Simulation-based power calculations for large cohort studies

Patrick Brown[1,2], and Hedy Jiang[1]
1: Population Studies and Surveillance, Cancer Care Ontario
2: Dalla Lana School of Public Health, University of Toronto

June 28, 2010

Abstract

A large number of factors can affect the statistical power and bias of analyses of data from large cohort studies, including misclassification, correlated data, followup time, prevalence of the risk factor of interest, and prevalence of the outcome. This paper presents a method for simulating cohorts where individual's risk is correlated within communities, recruitment is staggered over time, and outcomes are observed after different followup periods. Covariates and outcomes are misclassified, and Cox proportional hazards models are fit with a community-level frailty term. The Cox proportional hazards model is shown to produce unbiased tests in the presence of the aforementioned factors, and the effect on study power of changing various variables is shown.
Keywords: power calculations; survival models; correlated data; large cohort studies; nested case-control studies.

1 Introduction

Increasingly large cohort studies are being undertaken by epidemiologists in order to identify genetic and environmental risk factors for chronic diseases. Kaplan (2007) describes the motivations and limitations of Very Big Epidemiology, and states that the dominant motivation in undertaking such studies is obtaining sufficient statistical power to detect sometimes modest main gene effects as well as gene-environment and gene-gene interactions. Kaplan (2007) finds Burton et al. (2008)'s result that it will take up to 40 years to collect sufficient cancer cases in the 500,000 strong UK Biobank to detect gene-environment interactions disquieting, and points out that there is considerable uncertainty in the calculations.

A conventional power analysis, computing the expected number of incident cases and applying a contingency-table based power formula (see Woodward, 1999, ch. 8), would disregard key elements of the inferential complexity that arise from various aspects of the design, conduct and analysis of large cohort studies. Burton et al. (2008) use a simulation-based method for power calculations which accounts for: misclassification of covariates and outcomes; individual level variation in risk; and the effect of prevalence of the genetic and environmental

risk factors. Logistic regression is then used to perform inference on the parameters and the detection rate based on multiple simulations recorded.

This paper builds on the work of Burton et al. (2008) by developing a data analysis and power calculation methodology for the Ontario Health Study (OHS), a large cohort study investigating cancer and other chronic diseases. Currently recruiting participants, the OHS has a number of particular features and requirements necessitating careful statistical consideration. First, study participants will be sampled within communities with community-level environmental effects and random variation being a key component of the research questions to be considered. Second, followup times will be irregular due to staggered recruitment and censoring of observations due to non-cancer related deaths. Third, due to the high cost associated with genotyping blood samples, nested case-control analyses are likely to be used for a number of sub-studies.

These requirements resulted in developing a simulation methodology where each simulated cohort involves assigning individual's age, sex, and community. Community-level random variation around observed cancer rates is used to simulate cancer incidence dates and non-cancer death rate data is used to assign dates of removal from the cohort. A Cox proportional hazards survival model is used for parameter inference, with left truncation due to age at recruitment and right censoring from the end of followup periods or earlier removal from competing events (see Lawless, 2003). A frailty term allows for dependence within communities, and a subsample of the cancer-free individuals are considered for the nested case-control analyses.

This paper describes and illustrates the proposed power calculation methodology, using results to demonstrate the statistical consequence of various assumptions and design issues. More epidemiologically-focused treatments of the assumptions and implications for study design will hopefully build on this research and follow in due course. Section 2 describes the simulation and inference methods in detail. Section 3 presents results for one of the most common cancers (colorectal or bowel cancer) and the much rarer cancer of the stomach for both cohort and nested case-control study. Section 4 discusses the implications of the results for statistical power calculations and more rigorous epidemiological applications of the methodology.

The software developed and used for this project as been released as an `R` package and is available at `r-forge.r-project.org/????`.

2 Methods

Each power calculation can be divided into three stages: defining the cohort; simulation of cancer incidence and non-cancer deaths; and model fitting. The total sample size was fixed at 150000 but can of course be varied. Population figures used are from the 2001 Census of Canada, and rates for cancer incidence and non-cancer deaths were taken from the Ontario Cancer Registry for the year 2004 (the most recent available at the time).

The parameters which need specifying *a priori* to simulate a cohort are: the total sample size N, the number of communities sampled M, the prevalences of the genetic and environmental risk factors π_1 and π_2, the misclassification rates for these risk factors θ_1 and θ_2, the community level random effect variation σ^2, the individual level random variation τ^2, and

the genetic, environmental, and gene-environment interaction effect sizes β_1, β_2, and β_3. The age and sex distribution of the population by region, baseline incidence rates, death rates for causes unrelated to the event of interest, the schedule for subject recruitment, and the age of the target population are also required. This gives 11 parameters to vary, in addition to which population and recruitment profiles, followup times, significance levels, and methods of allocating individuals to communities can also be changed.

2.1 Cohort Simulation

The first task is defining the community each individual in the cohort belongs to. For the purpose of this exercise a community was taken as one of 183 census sub-divisions in Ontario, which roughly correspond to administrative municipalities, containing at least 3000 individuals in the target age-range of 35 to 69. The number of communities in the study, M, was varied between 40 and 100 but equal to 50 unless otherwise specified. The number of participants N_i in community i is, depending on the design under consideration, either constant at $N_i = 150000/M$ for a balanced design or with N_i proportional to the community's population size for an unbalanced design. The sex and 5 year age group of each individual at the time of recruitment are simulated according to the population distribution within that community (aged 35 - 69) with replacement. A continuously valued age e_{ij} is simulated uniformly within the census age groups.

As recruitment into the study will be phased in over 4 years, the time of enrolment in the study is needed to allow for differing follow-up times for each individual. It is anticipated that 20000, 40000, 50000 and 40000 participants will be recruited by the end of year 1, 2, 3 and 4 respectively. Year of enrolment is simulated in proportion to the anticipated number recruited for that year, and time within the year simulated uniformly.

Individual j in community i is then ascribed a genetic risk factor X_{ij1}, and the community is given an environmental risk factor X_{i2}, which are binary-valued and present with probabilities π_1 and π_2 respectively. The covariates are then misclassified to produce observed covariates $\tilde{X}_{ij1}$ and $\tilde{X}_{i2}$ which are different from the true covariates with probabilities θ_1 and θ_2 respectively.

Incidence rates for the cancer of interest are calculated by dividing the total incidence count for the province by the total population of the group. Age groups are intervals of 5 years, with the exception of the oldest group being age 80 and above. The baseline rates $\hat{\lambda}_f(t)$ and $\hat{\lambda}_m(t)$ at age t for females and males respectively, are then used to compute individual rates after accounting for fixed and random effects. An individual's risk function $\lambda_{ij}(t)$ for individual j in region i is computed as:

$$\begin{aligned}
\lambda_{ij}(t) =& \lambda_{\cdot}(t) R_{ij} \\
\log(R_{ij}) =& X_{ij1}\beta_1 + X_{i2}\beta_2 + X_{ij1}X_{i2}\beta_3 + U_i + V_{ij} - (\sigma^2 + \tau^2)/2 \\
U_i \sim& \mathrm{N}(0, \sigma^2). \\
V_{ij} \sim& \mathrm{N}(0, \tau^2) \\
X_{ij1} \sim& \mathrm{Bern}(\pi_1) \\
X_{i2} \sim& \mathrm{Bern}(\pi_2),
\end{aligned}$$

with $\lambda_.$ being λ_f or λ_m depending on the sex of the individual concerned. Here β_1, β_2, and β_3 are the effect sizes of gene, environment, and interaction effects respectively. The U_i and V_{ij} terms are the community level and individual level random effects, allowing different cancer rates for each individual and some similarity between two individuals within the same community. Half the total variance $\sigma^2 + \tau^2$ is subtracted from the risk so that $E(\lambda_{ij}) = \lambda_.(t)$ for the baseline group, and this expectation would increase with the variance of the random effects were the subtraction not made.

2.2 Cancer incidence

Simulation of cancer incidence of a particular type (i.e. colon cancer) is accomplished by simulating three events in turn: non-cancer death dates; possible incidence of cancers other than the specific cancer of interest (i.e. cancer other than colon cancer) between recruitment and non-cancer death; and possible incidence the cancer of interest prior to other cancer incidence.

Non-cancer death figures for Ontario are calculated by subtracting the cancer deaths from the total deaths by age and sex group for 2004. A Weibull distribution is fit to the male and female data separately, giving scale and shape parameters $\hat{\psi}_m$, $\hat{\nu}_m$ for males and $\hat{\psi}_f$ and $\hat{\nu}_f$ for females. An individual's non-cancer death date d_{ij} is simulated as

$$D_{ij} \sim \text{Weibull}(\hat{\psi}_., \hat{\nu}_.)$$

with the parameters for males or females substituted according to the sex simulated for the participant. Conditioning on survival up to the age of enrolment is accomplished by re-drawing any death dates which occurred before the age of enrolment. A parametric Weibull is used instead of simulating from the empirical distribution of death data because of difficulties in estimating the tails at the highest ages where data are sparse.

The non-cancer death times can be interpreted as the age of death which each subject would have experienced were it not for cancer onset. Cancer incidence unrelated to the cancer site of interest and the age of onset (if any) is simulated based on the probability of contracting the cancer of interest between the age at recruitment and the age of non-cancer death. These dates are simulated using the population average rates for all other cancers by 5 year age and sex group, and not using the relative risk λ_{it}. The time of removal from the cohort, d_{ij}, is either the non-cancer death date or the "unrelated" cancer incidence data if present.

An individual's age at first cancer of interest Y_{ij} is simulated as a Poisson process in time with rate $\lambda_{ij}(t)$. Numerically, this was accomplished by simulating a number of potential first cancers N_{ij} between the age of enrolment e_{ij} and the 'removal' date d_{ij} as

$$N_{ij} \sim \text{Poisson}\left(\int_{e_{ij}}^{d_{ij}} \lambda_{ij}(u)du\right).$$

If $N_{ij} = 0$, then no cancer incidence is recorded, with $Y_{ij} = d_{ij}$ and the censoring variable is coded as $Z_{ij} = 1$. When N_{ij} is positive, N_{ij} events are independently simulated from a density proportional to $\lambda_{ij}(t)$ between e_{ij} and d_{ij}, Y_{ij} is set to earliest value drawn, and

$Z_{ij} = 0$. Care should be taken not to interpret the multiple events simulated when $N_{ij} \geq 2$ as multiple cancer incidences. The rates $\lambda_{ij}(t)$ are for first cancer incidence, and second and subsequent cancers can be different biological processes from the first. The procedure above is simply a mathematical solution to generating first incidence times.

Each cancer of interest was deemed to have been missed with probability ϕ_2 independent of age, and the removal date $d_{ij} reinstated. Each individual with no cancer event is given a false cancer with pro$ times the , with the age of the false cancer being uniform from the age of enrolment e_{ij} and death d_{ij}. Each individual with a cancer event has that event removed with probability ϕ_2 and their event date set to their death date $Y_{ij} = d_{ij}$ (and censoring variable $Z_{ij} = 1$). Gene and environment effects are misclassified (switched from a 0 to 1 or vice versa) with probabilities ψ_1 and ψ_2 respectively.

For each simulated (real or false) cancer event, a cancer site S_{ij} (i.e. colorectal, stomach, lung ...) was simulated with probabilities proportional to the number of cancers recorded in 2004 at that site, in the age and sex group of the individual concerned.

2.3 Parameter values

The results presented in Section 3 use a baseline set of parameter values and graph effect on power of changing one parameter at a time. The baseline parameter values are based on those used by Burton et al. (2008) and give roughly 70% power for detecting an effect on colon cancer after 20 years of followup, and are varied to show a range of low and high powers. Unless otherwise specified, the gene, environment, and gene-environment effects had relative risks of $\beta_1 = 1.5$, $\beta_2 = 1.5$, and $\beta_3 = 2$ respectively; and the prevalences of the individual level genetic and communities level environmental risk factors were $\pi_1 = 0.2$ and $\pi_2 = 0.3$ and ascribed to individuals rather than communities. Recruitment reflected a current proposal for the study, with an age range of 35 to 69 years; the design was unbalanced with 50 communities; the sample size is 150000; and the number of individuals recruited is 20000, 40000, 50000 and 40000 in each of years 1 to 4 of the study.

Motivated by Burton et al. (2008), we take $\sigma^2 + \tau^2 = 0.35$ resulting in the 5% most susceptible of the population having 10 times the relative cancer risk of the 5% least susceptible. The proportion of variation in risk attributable to the community level to individual level $\sigma^2/(\sigma^2 + \tau^2)$ is 20%, chosen as a moderate value between the close to zero power observed with 50% community variation and nearly 100% power resulting from independent samples demonstrated in Figure 4b.

The cancer false positive and missed cancer probabilities are $\phi_1 = 0.00045$ and $\phi_2 = 0.15$ respectively, which combined with the assumption that 2% of the subjects will contract cancer (of any type) throughout the course of the study gives a specificity of 97% as used by Burton et al. (2008). The misclassification rates for gene and environment risk factors are $\psi_1 = 0.1$ and $\psi_2 = 0.1$, which more optimistic than the 0.2 value used as a baseline in Burton et al. (2008). This greater confidence in the environmental measurement was chosen to reflect the relative ease in collecting environmental information for a small number of communities rather than at the individual level. The lower genetic misclassification was chosen to be roughly in line with past experience with simple presence/absence genetic traits.

The significance level of 0.001 was used as a compromise between the 0.05 or 0.01 level that would be the standard values for analyses of environmental effects and the much smaller

p-values of 10^{-4} or 10^{-7} often used in genetic studies as a result of multiple testing.

2.4 Data Analysis

For each simulated cohort, data were analysed at follow-up periods of 5, 10, 20 and 30 years. For a given cancer site S and followup time F, each individual's event time $\tilde{Y}_{ij}$ and censoring indicator $\tilde{Z}_{ij}$ are constructed as follows.

Event : The individual had cancer at the site in question within the followup period when $Z_{ij} = 0$, $S_{ij} = S$, and $Y_{ij} - e_{ij} \leq F$. This gives $\tilde{Y}_{ij} = Y_{ij}$ and $\tilde{Z}_{ij} = 0$.

Censored from competing event : An event within the followup period was unrelated to the outcome of interest effectively withdraws the subject when $Y_{ij} - e_{ij} \leq F$, and either $S_{ij} \neq S$ or $Z_{ij} = 1$. This gives $\tilde{Y}_{ij} = Y_{ij}$ and $\tilde{Z}_{ij} = 1$.

Censoring from the end of followup : When there was no event during the followup period ($Y_{ij} - e_{ij} > F$), we set $\tilde{Y}_{ij} = e_{ij} + F$ and $\tilde{Z}_{ij} = 1$.

A Cox proportional hazards model (see Lawless, 2003) is fit to the resulting data of left truncation times e_{ij}, event times $\tilde{Y}_{ij}$ and censoring indicators $\tilde{Z}_{ij}$. This model is similar to that used to simulate the data, with the difference that the baseline intensities $\lambda_{\cdot}(t)$ are not estimated but rather a profile likelihood which does not depend on $\lambda_{\cdot}(t)$ is used to estimate the other parameters. Here we assumed each individual's hazard function has the form

$$\begin{aligned} \lambda_{ij}(t) =& \lambda_{\cdot}(t) \exp\left(\mu + X'_{ij}\beta\right) U_i \\ U_i \sim& \text{Gamma}(1/\nu, \nu) \end{aligned}$$

Here U_i is a Gamma-distributed frailty term, allowing for similar hazard rates for individuals from the same community. Notice that an individual random effect term was not included, as there is only one observation per subject and individual-specific terms are not identifiable. The vector of covariates X_{ij} contains indicator variables for gene, environment, and their interaction.

For nested case-control analyses, a subset of the individuals not exhibiting the outcome of interest during the followup period was randomly selected with equal probabilities. The `coxph` function in `R` is used for model fitting and inference, using the default options (Gamma distributed frailty, EM estimation, zeros as initial values, with the Efron method for ties). Other inference methods could be considered (see Xu & Donohue, 2008), in particular for the case-control analysis (i.e. Gorfine et al., 2009), and the exact inference method might vary depending on the specific research questions considered in the future.

For each analysis, the p-values for each parameter are recorded. For each set of parameter values, 500 datasets are simulated and each dataset analysed at each followup period. The power for each β parameter is computed as the proportion of p-values below the significance level 0.001. In addition to simulating data with positive effect values for the β, the type 1 error rate was assessed by simulating data with no effect ($\beta = \mathbf{0}$) and the proportion of p-values below 0.001 as recorded. An unbiased inference procedure should have uniformly distributed p-values in this case and the proportion of significant p-values equal to the significance level.

The community-level fixed effect could not be estimated from a surprising number of simulations due to singular design matrices. The problem was most pronounced when the variance of the community-level random effect was large, the followup period was long, and the prevalence of the cancer of interest was high. An automated procedure for overcoming this problem, suitable for including in a loop of thousands of simulations, became necessary. When this problem occurred, 5% of the individuals without events in the followup period were removed randomly. If singular design matrices resulted from 10 different subsamples, then a nested case control design was used samling 8 controls per case. If 40 such analyses still produced singular matrices, equal numbers of cases and controls were used with the samling repeated until a non-singular design matrix resulted. The choice of 8 controls per case resulted from the results in Figure 5 showing that power with this number of controls is generally as high as power from the full cohort. Using the baseline parameter vaules, 100 simulations resulted in 5% resampling performed sucessfully 4 times, 8:1 case control design necessary and successful 10 times, and 1:1 case control designs used 6 times.

3 Results

The following figures illustrate the effect on power of variations in some of the model parameters. More extensive results are available from the authors on request.

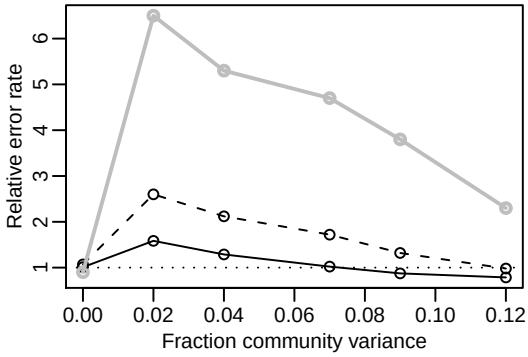
Years	**Colon**		**Stomach**	
	F	M	F	M
5	33 (20 , 45)	52 (37 , 71)	4 (0 , 8)	9 (4 , 16)
10	144 (109 , 185)	207 (158 , 257)	16 (8 , 24)	38 (26 , 52)
20	556 (455 , 678)	685 (540 , 817)	63 (44 , 82)	123 (94 , 155)
30	1095 (913 , 1299)	1221 (1004 , 1462)	125 (96 , 157)	221 (177 , 266)

Table 1: Average incidence numbers (with 95% prediction intervals) for male and female colon and lung cancer after various followup periods.

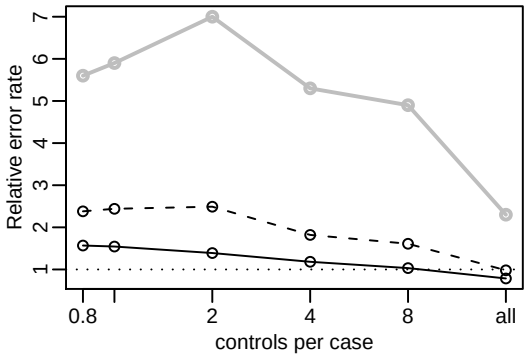
Table 1 shows the distribution of the number of colon and stomach cancer cases simulated at various followup periods, and the variations in power by cancer type and followup period should be interpreted in conjunction with these incidence numbers. The baseline parameter values were chosen to give roughly 70 % power for detecting a gene-environment interaction after 30 years of followup for colorectal cancer, which encompasses between 2000 and 2800 simulated cases. While colon cancer is one of the four most common cancers, stomach cancer is rarer and is included to illustrate the effect of disease incidence rates on study power.

3.1 Type 1 error rate

The following graphs show the type 1 error rate, obtained by simulating cohorts with no gene, environment, or interaction effects and fitting Cox proportional hazards models with such effects included. Rejecting the null hypothesis of no effect more often than the significance



(a) Full cohort, varying proportion of community variation

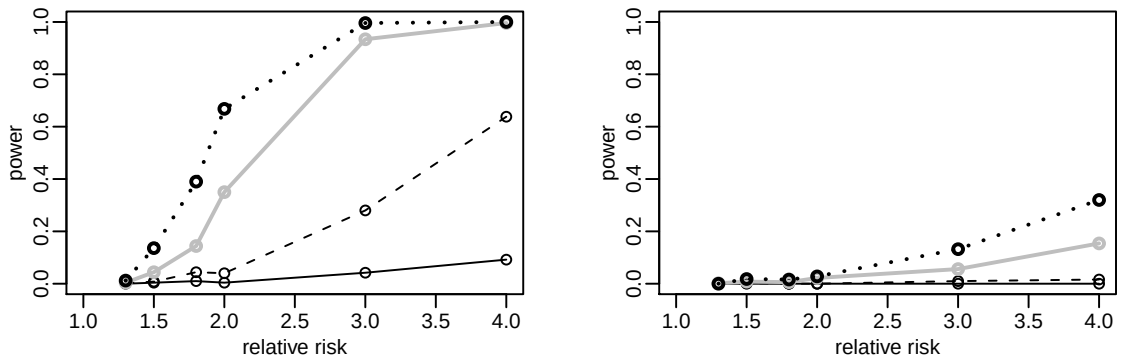


(b) Nested case-control study, varying number of controls

Figure 1: Type 1 error rate relative to the significance with levels of 0.05 (—), 0.01 (- - -), 0.001 (▬) for colorectal cancer after 30 years of followup.

level of 0.001 that is undesirable and indicates a high "false positive" possibility of finding effects which are non-existent. With the total number of communities fixed at 50, Figure 1a shows the type 1 error rate varying as the proportion of variation risk at the community level $\sigma^2/(\sigma^2 + \tau^2)$ increases. Significance tests for individual level genotype effects, group level environment effects, and interactions appear to be unbiased, with the type 1 error rate fluctuating around the value expected. To reveal the accuracy and efficiency of a nested case-control design, Figures 1b shows type 1 error rate as a function of different numbers of controls per case. Again, significance testing of all effects are unbiased, with the possible exception of the group-level environment with small numbers of controls.

3.2 Relative risk

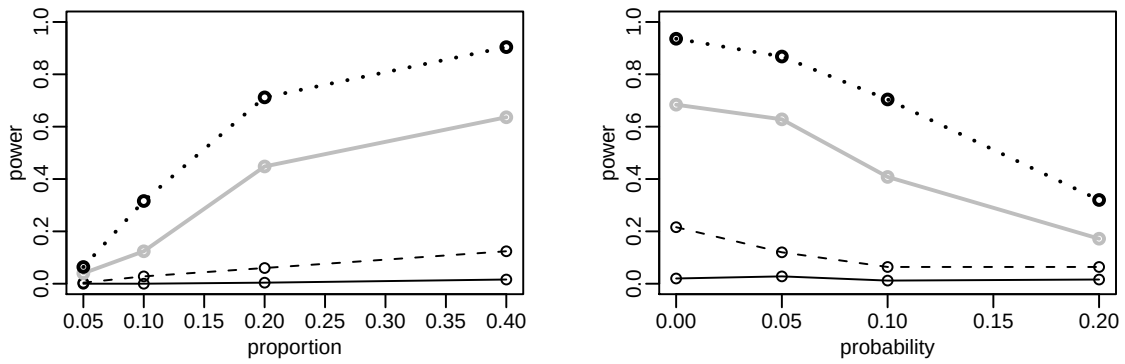


(a) Colorectal (b) Stomach

Figure 2: Power as a function of effect size for the gene-environment interaction effect, for stomach cancer and colorectal cancer. Follow up times of 5 (—), 10 (- - -), 20 (▬), 30 (⋯) years. Significance level is 0.001.

Figure 2 shows the power for detecting an individual-group interaction as a function of the effect size and followup time. Colorectal cancer, being one of the most common cancers, has much higher power than the much rarer stomach cancer. As expected, power increases with effect size and with followup time, with the biggest jump being between 10 and 20 years when relative risk is greater than 2.0. Effect size, as expected, has a large impact on power, though even large effect sizes don't appear to produce adequate power when the number of cases is small due to short followup (10 years or under) or low prevalence (stomach cancer).

3.3 Prevalence and Misclassification



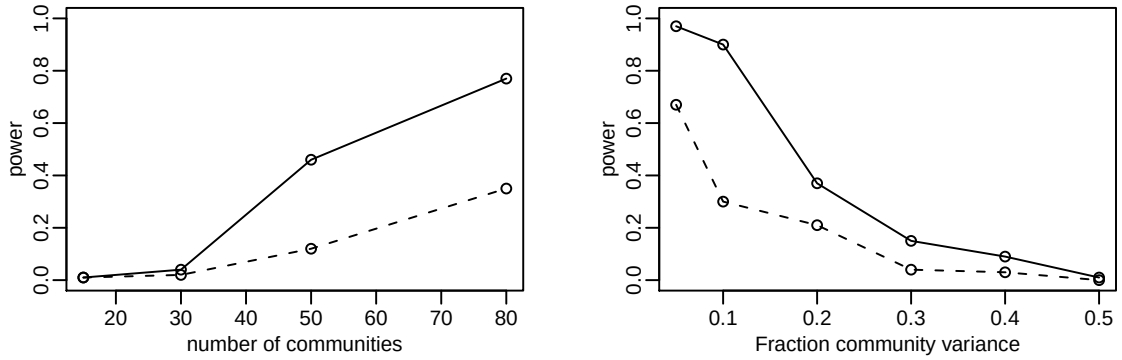
(a) Genotype prevalence

(b) Environmental misclassification rate

Figure 3: Power as a function of genotype prevalence and the probability of an individual's environment being misclassified, for the gene-environment interaction effect, for colorectal cancer. Follow up times of 5 (—), 10 (- - -), 20 (▬), 30 (⋯) years. Significance level is 0.001.

Greater prevalence of risk factors and lower misclassification of risk factors and cancers increase power substantially, and the result of varying some of these parameters are shown in Figure 3. Figure 3a shows power as a function of genotype prevalence for the gene-environment interaction effect on colorectal cancer. Again, followup times of 5 or 10 years show little power regardless of other parameter values. Power rises fairly steeply with prevalence, with power jumping from under 40% to over 70% after 30 years when prevalence increases from 10% to 20%. Figure 3b shows the power (colorectal, gene-environment) as a function of the probability of an individual's environment being misclassified. A misclassification above 10% results in low power regardless of the followup time.

3.4 Community effects and power



(a) Changing communities

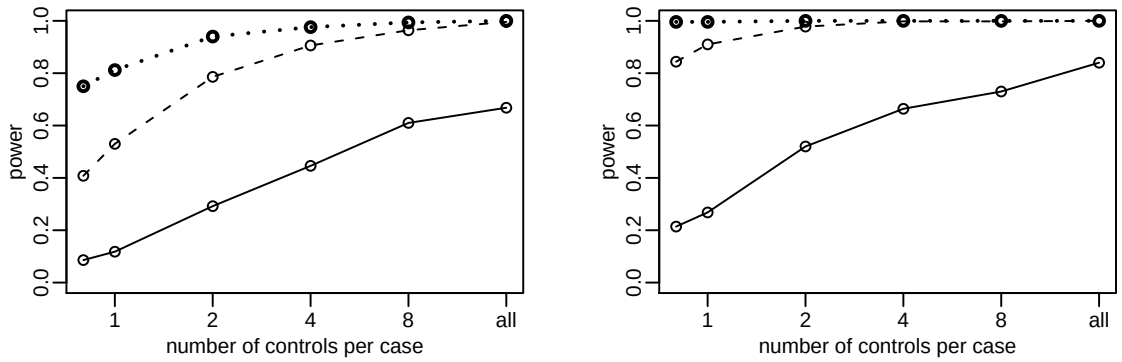
(b) Changing $\sigma^2/(\sigma^2+\tau^2)$

Figure 4: Power for detecting a community-level environmental effect on colorectal cancer after 30 years of followup, with a balanced (—) or unbalanced (- - -) allocation of individuals to communities, as a function of the number of communities. Significance level is 0.001.

Figures 4a and 4b show the power for detecting a community-level environmental effect on colorectal cancer as a function of the number of communities and the proportion of residual variation at the community level, or $\sigma^2/(\sigma^2+\tau^2)$. By allocating the environmental risk factor to communities, rather than to individuals, it would be expected that varying the number of communities or the effect of communities would have a bigger effect on the power of detecting a community-level effect than on an individual-level effect or on the interaction term. Two different sampling methods are compared: a balanced design with an equal number of subjects per community; and an unbalanced design with a random number of subjects per community (varying between simulations) and sampling probabilities proportional to each community's population.

Figure 4a shows that increasing the number of communities increases power for both balanced and unbalanced designs. Both designs perform similarly when the number of communities is small, with balanced sampling resulting in much higher power when the number of communities is large. Figure 4b keeps the total variation in risk $\sigma^2+\tau^2$ constant, but changes $\sigma^2/(\sigma^2+\tau^2)$, the proportion of the residual variation which is at the community level. High values of this proportion correspond to strong correlation between individuals in the same community. The dependence has a strong effect on the power, and as the number of communities is fixed at 50 the balanced design achieve higher power than the unbalanced one especially when the correlation between individuals in the same group is low.

3.5 Nested Case-Control



(a) Gene-environment interaction 2 (—), 3 (- - -), 4 (· · ·)

(b) Geneotype 1.5 (—), 2 (- - -), 3 (· · ·)

Figure 5: Power for detecting the genetic and the gene-environment interaction effect for colorectal cancer in a nested case-control study with various proportion of controls after 30 years of followup. Significance level is 0.001.

Figure 5 shows, for a nested case control study, the power for detecting the genetic and gene-environmental interaction effect as a function of the number of controls. The number of controls varies from 0.8 per case to 8 per case, with the power for the full cohort shown as the rightmost point. For the smallest effect sizes, power continues to increase with the number of cases, even beyond 4 controls per case. For the largest effect sizes, 4 controls per case are sufficient to achieve the same power as the full cohort does.

4 Discussion

This paper has demonstrated the effectiveness of a data analysis and power calculation methodology for large cohort studies with within-community dependence and varying ages and followup. The inclusion of a frailty term in a Cox proportional hazards survival model produced unbiased significance tests whereas, unsurprisingly, fitting a survival model without a frailty term was prone to detect significant effects when none existed (results not shown). Increasing dependence (Figure 4) or introducing misclassification of explanatory variables (Figure 3b) and outcome variables (not shown) reduce power but do not increase the Type 1 error rate.

Perhaps the contribution of most immediate practical use for the Ontario Health Study and other similar cohorts regards the effect of varying the number of controls in nested case-control studies. While the total sample size and number of communities in the OHS has necessarily been determined by financial and operational constraints, decisions on the amount and nature of genotyping to be done will be made at a later date and may vary between the various sub-studies. The number of controls required to obtain sufficient power, under a variety of assumptions and significance levels, can help determine future funding requirements and justify resource allocations.

While most of the power calculations presented here are intuitive (more cases, larger effects and higher risk factor prevalence result in higher power), there is room for more careful consideration of the subtleties involved in nested substudies. An upper bound for the power of a nested substudy is obtained by calculating the power of the full cohort, and Figure 5 shows that the benefit of adding additional controls is negligible in some circumstances, though a genotype effect of 1.5 shows continued benefit of adding controls beyond even the ratio of 8 to 1. Deliberately oversampling controls in the age groups where cases are more common is likely to result in higher power than the random selection of controls presented here, and one to one matching on age, community, and followup time could result in a yet more powerful logistic regression analysis. However, the more use is made of matching the greater the risk of bias due to unintentionally matching on the variable of interest (see i.e. Hein et al., 2009). Further, the inclusion of time-varying covariates (such as smoking status) complicates inference further (see Leffondre et al., 2003) and power and bias for such analyses should also be checked with simulations. Finally, nested case-cohort studies have not been considered here, and a comparison of case-control and case-cohort methods with dependent time-to-event data would be of further value in planning large cohort studies.

More generally, comprehensive power calculations allowing for a wide variety of factors, including community- or family-level dependence, will be a necessary component of interpreting the analyses of genotype and disease incidence data from large cohort studies. There are studies which have convincingly identified or replicated genetic and environmental association for a range of chronic diseases, though the conclusions are often inconsistent (see Rebbeck et al., 2007). In addition to possible scientific and technical issues which can cause spurious correlations, statistical inference methods which ignore dependence will underestimate parameter uncertainty and be prone to signaling artificial effects. Conversely, overly optimistic power calculations which ignore misclassification of outcomes and confounders might lead to researchers being too quick to dismiss the significant findings of previous studies which are not reproduced. The methods presented here are able to partially address the Statistical aspects of the reproducibility debate by:

- Demonstrating the unbiasedness of tests using the Cox proportional hazards survival model and providing a method for researchers to re-check the unbiasedness of this inference methods using values for prevalence, misclassification, and dependence which are felt to reflect their particular circumstances; and

- Yielding power estimates which reflect a wide variety of complications which can reduce the chance of detecting a risk factor, including staggered recruitment, within-community dependence, and misclassification.

Developing a more comprehensive power calculation has come at the expense of increasing the number of parameters which must be pre-specified. As such, it adds to (rather than addresses) the concerns expressed by Kaplan (2007) that such calculations are uncertain. The resulting model has 11 continuous parameters required for simulating cohorts, and although they are varied in this analysis a comprehensive treatment involving all possible combinations of 3 values for each parameter would prove infeasible in computational time and in interpretability. Finding sensible, scientifically sound values for all 11 parameters will

be difficult, and even with such knowledge sensitivity to assumptions about the parameter values should be explored.

Acknowledgements

This study was conducted with the support of the Ontario Institute for Cancer Research through funding provided by the Province of Ontario. Production of this paper has been made possible through a financial contribution from the Canadian Partnership Against Cancer and Health Canada. Patrick Brown thanks the Natural Sciences and Engineering Research Council of Canada for funding his Discovery Grant. Thanks are due to Peggy Sloan for helpful discussions.

References

Burton, P., Hansell, A., Fortier, I., Manolio, T., Khoury, M., Little, J., & Elliott, P. (2008). Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *International Journal of Epidemiology*.

Gorfine, M., Zucker, D., & Hsu, L. (2009). Case-Control Survival Analysis with a General Semiparametric Shared Frailty Model-a Pseudo Full Likelihood Approach. *Annals of statistics*, *37*(3), 1489.

Hein, M., Deddens, J., & Schubauer-Berigan, M. (2009). Bias from matching on age at death or censor in nested case-control studies. *Epidemiology*, *20*(3), 330.

Kaplan, G. (2007). How big is big enough for epidemiology? *Epidemiology*, *18*(1), 18.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley.

Leffondre, K., Abrahamowicz, M., & Siemiatycki, J. (2003). Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. *Statistics in Medicine*, *22*(24), 3781–3794.

Rebbeck, T., Khoury, M., & Potter, J. (2007). Genetic association studies of cancer: where do we go from here? *Cancer Epidemiology Biomarkers & Prevention*, *16*(5), 864.

Woodward, M. (1999). *Epidemiology: study design and data analysis*. CRC Press.

Xu, R. & Donohue, M. (2008). Proportional hazards mixed-effects models and applications. In R. Khattree & D. N. Naik (Eds.), *Computational Methods in Biomedical Research* (pp. 297–322). Taylor & Francis.