the generalized estimating equations approach. On average, melanomas were diagnosed slightly earlier in relatives (47.5 years) than in probands (50.2 years). Among relatives, melanomas were diagnosed at younger ages in later generations.

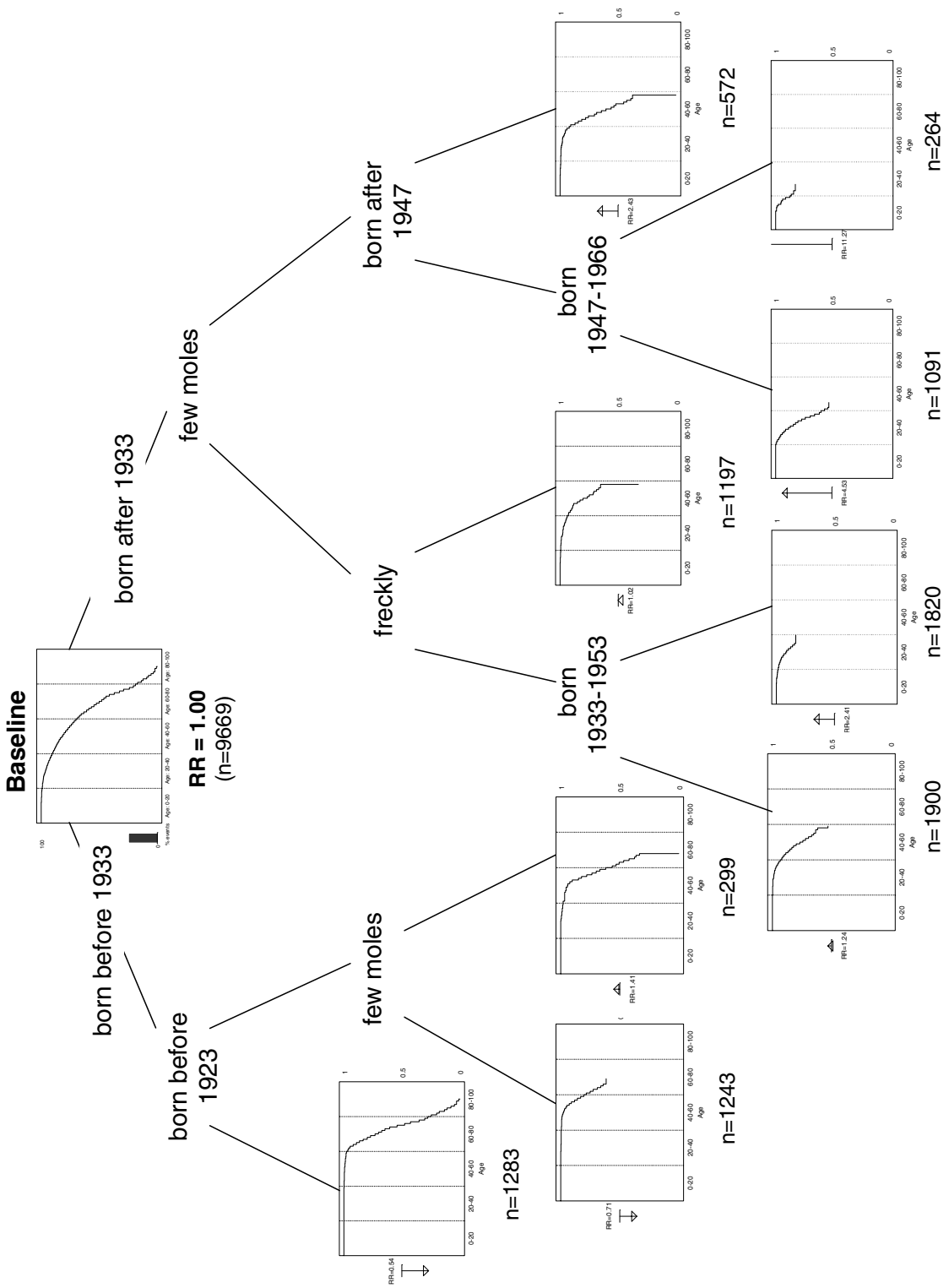

Figure 1
Survival tree for the melanoma data consisting of 9669 individuals from 1912 Queensland families. Terminal nodes are rectangular and display the Kaplan-Meier curve for the corresponding subgroup.

Table 2

Concordant and Discordant Pairs of Relatives in 1912 Families from the Queensland Familial Melanoma Project. Probands Were Not Included for the Calculation of Concordance

	Sib–sib	Parent–child	Second/Others	Total
++	49	15	41	105
+–	763	536	1200	2499
––	7011	1078	9817	17,906
Total	7823	1629	11,058	20,510

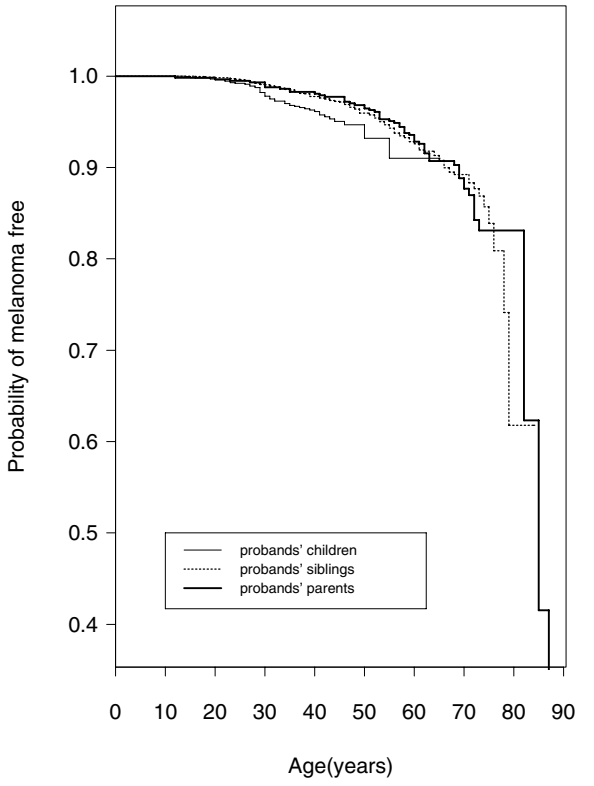

Figure 2

Cumulative melanoma-free survival among first-degree relatives of confirmed cutaneous melanoma cases (probands) diagnosed in Queensland, Australia, 1982–1987, according to the relative's relationship to the proband.

To account for the different ages at censoring in each generation due to termination of the study or death from causes other than melanoma, we examined the disease-free survival distribution for each generation using the standard failure-time analysis technique (Figure 2). The median age at diagnosis of melanoma was 64 among parents of probands, 50 among siblings of probands, and 33 among children of probands. The disease-free survival functions differed significantly between the children generation from earlier generations (log-rank test with $p < .01$); but there was no significant differences in age-at-onset of melanoma between the siblings and parents of probands (log rank test with $p > .9$).

Preliminary Segregation Analysis

A preliminary segregation analysis was conducted (Aitken et al., 1998) to investigate whether the familial clustering of cutaneous melanoma is consistent with Mendelian inheritance of a major autosomal gene. Analyses were performed with the SAGE (1992) statistical package, using the maximum likelihood REGTL program for a binary trait with a variable age of onset. The hypothesis of co-dominant Mendelian inheritance gave a significantly better fit to the data than either dominant or recessive Mendelian inheritance. Overall, both Mendelian inheritance of a single major gene, and purely environmental transmission were rejected. However, there was strong evidence of familial dependence in melanoma occurrence.

The inclusion of risk factors in the models may reveal whether all or a combination of these explains the familial dependence that was demonstrated by the segregation analysis. If a major gene exists that operates independently on these covariates, including them in the models may reduce residual variation and increase the power of the analysis to detect a major gene effect.

Estimating Equations Approach

Let $y = (\delta_{ki}, t_{ki}, Z_{ki})$ denote the data collected for the i^{th} member in the k^{th} family ($k = 1,\ldots,K$ and $i = 1,2$) where $\delta_{ki} = 0$ if the observation is censored, t_{ki} is either the recorded age at diagnosis of melanoma or the age at the most recent follow-up for unaffected people, and Z_{ki} is a vector of measured covariates. We assume that censoring time, age at diagnosis of melanoma and the covariates are independently distributed. These assumptions can be relaxed in more general models, subject to identification constraints. The hazard rate for melanoma is the instantaneous probability that melanoma is diagnosed immediately after time t, given that the person is unaffected at time t. The hazard rate under the Cox proportional hazards model (Cox, 1972) is given by

$$\lambda(t_{ki}) = \lambda_0(t_{ki})\ exp(\beta' Z_{ki})$$

where $\lambda_0()$ is the baseline hazard function, and β is a vector of regression coefficients.

For a specific pair of relatives, we follow Clayton (1978) in modeling the bivariate survivor function

$$F(t_{k1}, t_{k2}) = (F_1(t_{k1})^{-\theta} + F_2(t_{k2})^{-\theta} - 1)^{-1/\theta}$$

where F_1 and F_2 are univariable survivor functions, θ is a scalar parameter that measures the degree of dependence between the relatives' times at onset, independence being implied by $\theta = 0$, and positive association by $\theta > 0$. The Clayton model allows negative dependencies and has the property that failure times are absolutely continuous for $\theta > -0.5$. In addition, the cross-ratio (or odds-ratio) function as studied by Oakes (1989) is

$$c(t_{k1}, t_{k2}) = \lambda(t_{k1} \mid T_{k2} = t_{k2}) / \lambda(t_{k1} \mid T_{k2} \geq t_{k2}) = 1 + \theta.$$

This is equivalent to assuming that the odds-ratio is invariant over the grid region that supports the data. Heuristically, the parameter $1 + \theta$ is an odds-ratio that depends on the degree of dependence between the onset ages of the two relatives. If genetic factors do influence the age at onset of melanoma, we would expect to see a higher concordance in the age of onset in first degree relatives who on average, share half their genes in common in comparison to second degree relatives. Under the current model, this translates as $\theta_{first\text{-}degree} > \theta_{second\text{-}degree}$.

We may use a standard method to estimate within pair correlations for 2×2 tables from odds ratios. Estimates of relation-pair correlations, $p_{sib\text{–}sib}$, $p_{parent\text{–}child}$, and $p_{second\text{-}order}$ are recovered from using the relationships; for example, $r_{sib\text{–}sib} = min(1, ln(1 + \theta_{sib\text{–}sib}))$. Testing for the presence of genetic factors underlying the age at diagnosis of melanoma is equivalent to testing H_0: $p_{first\text{-}order} = p_{second\text{-}order}$. We may test this hypothesis using a z-transform (Kendall, 1979, p. 315) of the point estimates of the correlation coefficients. Let n_1 and n_2 denote the number of first order and second order relatives, let $z_1$and z_2 denote the transformed statistics of $r_1$and r_2, the correlation estimates for first-order and second-order relative pairs respectively. Specifically, we reject H_0 when $E/D > Z_{1-\alpha}$, where

$$E = E(z_1 - z_2) = \frac{1}{2} log \left[\left(\frac{1 + r_1}{1 - r_1} \right) \left(\frac{1 - r_2}{1 + r_2} \right) \right]$$

$$D^2 = V(z_1 - z_2) = \frac{1}{(n_1 - 3)} + \frac{1}{(n_2 - 3)}$$

and $Z_{1-\alpha}$ is the standard normal deviate corresponding to the one-sided α significance level.

This approach has the advantage of providing a test for the presence of genetic effects through a single parameter (θ). However, it is limited in its ability to attribute the phenotypic variance to specific effects (e.g., additive gene action).

Mathematical details of the GEE model and the iterative procedure to estimate the regression coefficients β and specific degrees of dependence θ for the different types of relative pairs have been summarized previously in Do et al. (2000).

MCMC Analysis Using BUGS

The Bayesian Paradigm and Gibbs Sampling

Markov Chain Monte Carlo (MCMC) is an alternative Bayesian approach that provides estimates of likelihoods and associated parameter values when exact computation is infeasible (Hastings, 1970; Metropolis et al., 1953). MCMC methods can be used to draw samples from the underlying joint distribution of major genotypes and polygenic values, conditional on the observed data. From these samples, desired parameters and likelihoods can be estimated without the need to resort to exact computation. MCMC methods have been used for linkage analysis (Kong et al., 1992; Lange & Sobel, 1991), for estimation of parameters in the mixed model with and without covariates (Guo & Thompson, 1991; Thomas, 1992), for estimation of gene–smoking interaction and covariate imputation (Gauderman et al., 1997), for performing combined linkage and segregation analysis (Faucett et al., 1993; Guo & Thompson, 1992), and for mixed models of large complex pedigrees (Guo & Thompson, 1994).

In a general setting, let y be the observed data, and θ be everything not observed including parameters and latent variables. The implementation of Bayesian methods using realistic models and priors is computer-intensive and relies heavily on cunning computational tools to approximate integrals. The problem, in general terms, is to obtain the expected value of a function of interest s (.) under the posterior density $p(\theta \mid \mathbf{x})$

$$E[s(\theta)] = \frac{\int_\Theta s(\theta)\, p(\theta)\, p(\mathbf{x} \mid \theta)\, d\theta}{\int_\Theta p(\theta)\, p(\mathbf{x} \mid \theta)\, d\theta}$$

which cannot generally be found analytically. One method to carry out the integration on the RHS is to perform simulation of exact Bayesian posterior distributions using Markov chain Monte Carlo techniques such as Gibbs sampling. The Gibbs sampler (Geman & Geman, 1984) is the most popular algorithm used in MCMC applications to correlated data. Gibbs sampling was introduced to the main statistical community by Gelfand and Smith (1990), and has since been applied to an even wider array of problems. The Gibbs sampler is easy to implement because it only depends on the local neighborhood structure. In the context of pedigree analysis (Olshen & Wijsman, 1996), the basic procedure is a sequential updating of missing and latent data including the underlying and unobserved major genotypes, polygenic effects, and

environmental effects. Values for the missing or latent data are sampled from the local conditional distribution, a function of the observed individual data, the current sampled values of other missing/latent data for this particular individual such as polygenic and environmental effects, and the values for the sampled genetic effects in the immediate neighbors of an individual. Gibbs sampling basically consists of three main steps:

- *Step 1*: Setting initial values for unobserved quantities (parameters and latent variables),
- *Step 2*: For each parameter or latent variable θ_j, sample from its "full conditional distribution" given the current values of all other quantities in the model,
- *Step 3*: Examine sampled values of parameters and latent variables to monitor convergence and to provide summary measures.

Some of the most recent and popular packages that implement Gibbs sampling for analysis of pedigree data include BUGS (Gilks et al., 1994; Spiegelhalter et al., 1996a,b), Genetic Analysis Package (GAP, 1996), and MIXD (Olshen & Wijsman, 1996; Thompson, 1994). We have used BUGS mainly because of its flexibility in programming hierarchical models besides being a freeware product.

The Model

The aim here is to model the correlation structure within the family structure to satisfy the fundamental additive genetic model (Crow & Kimura, 1970; Falconer, 1990; Kempthorne, 1960) as follows. Consider a nuclear family structure consisting of 4 members: father, mother, and two children denoted by *F, M, S_1, S_2* respectively. Using similar notation as in Burton et al. (1999), a conventional mixed linear model consisting of fixed and random effects may be written in the form

$$Q_{ij} = \beta^T \mathbf{z} + A_{ij} + C_{ij} + C_{S_{ij}} + E_{ij}$$

where *Qij* is the observed value of a normally distributed continuous trait for the j^{th} individual in the i^{th} nuclear family; z_{ij} is a vector of observed covariates representing fixed effects, and β is a corresponding vector of unknown fixed regression coefficients; A_{ij}, C_{ij}, and $C_{S_{ij}}$ A denote random effects that represent additive polygenic, common family environment, and common sibling environment effects respectively. The variation in an individual response is represented by a composite covariance matrix, V_T, and is the sum of an additive genetic covariance matrix V_A, a common family environment matrix V_C, a shared sibling environment matrix V_{Cs}, and residual environmental effects. The different variance components are

$$V_A = \begin{array}{c} \\ F \\ M \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{cccc} F & M & S_1 & S_2 \end{array} \\ \begin{pmatrix} \sigma_A^2 & 0 & \frac{1}{2}\sigma_A^2 & \frac{1}{2}\sigma_A^2 \\ 0 & \sigma_A^2 & \frac{1}{2}\sigma_A^2 & \frac{1}{2}\sigma_A^2 \\ \frac{1}{2}\sigma_A^2 & \frac{1}{2}\sigma_A^2 & \sigma_A^2 & \frac{1}{2}\sigma_A^2 \\ \frac{1}{2}\sigma_A^2 & \frac{1}{2}\sigma_A^2 & \frac{1}{2}\sigma_A^2 & \sigma_A^2 \end{pmatrix} \end{array}$$

$$V_C = \begin{array}{c} \\ F \\ M \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{cccc} F & M & S_1 & S_2 \end{array} \\ \begin{pmatrix} \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 & \sigma_A^2 & \sigma_A^2 \end{pmatrix} \end{array}$$

$$V_{Cs} = \begin{array}{c} \\ F \\ M \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{cccc} F & M & S_1 & S_2 \end{array} \\ \begin{pmatrix} \sigma_{Cs}^2 & 0 & 0 & 0 \\ 0 & \sigma_{Cs}^2 & 0 & 0 \\ 0 & 0 & \sigma_{Cs}^2 & \sigma_{Cs}^2 \\ 0 & 0 & \sigma_{Cs}^2 & \sigma_{Cs}^2 \end{pmatrix} \end{array}$$

The overall total covariance matrix is

$$V_T = \begin{array}{c} \\ F \\ M \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{cccc} F & M & S_1 & S_2 \end{array} \\ \begin{pmatrix} \sigma_A^2+\sigma_C^2 & \sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 \\ +\sigma_{Cs}^2+\sigma_{EP}^2 & & & \\ \sigma_C^2 & \sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 \\ & +\sigma_{Cs}^2+\sigma_{EP}^2 & & \\ \frac{1}{2}\sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 & \sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 \\ & & +\sigma_{Cs}^2+\sigma_{EP}^2 & +\sigma_{Cs}^2 \\ \frac{1}{2}\sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 & \frac{1}{2}\sigma_A^2+\sigma_C^2 & \sigma_A^2+\sigma_C^2 \\ & & +\sigma_{Cs}^2 & +\sigma_{Cs}^2+\sigma_{EC}^2 \end{pmatrix} \end{array}$$

The *components of variance* $\sigma_A^2, \sigma_C^2, \sigma_{Cs}^2$ need not be positive as long as V_T is positive definite. A negative value for σ_{Cs}^2 simply implies that the realized covariance between siblings is *less* than the realized

covariance between parent and a child. A negative value for σ^2_{Cs} may suggest dominance. To optimize convergence in BUGS, Model (1) may be reparameterized as

$$Q_{ij} = \begin{cases} \alpha + \beta'\mathbf{z} + F_i + G_i + R^{P}_{ij} & \text{for fathers} \\ \alpha + \beta'\mathbf{z} + F_i - G_i + R^{P}_{ij} & \text{for mothers} \\ \alpha + \beta'\mathbf{z} + F_i + H_i + R^{C}_{ij} & \text{for children} \end{cases}$$

where F_i, G_i, H_i are independent additive random effects or latent variables; and are the residual error terms for parents and children respectively. If we model

$$F_i \sim N(0, \frac{1}{2}\sigma^2_A + \sigma^2_C),$$

$$G_i \sim N(0, \frac{1}{2}\sigma^2_A), H_i \sim N(0, \sigma^2_{Cs}),$$

then the basic genetics covariance model (additive genetic, common environment and unique environment for this particular four-member family structure is satisfied.

In survival models, unobserved or unmeasured explanatory variables, some of which may be genetic, are often referred to as frailties. The frailties take values restricted to the positive line and may be assumed to act multiplicatively on the hazard. Extending the above model to correlated family data with time-to-onset endpoint, a multiplicative individual heterogeneity or frailty term representing the latent genetic and common environment variables may be modeled as random effects simultaneously with the effects associated with observed covariates. Consider right censored time to onset of melanoma data $\{(T_{ij}, \delta_{ij}, z_{ij}); 1 \le j \le n\}$ from n relative pairs; here T_{ij} denotes the true age at onset of the j^{th} family member or the censored time depending on whether δ_{ij}=1 or 0 respectively, and z denotes a $p \times 1$ vector of covariates. A Weibull distribution may be used to model time to failure as

$$f(ti, zi) = e^{\beta' z_i} \gamma t_i^{\gamma-1} exp(-e^{\beta' z_i} t_i^{\gamma})$$

where β is a vector of unknown regression coefficients, and γ is the shape parameter of the Weibull distribution. This leads to a baseline hazard of the form

$$\lambda_0(t_i) = \gamma t_i^{\gamma-1}.$$

Re-parameterize by letting $\mu_i = e^{\beta' z_i}$ the conditional distribution of t_i given μ_i is then Weibull (γ,μ_i). We formulated a mixed model to represent the conditional distribution of t_{ij} given covariate effects, random additive genetic and common environment effects as

$$t_{ij} \mid \mu_{ij} \sim \text{Weibull}\ (\gamma, \mu_{ij})\ i = 1,\ldots n;\ j = 1,2$$

where

$$\log \mu_{ij} = \begin{cases} \alpha + \beta'\mathbf{z} + F_i + G_i + R^{P}_{ij} & \text{for fathers} \\ \alpha + \beta'\mathbf{z} + F_i - G_i + R^{P}_{ij} & \text{for mothers} \\ \alpha + \beta'\mathbf{z} + F_i + H_i + R^{C}_{ij} & \text{for children} \end{cases}$$

The regression coefficients and the precision of the random effects (τ_G, τ_F, τ_H) were given "non-informative" Normal and Gamma priors respectively. The shape parameter, γ, of the time to onset of melanoma distribution was also given a non-informative Gamma prior which was slowly decreasing on the positive real line.

The overall model (including random effects) may be described in a Bayesian graph (Figure 3) which simplifies sampling from full conditional distributions by exploiting partial independence properties (Spiegelhalter et al., 1996a). In this graph, each random quantity is represented by a node, which may be connected by directed or undirected links. Conditional independence assumptions are represented by the absence of such links.

We implemented the Gibbs sampler using the BUGS program (Gilks et al., 1994), (code in the Appendix). Imputation of missing data was handled naturally in the Gibbs sampling framework by treating missing values as additional unknown quantities and randomly sampling values from their full conditional distributions. We chose simple prior distributions for imputation, since the number of missing values for covariates was not large (complete for birth year, 16% missing for naevi and 21% missing for freckles) and there was no indication of non-random missingness in our data. Therefore imputation for missing naevi and freckles covariates were based on Bernouilli prior distributions with respective parameter values estimated from the complete observations. We performed an initial 10,000 burn-in iterations followed by an additional 20,000. Parameter estimates were the mean and standard deviation (SD) of all post convergence Gibbs samples with a thinning interval between 20 and 50; credible intervals were computed as the lower and upper $\alpha/2$ percentiles from the last 20,000 iterations. Convergence to the posterior distribution was confirmed by using the different criteria provided by the add-on CODA package including those of Gelman and Rubin (1992), Geweke (1992), and Raftery and Lewis (1992a, 1992b).

Results

GEE Approach

An inspection of residual plots following preliminary model-fitting provided no evidence for the failure of proportional hazards assumption and did not detect influential observations. We then proceeded to apply the GEE approach that could estimate regression coefficients while incorporating a dependence structure between relative pairs. The results are displayed

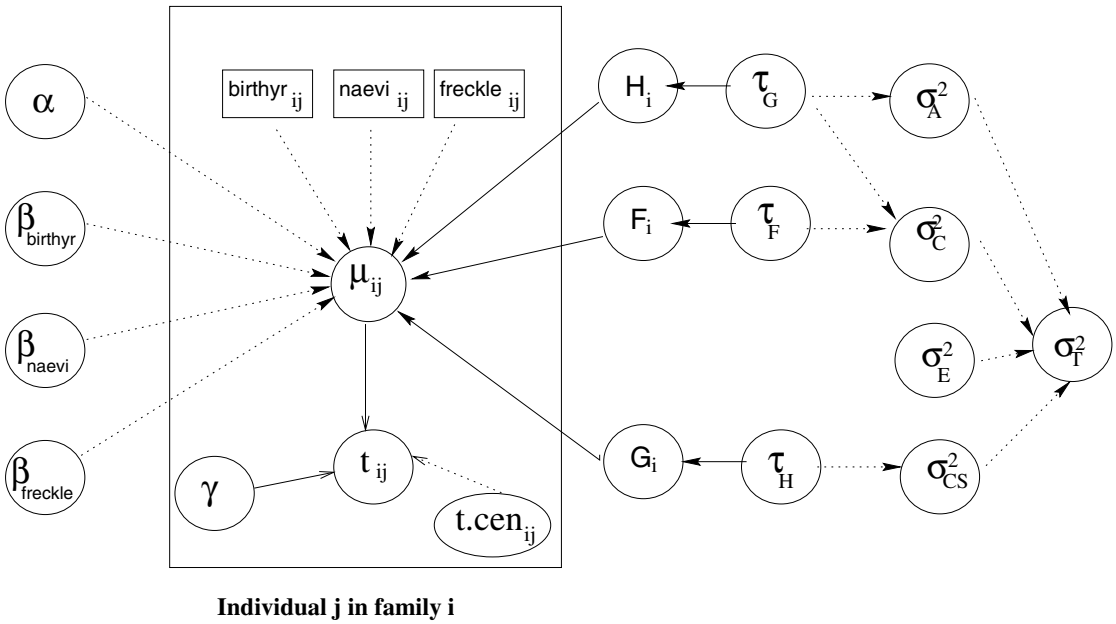

Figure 3

Graphical model of covariate and random family effects for an individual in a nuclear family structure. t_{ij} represents the observed failure time for the j^{th} individual in the i^{th} family with t.cen$_{ij}$ being an indicator variable of censoring status. Full arrows indicate stochastic links to which a probability is attached; broken arrows denote deterministic relationships; βs are regression coefficients, T is the precision of the prior distribution and equals the inverse of the variance; F_i, G_i, H_i are random effects modeled as $F_i \sim N(0,1\backslash 2\ \sigma^2_A + \sigma^2_C)$, $G_i \sim N(0, 1/2\ \sigma^2_A)$ and $H_i \sim N(0, \sigma^2_{Cs})$. Rectangles represent actual data values for the covariates; γ and μ_{ij} are shape and scale parameters for the underlying Weibull distribution.

in Table 3 which suggest that later birth year, having at least a moderate number of naevi and freckles were all simultaneously associated with later age at onset of melanoma. Table 3 also presents the estimated odds-ratios for quantifying the correlation between paired relatives of specific relationships.

The odds ratio for sib–sib pairs is 2.973 which is significantly different from 1 ($p < 0.01$). The odds ratio for parent–child pairs is 1.650 which is slightly greater than 1, indicating a mild dependency between these pairs, although they are not quite statistically significant. The odds ratio for second-degree and higher relative pairs is 1.155, indicating no dependence at all between these pairs. This pattern of familial aggregation is compatible with dominance variance as well as additive genetic variance (Falconer, 1990).

Bayesian Approach

We re-analyzed the age-at-onset of melanoma by using Gibbs sampling to impute missing covariates and to estimate subject-specific covariate effects, random additive genetic, common family environment, and shared sibling environment effects on the log scale. The estimated shape parameter γ was 4.3 with a 95% credible interval (CI) of (4.2, 4.6). The results are summarized in Table 4 (Model A). The residual plots did not indicate a gross departure from the underlying Weibull model and revealed no influential observations. We checked the sensitivity of the analyses to initial parameter values by re-running the Gibbs sampler five more times using different starting values. The resulting estimates did not differ by more than 5% from the values reported here. The mean estimate for (σ^2_A) was 0.452 with 95% CI = (0.348, 0.566), for (σ^2_C) was –0.053 with 95% CI = (–0.120, 0.019), and for (σ^2_{Cs}) was 0.467 with 95% CI = (0.393, 0.545).

A small negative value for the common family effect suggests that there may be a dominant effect, or that the current model is not quite appropriate, for example, that there are systematic effects or correlation structures that have not been accounted for. The results here indicate that additive genetics seem to impact equally on the variation of the age at onset of melanoma. Further exploration for alternative models is a focus of our future research.

In addition, we investigated the relative contribution of genetic and environmental effects on the expression of naevi (Model B) and freckles (Model C), which are known risk factors for melanoma. The expressions of naevi and freckles were coded as binary variables (none or few moles versus moderate or many moles; and no freckles versus one or more freckles). A hierarchical Bayesian binomial model was fitted to estimate the random variance components. The results in Table 4 indicate that a common family environment effect contributed the most to the expression of naevi ($\sigma^2_C = 0.704$) (relative to the contributions of additive genetic effect ($\sigma^2_A = 0.142$) and

Table 3

GEE Approach: Estimated Regression Coefficients in the Proportional Hazard Model and Estimated Odds Ratios for Quantifying Familial Aggregation in Age at Onset of Melanoma in Queensland Families (** Indicates Significance at the .05 Significance Level)

A. Mean effects				
Covariate	RR = e^{β}	Coefficient β	Robust se(β)	Z-statistic
Year of birth	1.142	0.132	0.051	2.588**
Naevi (Baseline = No or few moles)	1.765	0.568	0.073	7.781**
Freckling (Baseline = No freckles)	1.160	0.148	0.049	3.020**
B. Pattterns of familial aggregation				
Relationship		1 + θ	se(θ)	Z-statistic
Sib–sib		2.973	0.6217	3.17**
Parent–child		1.650	0.434	1.50
Second/Others		1.155	0.270	0.47

Table 4

Gibbs Sampling Approach: Estimated Regression Coefficients and Estimated Variance Components in a Melanoma Study of Queensland Families

Weibull Model: A. Mean effects — Response variable is Age-at-onset				
Covariate	RR = e^{β}	Coefficient β	Robust se(β)	95% CI of β
Year of birth	1.378	0.321	0.0027	(0.316,0.326) **
Naevi	1.126	0.119	0.0021	(0.058,0.185) **
Freckling	1.017	0.017	0.1400	(–0.005,0.085)
Weibull Model: B. Variance components — Response variable is Age-at-onset				
Latent effect		Mean from 5000 iterations	se(σ^2)	95% CI of σ^2
σ^2_A		0.452	0.054	(0.348,0.566) **
σ^2_C		–0.053	0.027	(–0.120,0.019)
σ^2_{Cs}		0.467	0.040	(0.393,0.545) **
γ		4.3	0.104	(4.2,4.6)
Binomial Model: Variance components — Response variable is Naevi				
Latent effect		Mean from 5000 iterations	se(σ^2)	95% CI of σ^2
σ^2_A		0.142	0.149	(0.002,0.498) **
σ^2_C		0.704	0.156	(0.403,1.010) **
σ^2_{Cs}		0.195	0.157	(0.0025,0.553) **
Binomial Model: Variance components — Response variable is Freckling				
Latent effect		Mean from 5000 iterations	se(σ^2)	95% CI of σ^2
σ^2_A		2.050	0.779	(0.835,3.570)
σ^2_C		2.600	0.418	(1.780,3.460)
σ^2_{Cs}		0.115	0.088	(0.011,0.312)

Note: ** Indicates Significance at the 0.05 level). Naevi Is a Binary Variable with Baseline 0 = No or Few Moles; Freckling Is Coded as a Binary Variable with Baseline 0 = No Freckles

of shared sibling effect(σ^2_{Cs} = 0.142), both of which were non negligible. In contrast, variation in the expression of freckles was largely explained by additive genetic and shared family effects(σ^2_A = 2.050, σ^2_C = 2.600), compared to a relatively small shared sibling effect(σ^2_C = 0.115).

Discussion

We applied two methods — generalized estimating equation and Bayesian analysis — to the genetic analysis of age at onset of melanoma based on a nuclear family structure. Under both approaches, the results suggest that additive genetic factors played an

important role in the age at onset of melanoma but that shared sibling environmental factors were not negligible. We focused attention on these approaches because they are more appropriate for modeling correlated age at onset data and they allow the inclusion of covariates in the analyses. Under both approaches, there were suggestions that earlier melanoma onset was influenced by later birth year, having a moderate number of naevi, and being freckly. The principal difference between the two approaches is in the interpretation of the regression coefficients. The GEE method uses a marginal approach resulting in regression coefficients that describe the average population response to changing covariates, whereas the Bayesian approach produces subject-specific coefficients. A secondary distinction is in the nature of the within-pair dependence. The GEE model only describes a common covariance among specific relative pairs, whereas the Bayesian approach can explicitly describe the source of this covariance. A third advantage of the Bayesian method is its flexibility in incorporating prior information, if available, for the covariates or latent effects by modifying their prior distributions. Further, the Bayesian method would permit a more accurate decomposition of the genetic variance into additive and dominant components, thus providing the means for a direct assessment of the no-dominance assumption. Finally, it is also interesting to record the amount of CPU time required for each method: 20 seconds for the GEE approach and approximately 3 hours (for binary traits) and 12 hours (for age at onset outcome) to run BUGS on a single-user Intel Pentium III 600 MHz personal computer with the Linux Mandrake 7.1 operating system. The amount of human and financial resources dedicated to collecting, maintaining, and updating the melanoma family database is extremely high, therefore, the extra CPU time requirement by the MCMC method is well worth the additional genetic information and flexibility that it provides. A BUGS program is included in the Appendix.

Acknowledgments

Data collection was supported by grants from the Australian National Health and Medical Research Council (NHMRC) ID 930223 and ID 961061, U.S. National Cancer Institute (CA88363). Programming resources were provided by start-up funds from the Department of Biostatistics at the M. D. Anderson Cancer Center. We acknowledge assistance with BUGS programming from Sang-Joon Lee.

References

Abel, L., & Bonney, G. E. (1990). A time-dependent logistic hazard function for modeling variable age of onset in the analysis of familial diseases. *Genetic Epidemiology, 12*, 391–407.

Aitken, J. F., Bailey-Wilson, J., Green A. C., MacLennan, R., & Martin, N. G. (1998). Segregation analysis of cutaneous melanoma in Queensland. *Genetic Epidemiology, 15*, 391–401.

Aitken, J. F., Green, A. C., MacLennan, R., Youl, P., & Martin, N. G. (1996). The Queensland familial melanoma project: Study design and characteristics of participants. *Melanoma Research, 6,* 155–165.

Allison, P. D. (1995). *Survival analysis using the SAS system — A practical guide.* Cary, NC: SAS Institute Inc.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.

Bernardo, J. M., Berber, J. O., David, A. P., & Smith, F. F. M. (Eds.). (1992).*Bayesian statistics 4*. Oxford: Clarendon Press.

Bonney, G. E. (1986). Regressive logistic model for familial disease and other binary traits. *Biometrics, 42*, 611–625.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth, California.

Cannings, C., & Thompson, E. A. (1977). Ascertainment in the sequential sampling of pedigrees. *Clinical Genetics, 12*, 208–212.

Clayton, D. G. (1978). A model for association in bivariate life-tables and its application in epidemiological studies of chronic disease incidence. *Biometrik*a, *65*, 141–151.

Clayton, D. G., & Cuziak, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A, 148*, 82–117.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B, 34*, 187–202.

Do, K-A., Broom, B. M., Kuhnert, P., Duffy, D. L., Todorov, A. A., Treloar, S. A., & Martin, N. G. (2000). Genetic analysis of the age at menopause by using estimating equations and Bayesian random effects models. *Statistics in Medicine, 19*, 1217–1235.

English, D. R., Armstrong, B. K., Kricker, A., et al. (1997). Sunlight and cancer. *Cancer Causes Control, 8,* 271–283.

Falconer, D. S. (1990). *Introduction to quantitative genetics* (3rd ed.). New York: Longman Group Ltd.

Faucett, C., Gauderman, W. J., Thomas, D., Ziogas, A., & Sobel, E. (1993). Combined segregation and linkage analysis of late-onset Alzheimer's disease in Duke families using Gibbs sampling. *Genetic Epidemiology, 10*, 489–494.

GAP (1996). Genetic Analysis Package, Release 1.0. Pasadena, CA: Epicenter Software.

Gauderman, W. J., Morrison, J. L., Carpenter, C. L., & Thomas, D. C. (1997). Analysis of gene-smoking interaction in lung cancer. *Genetic Epidemiology, 14*, 199–214.

Gauderman, W. J., & Thomas, D. C. (1994). Censored survival models for genetic epidemiology: A Gibbs sampling approach. *Genetic Epidemiology*, *11*, 171–188.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science, 7*, 457–511.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intelligence, 6*, 721–741.

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to calculating posterior moments*. In J. M. Bernardo, J. O. Berber, A. P. David, & F. F. M. Smith (Eds.). *Bayesian statistics 4*. Oxford: Clarendon Press.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modeling. *Statistica, 43,* 169–178.

Guo, S. W., & Thompson, E. A. (1991). Monte Carlo EM for the estimation of multiple random effects models on large pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology, 8*, 171–189.

Guo, S. W., & Thompson, E. A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics, 51*, 1111–1126.

Guo, S. W., & Thompson, E. A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics*, *50*, 417–432.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Heath, A. C. (1997). Markov Chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics, 61,* 748–760.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika, 73*, 671–678.

Hsu, L., & Prentice, R. L. (1996). On assessing the strength of dependency between failure time variates. *Biometrika*, *83*, 491–506.

Hsu, L., & Zhao, L. P. (1996). Assessing familial aggregation of age at onset, by using estimating equations, with application to breast cancer. *American Journal of Human Genetics, 58*, 1057–1071.

Kempthorne, O. (1960). *Biometrical genetics*. New York: Pergammon Press.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2; 4th ed.). Oxford: Oxford University Press.

Kong, K., Frigge, M., Cox, N., & Wong, W. H. (1992). Linkage analysis with adjustments for covariates: A method combining peeling with Gibbs sampling. *Cytogenetics and Cell Genetics*, *59*, 208–210.

Kuhnert, P., & Do, K-A. (2003). Fitting genetic models to twin data with binary and ordered categorical responses: A comparison of structural equation modeling and Bayesian hierarchical models. *Behavior Genetics*, *33*, 439–452.

Lange, K., & Sobel, E. (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics*, *49*, 1320–1334.

LeBlanc, M., & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, *48,* 411–425.

Li, H., & Thompson, E. A. (1997). Semiparametric estimation of major gene and random environmental effects for age of onset. *Biometrics*, *53*, 282–293.

Li, H., & Wijsman, E. A. (1998). Semiparametric estimation of major gene effects for age of onset. *Genetic Epidemiology, 15,* 279–298.

MacLennan, R., Green, A. C., McLeod, G. R., & Martin, N. G. (1992). Increasing incidence of cutaneous melanoma in Queensland, Australia. *Journal of the National Cancer Institute*, *84*, 1424–1432.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087–1092.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sorenson, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics, 19*, 25–43.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of American Statistical Association, 84,* 487–493.

Raftery, A. L., Lewis, S. (1992a). Comment: One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, *7,* 493–497.

Raftery, A. L. & Lewis, S. (1992b). *How many interactions in Gibbs sampler?* (pp. 763–774). In J. M. Bernardo, J. O. Berber, A. P. David, & F. F. M. Smith (Eds.), *Bayesian statistics 4*. Oxford: Clarendon Press.

S.A.G.E. (1992). *Statistical Analysis for Genetic Epidemiology, Release 2.1*. Cleveland, OH: Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University.

Scurrah, K. J., Palmer, L. J., & Burton, P. R. (2000). Variance components analysis for pedigree-based censored survival data unusing generalized linear mixed models (GLMMS) and Gibbs sampling in BUGS. *Genetic Epidemiology, 19*, 127–148.

Siegmund, K. D., & McKnight, B. (1998). Modeling hazard functions in families. *Genetic Epidemiology*, *15*, 147–171.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1996a). *Computation on Bayesian graphical models* (pp. 407–425). Oxford: Oxford University Press.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. (1996b). *BUGS 0.6 –Bayesian InferenceUsing Gibbs Sampling manual.* Cambridge: MRC Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D. (2003). *WinBUGS* (Bayesian Inference Using Gibbs Sampling) Windows version 1.4 user manual. Cambridge: MRC Biostatistics Unit.

Siskind, V., Aitken, J., Green, A., & Martin, N. (2002). Sun exposure and interaction with family history in risk of melanoma, Queensland, Australia. *International Journal of Cancer, 97*, 90–95.

Swerdlow, A., & Green, A. (1987). Melanocytic naevi and melanoma: An epidemiological perspective. *British Journal of Dermatology, 117*, 137–146.

Therneau, T. M., & Atkinson, E. (1997). *An introduction to recursive partitioning using the RPART routines* (Technical report). Rochester: Mayo Foundation.

Thomas, D. (1992). Fitting genetic data using Gibbs sampling: An application to nevus counts in 38 Utah kindreds. *Cytogenetics and Cell Genetics, 59*, 228–230.

Thompson, E. A. (1994). *Monte Carlo programs for pedigree analysis*. (Technical Report 267). Washington: Dept. of Statistics, University of Washington.

Vaupel, J. M., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16*, 439–454.

Appendix A

A BUGS Program to Implement the Bayesian MCMC approach in Modeling an ACE Model for Age at onset of Melanoma with Covariate Effects and Random Genetics/ Environmental Effects

```
# Data: melsub.dat
#       No. of Members  = 6819
#       No. of Families = 1450
#
# Response:  Age at Onset
#
model gene;

const   Members=6819,
        Fam=1450;

var     yrbth[Members],naevi[Members],freckle[Members],F[Fam],G[Fam],H[Fam],
        FamIND[Members],IndexF[Members],AgeFU[Members],Affect[Members],
        alpha,beta.yrbth,beta.naevi,beta.freckle,r,mu[Members],
        p.naevi[4],p.freckle[3],tauF,tauG,tauH,VF,VG,VH,VA,VC,VCS;

data    IndexF,FamIND,yrbth,naevi,freckle,AgeFU,Affect in "melsub.dat";

inits in "mel.in";
{

# Imputation of missing covariates using straightforward
# Bernoulli parameters estimated from complete data

for(i in 1:Members){
        naevi[i] ~ dbern(0.359);
        freckle[i] ~ dbern(0.669);
}

# The Model

for(i in 1:Members){
    AgeFU[i] ~ dweib(r,mu[i])I(Affect[i],);
    log(mu[i]) <- alpha + beta.yrbth*yrbth[i] + beta.naevi*naevi[i] +
                beta.freckle*freckle[i] +
                equals(FamIND[i],1)*(F[IndexF[i]] + G[IndexF[i]]) +
                equals(FamIND[i],2)*(F[IndexF[i]] - G[IndexF[i]]) +
                equals(FamIND[i],3)*(F[IndexF[i]] + H[IndexF[i]]);
}

for(j in 1:Fam){
    F[j] ~ dnorm(0.0,tauF);
    G[j] ~ dnorm(0.0,tauG);
    H[j] ~ dnorm(0.0,tauH);
}




# Priors

alpha ~ dnorm(0.0,0.001);
beta.yrbth ~ dnorm(0.0,0.001);
beta.naevi ~ dnorm(0.0,0.001);
beta.freckle ~ dnorm(0.0,0.001);
tauF ~ dgamma(0.001,0.001);
tauG ~ dgamma(0.001,0.001);
tauH ~ dgamma(0.001,0.001);
r ~ dgamma(0.001,0.001);

VF <- 1/tauF;
VG <- 1/tauG;
VH <- 1/tauH;

VA <- 2*VG
VC <- VF-VG
VCS <- VH
}
```