European Journal of Human Genetics (2008) 16, 1111–1116

www.nature.com/ejhg

ARTICLE

Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies

Jeroen B van der Net[1,2], A Cecile JW Janssens[1], Marinus JC Eijkemans[1], John JP Kastelein[3], Eric JG Sijbrands[2] and Ewout W Steyerberg*[,1]

[1]*Department of Public Health, Erasmus MC – University Medical Center Rotterdam, Rotterdam, The Netherlands;*
[2]*Department of Internal Medicine, Erasmus MC – University Medical Center Rotterdam, Rotterdam, The Netherlands;*
[3]*Department of Vascular Medicine, Academic Medical Center, Amsterdam, The Netherlands*

Cross-sectional genetic association studies can be analyzed using Cox proportional hazards models with age as time scale, if age at onset of disease is known for the cases and age at data collection is known for the controls. We assessed to what degree and under what conditions Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association analyses. Analyses were conducted in an empirical study on the association of 65 polymorphisms and risk of coronary heart disease among 2400 familial hypercholesterolemia patients, and in a simulation study that considered various combinations of sample size, genotype frequency, and strength of association between the genotype and coronary heart disease. We applied Cox proportional hazards models and logistic regression models, and compared effect estimates (hazard ratios and odds ratios) and statistical power. In the empirical study, Cox proportional hazards models generally showed lower *P*-values for polymorphisms than logistic regression models. In the simulation study, Cox proportional hazards models had higher statistical power in all scenarios. Absolute differences in power did depend on the effect estimate, genotype frequency and sample size, and were most prominent for genotypes with minor effects. For example, when the genotype frequency was 30% in a sample with size $n = 2000$ individuals, the absolute differences were the largest for effect estimates between 1.1 and 1.5. In conclusion, Cox proportional hazards models can increase statistical power in cross-sectional genetic association studies, especially in the range of effect estimates that are expected for genetic associations in common diseases.
European Journal of Human Genetics (2008) **16,** 1111–1116; doi:10.1038/ejhg.2008.59; published online 2 April 2008

Keywords: Cox proportional hazards models; logistic regression models; cross-sectional; genetic association studies; coronary heart disease; familial hypercholesterolemia

Introduction

Epidemiologic association studies are often analyzed using logistic regression models or Cox proportional hazards models. The choice between the two models is primarily based on the design of the study. Logistic regression models are used in cross-sectional and case–control studies, whereas Cox proportional hazards models are usually applied to prospective studies that have a follow-up period

***Correspondence: Professor EW Steyerberg, Department of Public Health – Ae236, Erasmus MC – University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands.**
Tel: +31 10 7038470; Fax: +31 10 7038475;
E-mail: e.steyerberg@erasmusmc.nl
Received 11 October 2007; revised 8 February 2008; accepted 19 February 2008; published online 2 April 2008

during which the occurrence of events is observed.[1] If follow-up data are available, Cox proportional hazards models are the recommended models as they have more statistical power than logistic regression models.[2,3] This is due to the fact that the Cox proportional hazards models take account of the time until events occur.[4] However, these models have not been compared in cross-sectional genetic association studies.

Genetic association studies that do not have follow-up time are generally analyzed using logistic regression models. However, since genotype status does not change over time and also represents genotype status at birth, age at event can be considered as follow-up time. If the age at event is known, genetic association studies could be analyzed with Cox proportional hazards models, even in the absence of prospectively studying follow-up. In the literature, there are various examples of studies in which logistic regression models were used, where Cox proportional hazards models could have been applied.[5–7]

We aimed to compare the statistical power of Cox proportional hazards models with that of logistic regression models in cross-sectional genetic association studies. We hereto performed an empirical study on the risk of coronary heart disease (CHD) in patients with familial hypercholesterolemia (FH), and a simulation study in which we examined the conditions under which additional statistical power can be achieved.

Methods

Study population

Empirical study We analyzed a retrospective, multicentre cohort study of patients with heterozygous FH who were recruited from 27 lipid clinics in the Netherlands between 1989 and 2002. Details of the study design and the study population have been published previously.[8,9] In brief, lipid clinics in the Netherlands routinely submit DNA of suspected FH individuals to a central laboratory for low-density lipoprotein (LDL) receptor mutation analysis. A total of 2400 unrelated patients who fulfilled the internationally established FH diagnostic criteria[8] were randomly selected from this database. Data on CHD were collected from the medical records by using a standardized protocol.[9] CHD was defined as the presence of at least one of the following: (i) myocardial infarction, (ii) percutaneous coronary intervention or other invasive procedures, (iii) coronary artery bypass grafting, or (iv) angina pectoris. Forty-nine percent of the FH patients were men, and mean age at the last visit to the lipid clinic was 50 years (SD 13 years). In total, 693 (29%) FH patients had proven CHD: 466 (19%) patients had a verified CHD event before study entry, and 227 (10%) incident CHD cases were observed during follow-up (median follow-up time without CHD was 3.1 years).

A previous association study considered 65 polymorphisms located in candidate genes for cardiovascular disease in our FH population.[10] Three polymorphisms had only wild-type alleles in our population and were therefore excluded from the present analyses. All patients gave informed consent and the ethics institutional review board of each participating hospital approved the protocol.

Simulation study We constructed a population of FH patients with sex, age at the first visit, and age at the last visit randomly sampled from the empirical data set. We simulated genotype status (for a single hypothetical polymorphism), age at event, and CHD status .

Genotype status was randomly assigned according to specified genotype frequencies. Although we recognize that individuals have one of three genotypes, we simulated only an at-risk genotype ('carrier') and another with the referent or baseline risk ('non-carrier') in our primary analysis. These two genotypes can be interpreted as dominant and/or recessive models of inheritance. In a secondary analysis, we repeated the simulations in the more complex setting of three genotypes. This yielded virtually identical results (data not shown).

Age at event was randomly drawn from distributions of age-, sex-, and genotype-specific CHD incidence rates for patients with FH. These distributions were obtained in three steps. First, we fitted Weibull distributions on age-specific CHD incidence rates in the general Dutch population, for men and women separately. The incidence rates were obtained from the National Institute for Public Health and the Environment (RIVM) [http://www.rivm.nl/vtv/object_document/o1320n17964.html]. Second, these distributions were adjusted to fit the age-specific CHD incidence in the FH patients of the empirical study, resulting in a cumulative CHD incidence of 29%. Finally, separate distributions were constructed for carriers and non-carriers by changing the average hazard according to the strength of association of the risk genotype, and assuming proportional hazards. CHD status was considered present when the simulated age at event was lower than the age at the last visit, and considered absent when the simulated age at event was higher.

Statistical analysis

Cox proportional hazards models and logistic regression models were fitted in the empirical and simulated data sets. With the term 'effect estimate', we refer to hazard ratios in the Cox proportional hazards models and odds ratios in the logistic regression models. All analyses were adjusted for sex and the logistic regression models were additionally adjusted for age (as a linear term), which was age at event or age at the last visit to the lipid clinic in the case of 'no event'. For the Cox proportional hazards models, we used age as time variable, thereby assuming that follow-up time started at birth and ended at the date of the first occurrence

of established CHD, or at the last visit to the lipid clinic. In the empirical data set, we assumed that each polymorphic allele had an additive contribution to the log-hazard/log-odds scale (additive genetic mode of inheritance). We compared the effect estimates and the *P*-values of the two models and calculated Spearman's rank correlation coefficient for the *P*-values of the two models.

In the simulation study, we varied the size of the population ($n=500$; $n=2000$; $n=5000$), the frequency of the risk genotype (10; 30; 50; 70%), and the strength of association between the polymorphism and risk of CHD (hazard ratio 1.0–2.0 with increments of 0.1) in separate scenarios. The hazard ratio of 1.0 was also simulated to check if type 1 error rates of the two approaches were as simulated, namely 0.05, and this was confirmed for all scenarios. The simulated and observed hazard ratios were slightly lower than the observed age-adjusted odds ratios in all scenarios (data not shown). Prospective data collection is assumed when Cox proportional hazards models are used. In retrospective data collection, early cases could be missed. To investigate whether missing data might influence our results, we simulated scenarios in which all prevalent cases (with an event before the date of the first visit to the lipid clinic) were missed, thereby excluding the prevalent cases from the analysis and considering only incident cases. Each scenario was repeated 5000 times. The statistical power of the two models was defined as the percentage of statistically significant ($P<0.05$) associations between the polymorphism and CHD status found for that regression model in the 5000 repeated scenarios. The gain in statistical power with the Cox proportional hazards models was expressed as the absolute difference in power and as the potential reduction in required sample size that can be obtained when the most powerful model would have the same power as the least powerful model.[11] Percentage reduction in required sample size was calculated as $100-100\ (Z_2/Z_1)^2$, where Z_1 and Z_2 are the Wald statistics of the most and least powerful model, respectively. These measures are independent of the effect estimate, the α-value and sample size.[11] Therefore, we calculated the mean percentage reduction in required sample size for each genotype frequency. All statistical analyses and simulations were performed using the R statistical package (version 2.5.1).[12]

Results

Empirical study

Figure 1 shows the effect estimates and *P*-values for the association between 62 polymorphisms and the risk of CHD obtained by Cox proportional hazards and logistic regression analyses. The effect estimates tended to be more extreme for the logistic regression models than for the Cox proportional hazards models (Figure 1a). The rank correlation coefficient for *P*-values was 0.54. Logistic regression analyses showed statistical significance for two polymorphisms, whereas four polymorphisms were statistically significant using the Cox proportional hazards models.

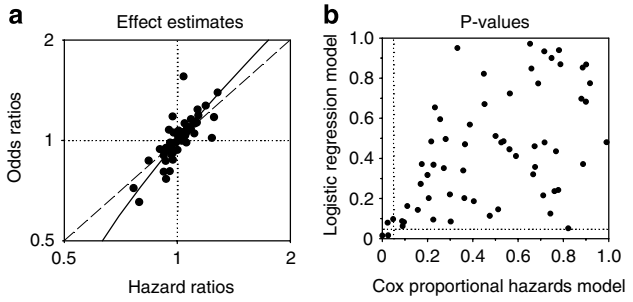

Figure 1 Effect estimates and *P*-values for 62 polymorphisms obtained by the Cox proportional hazards models and logistic regression models in the empirical study. Effect estimates are hazard ratios for the Cox proportional hazards models and odds ratios for the logistic regression models. Two outliers with effect estimates above 2.0 are not shown in these figures. (**a**) Dashed line represents the reference line for which the hazard ratio is equal to the odds ratio, solid line represents the linear regression line through the data points. (**b**) Dotted lines represent the significance threshold ($P=0.05$).

Simulation study

Figure 2 shows that Cox proportional hazards models had more statistical power than logistic regression models in all scenarios. The absolute difference in power was determined by the effect estimate, genotype frequency and sample size (Figure 3). The absolute difference in power was larger when sample size was low, the risk genotype was infrequent or the risk associated with the genotype was low. For example, when the genotype frequency was 30% and sample size was 2000, the absolute differences in power were most prominent for effect estimates between 1.1 and 1.5. The differences in power equaled to a reduction in required sample size ranging from 33% when the genotype frequency was 10%, to 18% when the genotype frequency was 70%. Risk sizes and reduction in sample sizes were similar when analyses were restricted to incident cases. However, the statistical power was lower in both models because of a lower number of events (data not shown).

Discussion

This study shows that Cox proportional hazards models may yield more statistical power than logistic regression models in cross-sectional genetic association studies. Differences in statistical power were most prominent for genotypes with minor effects in a range in which most genetic associations are expected.

The observation that Cox proportional hazards models have more statistical power than logistic regression models in association studies has been described previously.[2,3] For

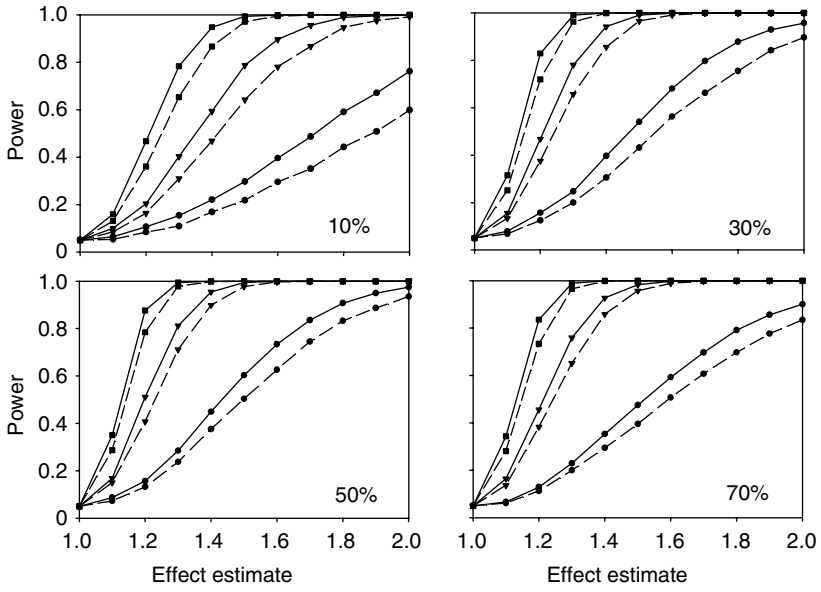

Figure 2 Statistical power of the Cox proportional hazards models and logistic regression models as a function of genotype frequency, effect estimate and sample size. Solid lines indicate power estimates of the Cox proportional hazards models and dashed lines of the logistic regression models. Power estimates are presented for sample sizes of 500 (●), 2000 (▼), and 5000 (■) patients.

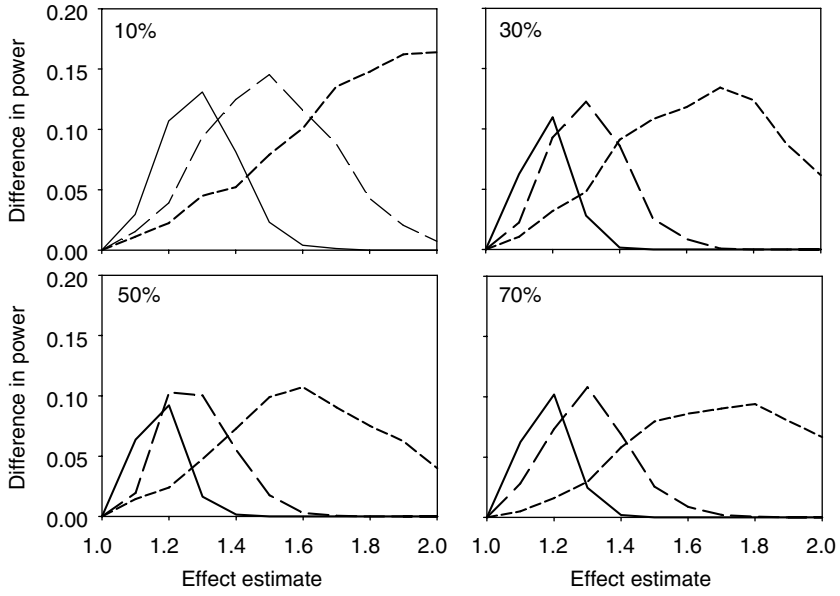

Figure 3 Absolute differences in statistical power between the Cox proportional hazards models and logistic regression models. Differences in power estimates were obtained by subtracting the power estimates of the logistic regression models from the estimates of the Cox proportional hazards models. Solid lines ($n = 5000$), long-dashed lines ($n = 2000$), small-dashed lines ($n = 500$).

example, the effect estimates will diverge when follow-up time is longer,[3,4] and effect estimates of logistic regression models are less precise, especially when the event is more common or when there is a strong relative risk.[13] This is in line with our findings in the empirical study, which showed that odds ratios tended to be more extreme than the hazard ratios. It has been previously shown that the Cox proportional hazards models give more conservative effect estimates than the logistic regression models, especially when the incidence of the disease is high,[2] as is the case in FH.

An explanation for the higher power of Cox proportional hazards models is that these models take into account the time until events occur, thereby changing the unit of

analysis from persons to person-years. Therefore, the interpretation of the results of the Cox proportional hazards models differs from those of the logistic regression model. While the logistic regression model tests whether a risk factor affects the odds of disease, the Cox proportional hazards model tests whether a risk factor affects the age of onset of the disease. Logistic regression models do not take into account the time until events occur, but give 'early' events and 'late' events the same weight in the analysis.[1,4] Young individuals who have had no event (yet) are classified as 'no event', while some would have experienced the event at an older age. This is a form of misclassification in terms of outcome. The superiority of the Cox proportional hazards models over the logistic regression models in analyzing longitudinal data has been mathematically proven for models, which consider one dichotomous covariate[2] and models with multiple covariates.[3]

A number of considerations regarding the generalizability of our results merit discussion. First, we simulated populations with a high risk of CHD, which implies that the absolute estimates of the statistical power of the different scenarios apply only to populations with similar disease risks. Statistical power and differences in statistical power were lower when the disease risks were lower, but still in favor of the Cox proportional hazards models (data not shown); a relative measure such as the potential percentage reduction in required sample size did not depend on the incidence of the disease. Second, analysis of a cross-sectional genetic association study with Cox proportional hazards models assumes that follow-up time starts with birth. In a retrospective design, Cox proportional hazards models only yield valid estimates compared to prospective studies, if there is no selective loss of follow-up. Our study only included patients who at least survived until a first visit to the lipid clinic, and early CHD cases could have been missed. Although these early cases might have been rare as demonstrated in a previous study,[14] we cannot exclude this possibility. We investigated the extreme scenario in which all prevalent cases were missed. This did not influence the effect estimates, because the analysis included all characteristics related to missingness of these cases (age, genotype status). Third, in our simulations, we did not adjust for covariables other than age. In cross-sectional genetic association studies, the use of Cox proportional hazards models is formally valid only when no adjustment is needed, or when adjustment is needed only for covariables that can be reliably assessed in retrospect, such as sex and education.

When there is no reason to expect selective loss of follow-up and no other variables than age and sex need to be adjusted for, Cox proportional hazards models are the preferred strategy for the analysis of genetic association studies. As the increase in power is independent of the type 1 error rate, Cox proportional hazards models may not only be preferred in association studies of candidate genes, but also in the statistical analysis of genome-wide association (GWA) studies, which generally consider lower type 1 error rates. Current GWA studies most often make simple and less powerful comparisons of genotype counts between cases and controls. Application of Cox proportional hazards models may lead to the additional identification of susceptibility genes with weaker effects that will remain undetected otherwise. In the case of cross-sectional genetic association studies in which adjustment for other variables than age and sex is needed, it is not immediately clear which is the model of choice. Additional variables, such as blood pressure and cholesterol levels, are more difficult to assess in retrospect. Ideally, these additional variables should be treated as time-dependent variables.[15] As the levels of these variables at the time of the event are often unknown in cross-sectional studies, the levels at the time of study conduction are frequently used as surrogates. Whether this will introduce a different size of bias in the two models is not clear. Yet, this potential bias is of lesser importance in gene-finding studies (as described in this study) than in risk prediction studies in which it is more important to accurately estimate the effect.[16]

We conclude that the advantage in terms of statistical power of the Cox proportional hazards models in comparison with the logistic regression models was most prominent for the range of effect estimates that are expected for most genetic associations. We recommend to consider the use of the Cox proportional hazards model in both cross-sectional genetic association studies and GWA studies.

References

1 Cox DR: Regression models and life tables (with discussion). *J R Stat Soc Ser B* 1972; **34**: 187–220.
2 Cuzick J: The efficiency of the proportions test and the logrank test for censored survival data. *Biometrics* 1982; **38**: 1033–1039.
3 Annesi I, Moreau T, Lellouch J: Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Stat Med* 1989; **8**: 1515–1521.
4 Green MS, Symons MJ: A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *J Chronic Dis* 1983; **36**: 715–723.
5 Gayther SA, Song H, Ramus SJ *et al*: Tagging single nucleotide polymorphisms in cell cycle control genes and susceptibility to invasive epithelial ovarian cancer. *Cancer Res* 2007; **67**: 3027–3035.
6 Jin Y, Mailloux CM, Gowan K *et al*: NALP1 in vitiligo-associated multiple autoimmune disease. *N Engl J Med* 2007; **356**: 1216–1225.
7 Raptis S, Mrkonjic M, Green RC *et al*: MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. *J Natl Cancer Inst* 2007; **99**: 463–474.
8 Jansen AC, van Aalst-Cohen ES, Tanck MW *et al*: The contribution of classical risk factors to cardiovascular disease in familial hypercholesterolaemia: data in 2400 patients. *J Intern Med* 2004; **256**: 482–490.
9 Jansen AC, van Aalst-Cohen ES, Hutten BA, Buller HR, Kastelein JJ, Prins MH: Guidelines were developed for data collection from medical records for use in retrospective analyses. *J Clin Epidemiol* 2005; **58**: 269–274.

10 Jansen AC, van Aalst-Cohen ES, Tanck MW *et al*: Genetic determinants of cardiovascular disease risk in familial hypercholesterolemia. *Arterioscler Thromb Vasc Biol* 2005; **25**: 1475–1481.
11 Hernandez AV, Steyerberg EW, Habbema JD: Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004; **57**: 454–460.
12 Ihaka R, Gentleman RR: A Language for Data Analysis and Graphics. *J Comput Graph Stat* 1996; **5**: 299–314.
13 Callas PW, Pastides H, Hosmer DW: Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *Am J Ind Med* 1998; **33**: 33–47.
14 Sijbrands EJ, Westendorp RG, Defesche JC, de Meier PH, Smelt AH, Kastelein JJ: Mortality over two centuries in large pedigree with familial hypercholesterolaemia: family tree mortality study. *BMJ* 2001; **322**: 1019–1023.
15 Kramer A, Jansen AC, van Aalst-Cohen ES, Tanck MW, Kastelein JJ, Zwinderman AH: Relative risk for cardiovascular atherosclerotic events after smoking cessation: 6–9 years excess risk in individuals with familial hypercholesterolemia. *BMC Public Health* 2006; **6**: 262.
16 Leffondre K, Abrahamowicz M, Siemiatycki J: Evaluation of Cox's model and logistic regression for matched case–control data with time-dependent covariates: a simulation study. *Stat Med* 2003; **22**: 3781–3794.