

# Methods for Testing Familial Aggregation of Diseases in Population-based Samples: Application to Hodgkin Lymphoma in Swedish Registry Data

R. M. Pfeiffer<sup>1,\*</sup>, L. R. Goldin<sup>1</sup>, N. Chatterjee<sup>1</sup>, S. Daugherty<sup>2</sup>, K. Hemminki<sup>3</sup>, D. Pee<sup>4</sup>, L. I. X<sup>5</sup> and M. H. Gail<sup>1</sup>

<sup>1</sup>National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd, Bethesda, MD, 20892-7244, USA

<sup>2</sup>The Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD, 21205-2179, USA

<sup>3</sup>Division of Molecular Genetic Epidemiology, German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

<sup>4</sup>Information Management Services Inc., Rockville, MD, 20852, USA

<sup>5</sup>Department of Biosciences at Novum, Karolinska Institute, 141 57 Huddinge, Sweden

## Summary

We use data on lymphoma in families of Hodgkin lymphoma (HL) cases from the Swedish Family Cancer Database (Hemminki *et al.* 2001) to illustrate survival methods for detecting familial aggregation in first degree relatives of case probands compared to first degree relatives of control probands, from registries that permit sampling of all cases. Because more than one case may occur in a given family, the first degree relatives of case probands are not necessarily independent, and we present procedures that allow for such dependence. A bootstrap procedure also accommodates matching of case and control probands by resampling the matching clusters, defined as the combined set of all first degree relatives of the matched case and control probands. Regarding families as independent sampling units leads to inferences based on “sandwich variance estimators” and accounts for dependencies from having more than one proband in a family, but not for matching. We compare these methods in analysis of familial aggregation of HL and also present simulations to compare survival analyses with analyses of binary outcome data.

Keywords: Familial Correlation, Cluster Data, Marginal Model, Bootstrap, Non-Hodgkin Lymphoma, Matched Design

## Introduction

Detecting familial aggregation of disease can provide an important clue to genetic etiology. This paper describes methods used to analyze data from the Swedish Family-Cancer Database (Hemminki *et al.* 2001a) which contains information on family structure and cancer outcomes, obtained from the Swedish Cancer Registry, for more than 10 million individuals. Multigenerational disease registries afford an opportunity to de-

tect familial aggregation but require special analytical tools.

One approach to assessing familial aggregation is to compute a quantity similar to a standardized incidence ratio by dividing the number of diseased first degree relatives of diseased individuals (probands) by the number expected based on standard population rates (Hemminki *et al.* 2001b, 2001c). This analysis should take into account the fact that every case in a family serves as a proband, which affects variances of the estimates (Goldgar *et al.* 1994). A potential disadvantage of this approach is that the standard population rates may not apply to the registry population, yielding biased results.

\*Corresponding author: Ruth Pfeiffer, National Cancer Institute, 6120 Executive Blvd EPS 8030, Bethesda, MD, 20892-7244. Fax: 301-4020081. E-mail: pfeiffer@mail.nih.gov

An internal comparison that avoids reliance on external rates can be based on the ratio of risk in first degree relatives of case probands to that in the first degree relatives of a random sample of control probands from the registry. Liang (1991) discussed such an analysis of survival data, namely the ages of onset of disease in first degree relatives of unrelated cases and controls. Even though the cases and controls are unrelated, allowance needs to be made for the fact that the relatives' ages at disease onset may be correlated within families (Liang, 1991).

We adapt the methods of Liang (1991) to the registry setting, in which every case serves as a proband. Because some families have more than one case, the case probands cannot be regarded as unrelated or independent as in Liang's work. In the design described below, case probands were matched to unaffected control probands. We present a bootstrap procedure that accounts both for the potential relatedness of case probands, and for the matching of controls, to calculate confidence intervals and test for aggregation.

An alternative analysis treats families sampled from the Swedish Family-Cancer Database, either through case probands or through control probands, as independent sampling units. This analysis accounts for having families with multiple probands but not for matching. We compare the resulting analysis based on a robust "sandwich" estimate of variance with the previously mentioned bootstrap analysis.

We also compare these survival analyses to a simpler analysis that treats the disease outcomes of first degree relatives of probands as binary random variables. We illustrate and compare these methods on Hodgkin Lymphoma (HL) data from the Swedish Family-Cancer Database, and on simulated data. The simulations also give an indication of how large shared frailties need to be in order to induce an appreciable relative risk of disease, comparing relatives of case probands to relatives of control probands.

## Materials and Methods

### Data

The Swedish Family-Cancer Database was constructed as described by Hemminki *et al.* (2001a). Briefly, Statis-

tics Sweden maintains a multigenerational register consisting of individuals born since 1932, linked to their biological parents. This database now contains 10.2 million individuals with defined family structures and has been merged with the Swedish Cancer Registry (1958–1998). Thus, the familial distribution of all registered cancers can be assessed. The database has also been merged with census databases to obtain some demographic information, and with the death notification database to incorporate vital status on all individuals.

The data on individuals who had no children born in or after 1932, and the data on all offspring who died before 1960, are missing from the Swedish Family-Cancer Database. Among offspring who died before 1991, about half are not linked to their parents. 75% of all tumors registered in the Swedish Cancer Registry are included in the Family-Cancer Database. The cancer incidence rates (up to age 70) in the Family-Cancer Database are nearly identical to those in the Cancer Registry.

To study lymphomas in relatives of HL cases, we selected all cases of HL from the Swedish Family-Cancer Database. For each case, we randomly sampled two cancer-free controls, matched for gender, year of birth, and county of residence, from the Swedish Family-Cancer Database. County of residence was used as a matching criterion to allow for regional variability over time in reporting of cancers to the central registry. For each case and control, all first degree relatives were included in the data set. We call the cases and sampled controls "probands".

All members in the Swedish Family-Cancer Database have been partitioned into families, consisting of individuals who are related by blood relationships documented by data in the Swedish Family-Cancer Database. Every diseased individual belongs to one such family, though a family may contain more than one case proband. If there are multiple probands in the same family, some of their first degree relatives can be present in the data set multiple times. The most extreme example is a family with two siblings who are eligible probands. This family is duplicated in the data set, each time with a different sibling chosen as the proband. For any other type of relationship among probands usually some, but not all, of the relatives are replicated. A family or parts of it could also be represented in the data more than

once because of multiple control probands, or control and case probands, in the same family. The replication of individuals in the cluster data construction is needed to obtain unbiased estimates of risks in the first degree relatives of all case probands and of their matched control probands.

In the next section we describe how to account for these various types of dependencies in the statistical procedures.

### Statistical Methods

The basic statistical approach is to compare measures of disease risk in first degree relatives of all case probands in the population with those in first degree relatives of matched control probands.

### Marginal Survival Model

First we present a method that uses ages at cancer onset for the relatives of the probands. We are interested in the population average effect of the proband's disease status on the hazard of disease in first degree relatives, and therefore use a marginal model of the hazard for the  $j$ th individual in matching cluster  $i$ ,

$$\lambda_{ij}(t_{ij} | X_{ij}, Z_{ij}) = \lambda_0(t_{ij}) \exp(\beta X_{ij} + \gamma Z_{ij}) \quad (1)$$

where  $t_{ij}$  is the age of disease onset or censoring,  $Z_{ij} = 1$  if the individual is a first degree relative of a case proband,  $Z_{ij} = 0$  if the individual is a first degree relative of a control proband, and  $X_{ij}$  is a vector of measured covariates. In the examples,  $X_{ij}$  denotes gender, and  $\lambda_0$  denotes the arbitrary baseline hazard function. For each individual the outcome data consist of  $(t_{ij}, \delta_{ij})$ , where the binary variable  $\delta_{ij}$  indicates whether an individual developed disease,  $\delta_{ij} = 1$ , or not,  $\delta_{ij} = 0$ , at age  $t_{ij}$ .

Our data are also left truncated because of restrictions in cancer registry coverage in the Swedish Family-Cancer Database. Thus, a person is at risk in a partial likelihood analysis of the hazard (1) only at or beyond the age of left truncation; namely the individual's age in 1958 if the person is born before 1958, and 0 (no truncation) if the person is born in 1958 or after. The main aim of the analysis is to estimate and test the null hypothesis  $H_0 : \gamma = 0$  of no association between risk and the proband's disease status. Note that we only use the proband's disease status but not his or her failure time in the model.

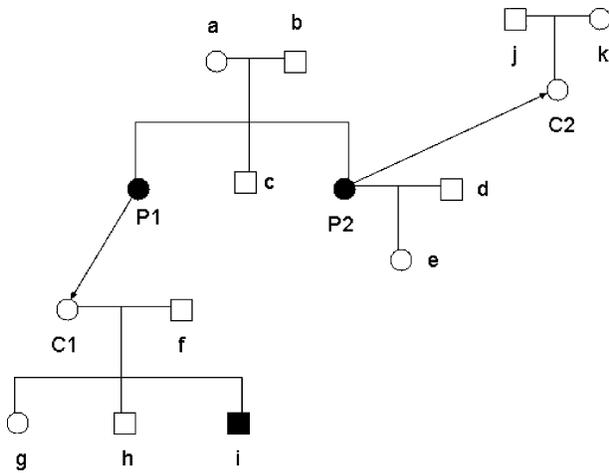
The parameters in the survival model (1) are estimated under a "working independence" assumption, using Proc Phreg, SAS 8.0. The left truncation is taken into account with the *entrytime* statement.

A matching cluster of individuals to be analyzed according to equation (1) is defined as follows. Consider a family that contains one or more case probands. All the first degree relatives of each case proband are included in the cluster, and, if an individual is a first degree relative of more than one proband, that person is entered into the cluster once for each such first degree relationship to a case proband. The cluster also contains the first degree relatives of each control proband who was matched to the case proband(s) in the family. Figure 1 illustrates the construction of such a matching cluster, ignoring replication of first degree relatives. We defined a matching cluster in this way to allow for correlations among all first degree relatives of all probands in a family, as well as with first degree relatives of the matched control. This is analogous to treating the pair as independent sampling units in data for a paired t-test. This definition of a matching cluster also allows for the repeated use of a relative, once for each proband to which he or she is related, in order to obtain unbiased estimates of relative risk for first degree relatives of case versus control probands.

Notice that matching clusters as defined in Figure 1 may overlap. For example, individual  $i$  in Figure 1 is a case proband for his family. He therefore induces another matching cluster containing individuals C1, f, g and h as first degree relatives, together with the first degree relatives of the control probands (not shown) matched to case  $i$ . This second cluster thus includes individuals f, g and h in common with the cluster in Figure 1. However, there is little chance of overlap in our data in clusters defined as in Figure 1, and we therefore treat such matching clusters as independent sampling units in the bootstrap and jackknife procedures described next, to compute the covariance of the estimates.

### Bootstrapping Matching Clusters

Consider the set of matching clusters defined as in Figure 1 as the combined set of all first degree relatives of the matched case and control probands in families that contain at least one case proband. Sample these



**Figure 1** A Matching Cluster of Individuals in the Analysis File. Case probands P1 and P2 in the family lead to the inclusion of first-degree relatives (a,b,c) and (a,b,c,d,e) respectively in the matching cluster. Note that a, b and c are included twice. The corresponding matched controls C1 and C2 lead to the inclusion of individuals (f,g,h,i) and (j,k) respectively. Solid symbols denote diseased individuals, circles females, and squares males. Arrows indicate matching. Note that because individual i is a case, he will form the basis of another matching cluster based on his family.

clusters with replacement in bootstrap replication  $b$  and obtain the estimates  $(\hat{\beta}_b, \hat{\gamma}_b)$  from this sample. Repeat for bootstrap samples  $b = 1, 2, \dots, B$ . The upper and lower 2.5% percentiles of the bootstrap distribution are used to produce 95% confidence intervals on the parameters. Alternative confidence intervals could be based on  $\hat{\gamma} \pm 1.96 \widehat{SE}(\hat{\gamma})$ , where the estimated standard error  $\widehat{SE}(\hat{\gamma})$  is the square root of the sample variance computed from  $\hat{\gamma}_b$  for  $b = 1, 2, \dots, B$ . One can also estimate the variance from a jackknife as  $\sum (\hat{\gamma}_{(i)} - \bar{\gamma})$ , where  $\hat{\gamma}_{(i)}$  is the estimate obtained with cluster  $i$  excluded from the sample, and  $\bar{\gamma}$  is the mean of the  $\hat{\gamma}_{(i)}$ . As there are thousands of matching clusters in our data, the jackknife computations were too slow to be practical, and we do not present results for this. In the studies below we used  $B = 1000$  bootstrap samples.

### Sandwich Estimates of Variance Based on Families as Independent Sampling Unit

An alternative approach to variance estimation regards families, which are mutually exclusive, as independent sampling units, rather than the previously defined

matching clusters. For each selected family, the analysis data set included each relative once for every proband in the family to which he or she was related in the first degree. The sandwich estimate of the covariance matrix (Wei *et al.* 1989) was obtained from a SAS/IML program, which performed matrix multiplication using the score residuals from *dfbeta* from Proc Phreg, that were sorted and averaged by family as input (Therneau & Hamilton, 1997).

Confidence intervals were based on  $\hat{\gamma} \pm 1.96 \widehat{SE}(\hat{\gamma})$  and two-sided  $p$  values were estimated from asymptotic normal theory based on  $\hat{\gamma}$  and  $\widehat{SE}(\hat{\gamma})$ . While the sandwich variance does not fully account for the matched design, it is computationally very simple.

### Marginal Model for Binary Diseases Status

For rare outcomes, the marginal survival model presented in the previous section is closely related to a marginal model of binary outcomes analyzed with generalized estimating equations (GEE). We thus compared the results of the survival analyses with those obtained from a GEE model that treats the outcomes as binary random variables, as outlined by Liang & Beauty (1991). As in the sandwich method for survival analysis we treat the family as the independent sampling unit. Again,  $X_{ij}$ , which includes an intercept, stands for measured covariates for relative  $j$  in family  $i$ , and  $Z_{ij}$  denotes the indicator of the proband's disease status. Relatives were replicated in the family's analysis file once for each first degree relationship to a proband in a family. The disease status of the  $j$ th relative of the  $i$ th family is  $Y_{ij}$ , where  $Y_{ij} = 1$  if the individual develops disease and 0 otherwise, and is modelled using logistic regression

$$P(Y_{ij} = 1 | X_{ij}, Z_{ij}) = \frac{\exp(\beta X_{ij} + \gamma Z_{ij})}{1 + \exp(\beta X_{ij} + \gamma Z_{ij})}. \quad (2)$$

To accommodate the left truncation in the data we included years at risk, defined as age at censoring or age minus the age of the individual in 1958 (the truncation date) into the model as one of the covariates  $X$ . While this procedure does not adjust for the left truncation completely, it is an appropriate adjustment if the distribution of ages among case relatives is similar to the age distribution among control relatives. This assumption is reasonable, because the probands were matched

on age. As in the survival methods above, we also included gender as a covariate  $X$ . The null hypothesis of no aggregation again corresponds to  $H_0 : \gamma = 0$ . The model parameters were estimated under a working independence covariance structure to make the odds ratio estimates from the GEE model comparable to the hazard ratio estimates from the marginal survival model. The analysis was performed with Proc Genmod, SAS 8.0. The 95% confidence interval on  $\gamma$  was  $\hat{\gamma} \pm 1.96\widehat{SE}(\hat{\gamma})$ , where  $\widehat{SE}(\hat{\gamma})$  denotes the square root of the sandwich estimate of  $\text{var}(\hat{\gamma})$ , and two-sided p values were computed using asymptotic normal theory.

### Simulation Study

We used simulations to study two aspects of the procedures described previously. We assessed the efficiency of the survival approach compared to the binary model, both with sandwich estimates of variance, and examined how strong familial correlation has to be in order to induce strong aggregation, measured by the hazard ratio associated with proband status.

To create dependency within a family we simulated multivariate survival data from a shared frailty model. This random family frailty multiplies the hazard function for members of a family, resulting in individuals who share the same baseline hazard within a family, but have different baseline hazards from members of different families, inducing intrafamilial correlation. Letting  $W_i$  denote the frailty term for the  $i$ th family, the conditional hazard function of individual  $j$  in the  $i$ th family given  $W_i$  was

$$\lambda_{ij}(t_{ij} | X_{ij}) = W_i \lambda_0(t_{ij}) \exp(\beta X_{ij}). \tag{3}$$

We assumed that  $\lambda_0(t_{ij}) = \lambda_0$ , a constant, and  $\beta = 0$ . Then the survival times for the  $j$ th member of the  $i$ th family followed an exponential distribution with hazard  $W_i \lambda_0$ .  $W_i$  had a gamma distribution  $F(W; \alpha)$  with a mean of 1 and variance  $1/\alpha$ . For all simulations, the independent competing risks of mortality were modelled by a Weibull distribution,  $S(t; \lambda, \rho) = \exp(-\lambda t)^\rho$ , with a median of 70 and a standard deviation of 10, which correspond to the scale parameter  $\lambda = 0.014$  and shape parameter  $\rho = 8.21$ . Accounting for such competing risks, we obtain the cumulative ab-

solute risk to age  $t$  as

$$F(t) = \int_0^t \int_0^\infty \gamma \lambda_0 \exp(-\lambda_0 \gamma x) dF(\gamma; \alpha) S(x; \lambda, \rho) dx.$$

We simulated a population of 10000 families each with six family members, all assumed to be related in the first degree, for various values of parameters  $\alpha$  and  $\lambda_0$ . We chose all the cases from the simulated population, and an equal number of unmatched controls. Note that the model does not include covariates and the match was random; thus sandwich estimates of the variances are valid. Using the relatives of each proband with appropriate replication for multiple probands in a family, we fitted the marginal survival model (1), and the binary model (2) with proband status  $Z$  as the only covariate. It is easy to see (for example by using moment generating functions) that no familial correlation, i.e.  $1/\alpha = 0$ , will result in estimates  $\gamma = 0$  in (1) and (2). We thus expect  $\gamma$  to decrease as  $\alpha$  gets larger.

To compare the power of the two procedures, we used the McNemar test of the difference in the percentage of rejections of the null hypothesis of no familial aggregation,  $H_0 : \gamma = 0$ .

To illustrate the importance of accounting for correlations among family members in the variance calculations, we performed one simulation with 1000 repetitions, in which we computed the standard deviation of the log relative risk estimate, firstly by treating the data as independent and secondly by using the robust sandwich estimate.

## Results

### Simulations

We present data on the power, mean log relative risk and log odds ratio estimates based on independent simulation studies with 1000 replicates, each for various choices of baseline hazard  $\lambda_0$  and  $\alpha$  (Table 1). As mentioned previously, small values of  $\alpha$  correspond to large variation in frailty and large intrafamilial correlation. For fixed  $\lambda_0$ , the mean log relative risk estimate increases from about 0.16 for  $\alpha = 5$  to 1.1 for  $\alpha = 0.5$ , and power increases as  $\alpha$  decreases accordingly. For the frailty to induce a two-fold proband effect on the hazard of

**Table 1** Mean log relative risk and log relative odds in 1000 simulations (mean standard error in parenthesis) together with number of rejections of  $H_0 : \gamma = 0$

$\lambda_0$	$\alpha$	$F(70)^a$	Survival Model		Binary Model	
			log relative risk	# rejected	log odds ratio	# rejected
0.000143	5.003	0.00	0.160 (0.392)	169	0.161 (0.323)	165
0.000143	2.00	0.09	0.394 (0.248)	380	0.397 (0.253)	376
0.000143	1.5	0.09	0.512 (0.243)	566	0.515 (0.248)	559*
0.000143	1.00	0.09	0.693 (0.238)	801	0.697 (0.242)	799
0.000143	0.5	0.08	1.125 (0.229)	995	1.135 (0.234)	995
0.0004	6.0	0.00	0.155 (0.093)	404	0.158 (0.097)	387*
0.0004	3.0	0.23	0.285 (0.091)	827	0.289 (0.095)	817*
0.0004	2.0	0.22	0.414 (0.090)	986	0.421 (0.095)	984
0.0004	0.5	0.19	1.124 (0.089)	1000	1.132 (0.084)	1000
0.00073	6.0	0.38	0.156 (0.052)	799	0.161 (0.056)	769*
0.00073	5.0	0.37	0.189 (0.055)	913	0.194 (0.056)	892*
0.00073	2.0	0.35	0.415 (0.0514)	1000	0.428 (0.056)	1000
0.00073	1.0	0.33	0.714 (0.051)	1000	0.740 (0.056)	1000
0.00073	0.5	0.29	1.142 (0.059)	1000	1.188 (0.070)	1000

<sup>a</sup> cumulative absolute risk to age 70

\* significantly different number of rejections between the models ( $p \leq 0.01$  for McNemar test)

disease in relatives,  $\alpha$  must be less than 1.0, corresponding to  $Var(W) \geq 1.0$  (Table 1).

The other factor that affects power and the precision of the log hazard ratio estimates is the number of uncensored events ( $\delta_{ij} = 1$ ). The value of  $\lambda_0 = 0.000143$ , which corresponds to a low cumulative absolute risk up to age 70, is associated with lower power and wider average standard errors of the log relative risk estimate than are the values  $\lambda_0 = 0.0004$  and  $\lambda_0 = 0.00073$  (Table 1).

Very similar results (Table 1) are observed for the GEE analysis of binary outcomes from model (2). The mean log odds ratio estimates are very slightly larger than the corresponding mean log hazard ratio estimates from model (1), however. This result is not surprising, because the odds ratio from a single  $2 \times 2$  table exceeds the corresponding relative risk. The power of the analysis of binary outcomes was less than that from analyzing the survival data in every case, and the difference in power was statistically significant based on a McNemar test for five parameter settings (Table 1):  $\lambda_0 = 0.00073$  with  $\alpha = 6.0$  and  $\alpha = 5.0$ ,  $\lambda_0 = 0.0004$  with  $\alpha = 6.0$  and  $\alpha = 3.0$ , and for  $\lambda_0 = 0.000143$  and  $\alpha = 1.5$ . We also repeated one of the simulations with 5000 replicates to assess the robustness of our findings. For  $\lambda_0 = 0.0004$  with  $\alpha = 3.0$ , the power of the survival method was 85.17%, and for the GEE the power was statistically significantly lower at 83.63%. The corresponding mean log hazard ratio and mean odds ratio estimates were 0.293

(0.0913) and 0.298 (0.0955), respectively. Not surprisingly, the survival method is more powerful for more common outcomes with small relative hazards, while the powers of the two methods were similar for rare outcomes.

For  $\lambda_0 = 0.0004$  with  $\alpha = 3.0$ , the mean standard deviation of the log relative risk estimate for 1000 repetitions based on the sandwich estimate was 0.0912 (0.082), while it was 0.0898 (0.080) assuming independent observations. Though the difference is small, it is sufficient to raise the nominal test size from 0.050 to 0.054. A paired t-test demonstrated that the difference between these estimates of standard deviations was statistically significant with  $p < 0.0001$ .

### Hodgkin lymphoma (HL) and Non-Hodgkin Lymphoma (NHL)

We studied familial aggregation of HL by assessing if first degree relatives of probands diagnosed with HL are at 1) an increased risk of HL, and 2) at an increased risk of NHL. A more detailed analysis of these data is the subject of another paper (Goldin *et al.* 2004). The data consisted of a total of 15799 first degree relatives of 5047 HL probands, and 32117 first degree relatives of 10078 control probands. The probands were born between 1897 and 1994. 59% of all HL cases were male. Spouses of probands were not included in the

**Table 2** HL and NHL cases among first degree relatives of HL case and control probands (proportions given below the counts)

Diagnosed with	relative of control proband	relative of case proband
HL	18/32117 0.06%	32/15799 0.20%
NHL	70/32117 0.22%	46/15799 0.29%

analysis. Table 2 shows the numbers of HL and NHL cases among first degree relatives of case and control probands. Crude relative risks comparing case to control proband are  $0.20/0.06 = 3.7$  for HL and  $0.29/0.22 = 1.3$  for NHL (Table 2).

Eighty-three family members from control families and 86 from case families were duplicated in the data, because they were first degree relatives of more than one proband. Ninety-eight subjects were excluded from the subsequent survival analysis as they were less than a year old when they died. When not specified otherwise, the reported CIs are based on the “sandwich” variance estimate and asymptotic normality.

For HL, we estimated the relative hazard associated with proband status ‘case’, in a model that adjusted for gender, to be 3.50 with 95% CI (1.97, 6.22) (95% bootstrap CI (1.54, 5.22)), and the corresponding odds ratio from the GEE model of binary outcomes, also adjusted

for age, to be 3.74 (1.29, 7.36) (95% bootstrap CI (1.64, 5.44); Table 3). Women were at a significantly lower risk than men (Table 3). For NHL, the relative hazard estimate associated with proband status ‘case’ in the family was 1.33 (0.92, 1.94) (95% bootstrap CI (.96, 1.82)), and the corresponding odds ratio estimate was 1.31 (0.90, 1.91). Again, women were at a lower risk than men (Table 4).

The CIs for the relative hazards based on bootstrapping the matching clusters are 13% narrower for HL than those derived from the sandwich method that regards families as independent sampling units but ignores matching (Table 3). For NHL, the bootstrap CI was 16% narrower than the CI based on the sandwich method (Table 4).

The large relative hazards for HL indicate strong intrafamilial correlative of responses, and, if this were induced by a common frailty, the data in Table 1 suggest that the coefficient of variation would need to be greater than  $(0.5)^{-1/2} = 1.41$ .

To assess possible effects of the proband’s age-at-onset on the disease risk of the relatives, we plotted three hazards for the first degree relatives of probands for HL and NHL, respectively (Figures 2 and 3): one for the relatives of control probands, one for the relatives of early age-at-onset (< 41 years) case probands, and one for the relatives of late age-at-onset ( $\geq 41$  years) case probands.

Variable	Survival Model relative risk (95% CI)	GEE odds ratio (95% CI)
<b>Proband’s status</b>		
Control	1.00 (referent)	1.00 (referent)
Case <sup>1</sup>	3.50 (1.97, 6.22)	3.74 (1.29, 7.36)
Bootstrap CI <sup>2</sup>	(1.54, 5.22)	(1.64, 5.44)
<b>Gender</b>		
Male	1.00 (referent)	1.00 (referent)
Female	0.45 (0.25, 0.81)	0.50 (0.26, 0.99)
Bootstrap CI <sup>2</sup>	(0.24, 0.74)	(0.25, 0.78)
<b>Age</b>		
<39 years	NA	0.39 (0.18, 0.85)
Bootstrap CI <sup>2</sup>		(0.12, 0.42)
39–53 years	NA	0.10 (0.03, 0.34)
Bootstrap CI <sup>2</sup>		(0.06, 1.55)
53–66 years	NA	0.23 (0.08, 0.66)
Bootstrap CI <sup>2</sup>		(0.07, 0.53)
>66 years	NA	1.00 (referent)

<sup>1</sup>CI based on asymptotic normal approximation and sandwich estimate.

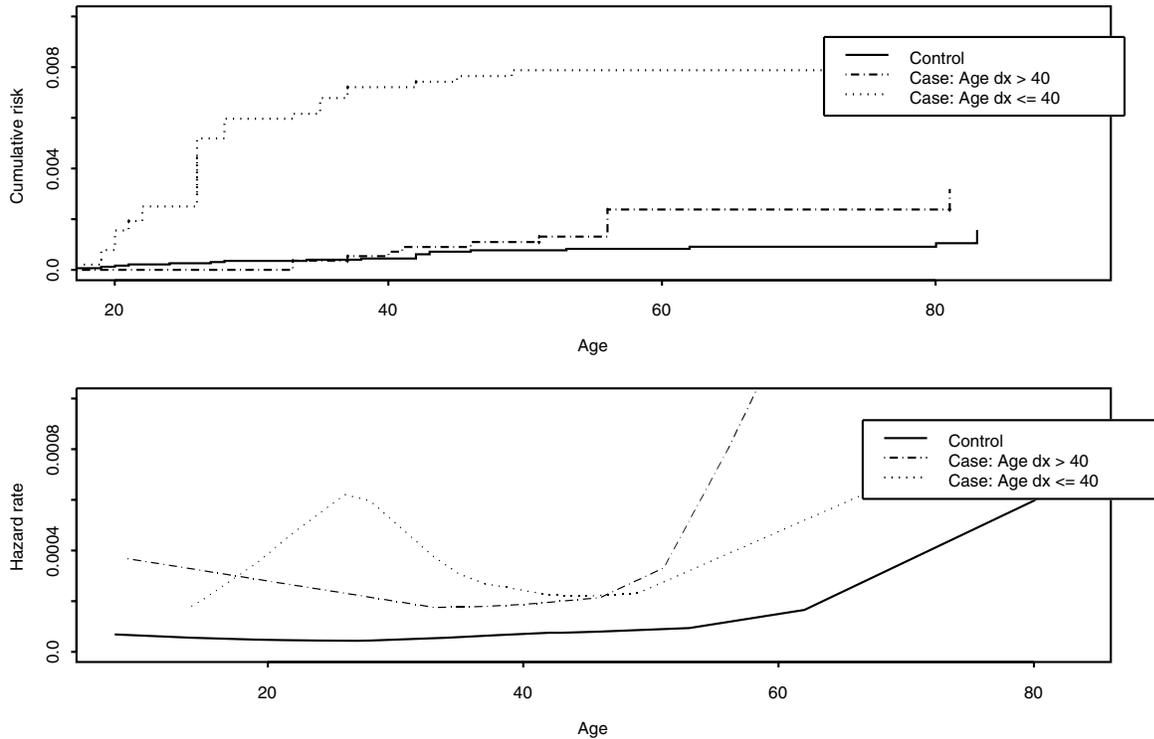
<sup>2</sup>CI based on empirical distribution function with  $B = 1000$  bootstrap replications.

**Table 3** Estimates for risk of HL among first degree relatives and 95% confidence intervals (CI)

**Table 4** Estimates for risk of NHL among first degree relatives and 95% confidence intervals (CI)

Variable	Survival Model relative risk (95% CI)	GEE odds ratio (95% CI)
<b>Proband's status</b>		
Control	1.00 (referent)	1.00 (referent)
Case <sup>1</sup>	1.33 (0.92, 1.94)	1.31 (0.90, 1.91)
Bootstrap CI <sup>2</sup>	(0.96, 1.82)	(0.91, 1.76)
<b>Gender</b>		
Male	1.00 (referent)	1.00 (referent)
Female	0.69 (0.48, 1.00)	0.70 (0.49, 1.02)
Bootstrap CI <sup>2</sup>	(0.56, 0.92)	(0.56, 0.94)
<b>Age</b>		
<39 years	NA	1.61 (0.41, 6.34)
Bootstrap CI <sup>2</sup>		(0.67, 5.21)
39–53 years	NA	4.21 (1.30, 13.69)
Bootstrap CI <sup>2</sup>		(2.18, 12.63)
53–66 years	NA	17.26 (5.42, 54.95)
Bootstrap CI <sup>2</sup>		(8.21, 28.74)
>66 years	NA	1.00 (referent)

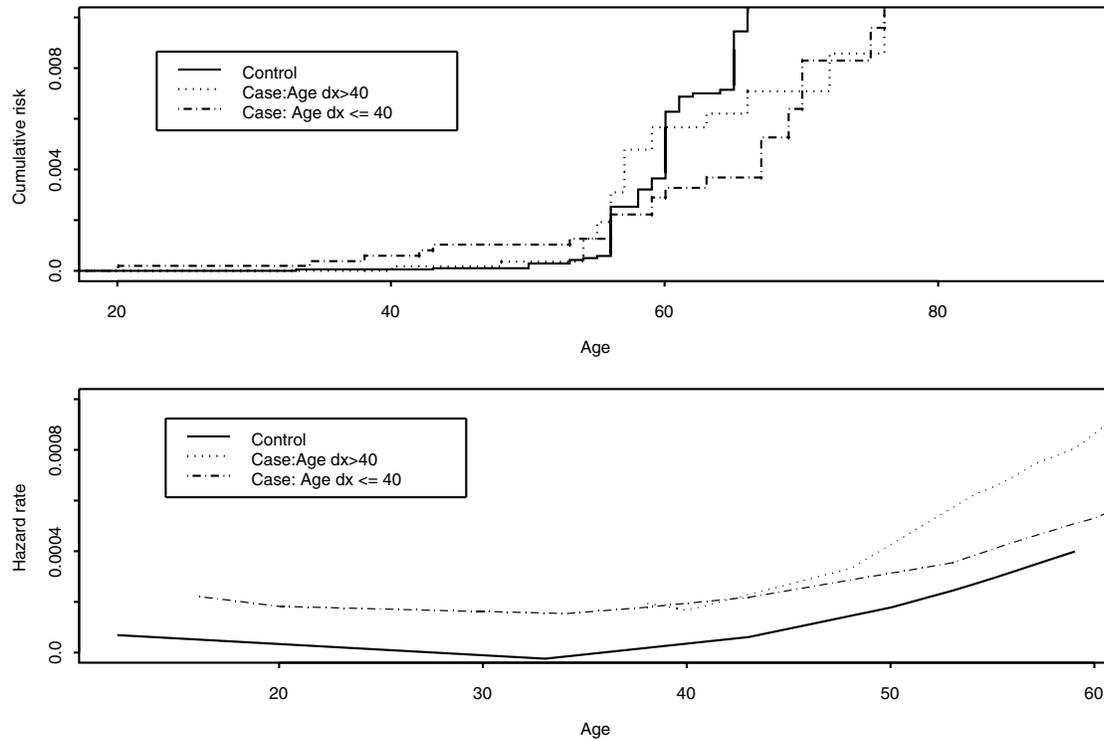
<sup>1</sup>CI based on asymptotic normal approximation and sandwich estimate.  
<sup>2</sup>CI based on empirical distribution function with  $B = 1000$  bootstrap replications.



**Figure 2** Plots of hazards and cumulative risk for Hodgkin Lymphoma for the first degree relatives of control probands, for first degree relatives of early age-at-onset (<41 years) case probands and for first degree relatives of late age-at-onset ( $\geq 41$  years) case probands.

The cutoff point of 41 for the proband's age-at-onset was chosen as it was the midpoint of the bimodal age distribution among HL probands in our population. While Figure 2 shows a strong difference in risk of HL for relatives

of early versus late age-at-onset probands, it also reveals that the proportional hazard assumption does not hold. In fact, the hazard plots for the relatives of case probands cross at around age 41. Thus the relatives of



**Figure 3** Plots of hazards and cumulative risk for Non-Hodgkin Lymphoma for the first degree relatives of control probands, for first degree relatives of early age-at-onset (<41 years) case probands, and for first degree relatives of late age-at-onset (>=41 years) case probands.

early age-at-onset cases have a higher risk early in life, while the relatives of late age-at-onset case probands have a higher risk later in life. For NHL a similar pattern is observed (Figure 3), with the hazards for the relatives of the late and early age-at-onset probands crossing at around age 41.

To capture the potential interaction between proband's age-at-onset and the age-specific hazard of the relatives, we considered a time-dependent proportional-hazards model that allows the effect of the proband's age-at-onset on the hazard of a relative to be different before and after the age of 41 years. We introduced a time-dependent exposure variable,  $E(t)$  so that  $E(t) = 1$  for  $t < 41$  and  $E(t) = 0$  for  $t \geq 41$ . Each relative whose exit-age was greater than 41 years was split into two separate observations, one corresponding to the follow-up from the entry-age to 41 years, and one corresponding to the follow-up from 41 years to the exit-age. The binary exposure  $A$  was defined as  $A = 1$  for proband's age-at-onset < 41, and  $A = 0$  otherwise, leading to four exposure groups among first degree relatives

of case probands, ( $A = 0, E = 0$ ), ( $A = 0, E = 1$ ), ( $A = 1, E = 0$ ) and ( $A = 1, E = 1$ ).

We then fitted model (1) with  $Z = 1$  if the proband was a case, and  $X = gender$

$$\lambda(t|X, A, E(t), Z) = \lambda_0(t) \exp(\gamma Z + \beta_0 ZA + \beta_1 ZE(t) + \beta_2 ZE(t)A + \beta_3) \tag{4}$$

Similar models have been used in studies of risk factors for early onset hypertension (Liang *et al.* 1990). To obtain variance estimates we employed the bootstrap procedure described in the statistical methods section.

Fitting model (4) to HL outcomes in the relatives, we obtained the following estimates (with 95% bootstrap CIs in parenthesis):  $\gamma = 1.28(-10.70, 1.84)$ ,  $\beta_0 = -0.54(-13.48, 0.00)$ ,  $\beta_1 = -0.55(-2.08, 10.72)$ , and  $\beta_2 = 1.33(-12.32, 2.76)$  and gender effect  $\beta_3 = -0.98(-1.67, 0.55)$ . The wide CIs reflect the small numbers of events among the relatives. Table 5 shows the corresponding hazard ratios for the four exposure groups compared to the hazard for the

**Table 5** Hazard ratios for HL and NHL stratified by age of onset of proband and relative

Age of relative	Age of proband	
	<41	≥ 41
	HL	
<41	4.55	2.05
≥ 41	2.09	3.58
	NHL	
<41	1.259	1.309
≥ 41	1.193	1.234

relatives of controls. For a relative of an early age-at-onset proband the hazard ratio is 4.55 before age 41, and 2.05 after age 41, compared to the relative of a control. For the relative of a late age-at-onset case, the hazard ratio is 2.09 before age 41, and 3.58 after age 41, compared to the relative of a control.

For NHL among the relatives, model (4) resulted in the estimates  $\gamma = 0.21(-0.41, 0.64)$ ,  $\beta_0 = 0.06(-0.01, 0.36)$ ,  $\beta_1 = -0.03(-0.70, 0.67)$ , and  $\beta_2 = 0.05(-0.24, 0.21)$ , and gender effect  $\beta_3 = -0.34(0.78, 1.24)$ . with corresponding hazards given in Table 5. There is little evidence of deviation from the proportional hazards assumption or that the age-at-onset of the HL proband affects the relative risk of NHL in the relatives.

## Discussion

We have studied methods to assess familial aggregation in registry data. The basic approach is to compare cancer occurrence among first degree relatives of case probands with cancer occurrence among first degree relatives of control probands. Our main analysis is based on a marginal survival model that incorporates the proband's disease status as a covariate into a proportional hazards model.

To account for dependencies in the data that arise from complete ascertainment of the cases, which means that more than one case proband can occur in a family, and for intra-familial correlations of times to disease onset, we present a bootstrap procedure based on resampling matching clusters, and a computationally simpler alternative that allows application of standard software by treating families as independent sampling units. The latter procedure is appropriate for unmatched data but

can lead to under or overestimation of the variance of the estimated relative hazard when case probands are matched to control probands.

In our data example from the Swedish Family Cancer Database, case probands were matched to control probands by age, gender and county of residence. The most important matching criterion was county of residency, which was chosen to allow for regional variability over time in reporting of cancers to the central registry. For both HL and NHL, the bootstrap confidence intervals were narrower than the confidence intervals based on the sandwich estimates of the variance. The narrower bootstrap confidence intervals reflect the fact that the bootstrap accounts for the matched sampling. Therefore, if matching has been used, we recommend the bootstrap procedure with resampling of matching clusters. For unmatched designs, regarding families as an independent sampling unit is appropriate.

In principle it is possible to account for the matching of case and control probands using a robust variance estimate, by summing over the score contributions corresponding to a matching cluster instead of the family. However, standard software cannot readily be used if one wishes to accurately account for various layers of dependence within the matching cluster. Similarly, the bootstrap can accommodate an unmatched case-control design by resampling families as an independent unit instead of matching clusters.

Our small simulation study showed that the log relative hazards estimated from the survival model are very similar to the log odds estimated from binary outcomes, though when the disease becomes more common, the log relative hazards are noticeably smaller than the log relative odds estimates (Table 1). In our example the coefficient of variation of the frailty needed to be 1.0 or more to induce a relative hazard of 2.0 or more, comparing risks associated with case and control probands. For the more common disease scenario with such relative hazards, the survival analysis tended to be slightly more powerful than the binary data analysis. Survival methods also control for a potential bias from unequal follow-up times for relatives of case versus control probands. The survival methods are thus preferable to the GEE approach for detecting familial aggregation.

Our simulations also illustrate that ignoring familial dependencies leads to small, but statistically significant, underestimation of  $\text{var}(\hat{\gamma})$ .

We applied the methods to a data set of first degree relatives of Hodgkin lymphoma probands and found a significantly increased risk of HL in relatives of case probands. Men had higher risk than women (Table 3). We also observed an increase in risk of NHL among relatives of HL probands, but this increase was not statistically significant.

Incorporating a time dependent indicator for age,  $E(t)$ , as well as an indicator for the age-at-onset of disease in the proband, and appropriate interaction terms, allows us to assess effects of age-at-onset of the proband and to study variation in risk according to the ages of the relatives. To account for dependencies between age-at-onset of the case proband and the relatives, Shih & Chatterjee (2002) used the Clayton model for survival outcomes in matched case-control family studies, and Hsu *et al.* (1999) derived a set of estimating equations by establishing a connection between the cross ratio function and the relative risk function, in a stratified proportional hazards model. Our analyses indicate that the relative risk of HL is higher among early age-at-onset HL cases, and that the relative risk declines in older relatives of early age-at-onset probands. For later age-at-onset probands, the relative risk among relatives is higher in older (age  $\geq 41$ ) rather than younger relatives (Table 5). No such deviation from proportional hazard was found for NHL (Table 5), but further analyses are planned to examine risks of NHL in relatives of NHL probands.

## References

- Goldgar, D. E., Easton, D. F., Cannonalbright, L. A., Skolnick, M. H. (1994) Systematic population-based assessment of cancer risk in first degree relatives of cancer probands. *Journal of the National Cancer Institute*, **86** (21): 1600–1608.
- Goldin, L. R., Pfeiffer, R. M., Gridley, G., Gail, M. H., Li, X., Mellekjaer, L., Olsen, J., Hemminki, K., Linet, M. (2004) Genetic Etiology of Hodgkin Lymphoma. *Cancer*, **100**, 1902–1908.
- Hemminki, K., Li, X., Plna, L., Granstrom, C., Vahtinen, P. (2001a) The Nation-wide Swedish Family-Cancer Database. *Acta Oncol* **40** (6): 772–777.
- Hemminki, K., Li, X., Mutanen, P. (2001b) Familial risks in invasive and in situ cervical cancer by histological type. *Eur J Cancer Prev*, **10** (1): 83–89.
- Hemminki, K. & Li, X. J. (2001c). Increased cancer risk in the offspring of women with colorectal carcinoma - A Swedish register-based cohort study. *Cancer* **91** (5): 1061–1063.
- Hsu, L., Prentice, R. L., Zhao, L. P., Fan, J. J. (1999) On dependence estimation using correlated failure time data from case-control family studies. *Biometrika*, **86** (4): 743–753.
- Liang, K. Y., Self, S. G., Liu, X. H. (1990) The Cox proportional hazards model with change point an epidemiologic application. *Biometrics*, **46** (3): 783–793.
- Liang, K. Y. & Beaty, T. H. (1991) Measuring familial aggregation by using odds-ratio regression models. *Genet Epidemiol* **8**(6):361–70.
- Liang, K. Y. (1991) Estimating effects of probands' characteristics on familial risk: I. Adjustment for censoring and correlated ages at onset. *Genet Epidemiol*, **8**(5):329–38.
- Oakes, D. (1989) Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**, 487–493.
- Shih, J. H. & Chatterjee, N. (2002). Analysis of survival data from case-control family studies *Biometrics*, **58**, 3.
- Therneau, T. M. & Hamilton, S. A. (1997) rhDNase as an example of recurrent event analysis. *Stat in Med* **16**, 2029–2047.
- Wei, L. J., Lin, D. Y., Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Statist Assoc* **84** (408): 1065–1073.

Received: 12 September 2003

Accepted: 26 February 2004