METHODOLOGY

Conventional case–cohort design and analysis for studies of interaction

John Cologne,1* Dale L Preston,2 Kazue Imai,3 Munechika Misumi,1 Kengo Yoshida,3 Tomonori Hayashi3 and Kei Nakachi4

1Department of Statistics, Radiation Effects Research Foundation, Hiroshima, Japan, 2Hirosoft International Corporation, Seattle, WA, USA, 3Department of Radiobiology/Molecular Epidemiology, Radiation Effects Research Foundation, Hiroshima, Japan and 4Consultant, Radiation Effects Research Foundation, Hiroshima, Japan

*Corresponding author. Department of Statistics, Radiation Effect Research Foundation, 5-2 Hijiyama Park, Minami-ku, Hiroshima 732-0815 Japan. E-mail: jcologne@rerf.jp

Accepted 1 June 2012

Background The case–cohort study design has received significant methodological attention in the statistical and epidemiological literature but has not been used as widely as other cohort-based sampling designs, such as the nested case–control design. Despite its efficiency and practicality for a wide range of epidemiological study purposes, researchers may not yet be aware of the fact that the design can be analysed using standard software with only minor adjustments. Furthermore, although the large number of options for design and analysis of case–cohort studies may be daunting, they can be reduced to a few simple recommendations.

Methods We review conventional methods for the design and analysis of case–cohort studies and describe empirical comparisons based on a study of radiation, gene polymorphisms and cancer in the Japanese atomic bomb survivor cohort.

Results Stratified, as opposed to simple, random subcohort selection is recommended, especially for studies of gene–environment interaction, which are notorious for lacking statistical power. Methods based on the score-unbiased exact pseudo-likelihood (or its analogue with stratified case–cohort data) are recommended for use in conjunction with the asymptotic variance estimator.

Conclusions We present an example of how to implement case–cohort analysis methods using SPSS, a popular statistical package that lacks some of the features necessary to directly adapt and implement published methods based on other software platforms. We also illustrate case–control analysis using Epicure, which provides greater risk-modelling flexibility than other software. Our conclusions and recommendations should help investigators to better understand and apply the case–cohort design in epidemiological research.

Keywords case–cohort study, pseudo-likelihood, stratification, statistical interaction, gene–environment interaction


Published by Oxford University Press on behalf of the International Epidemiological Association © The Author 2012; all rights reserved. Advance Access publication 18 July 2012

Downloaded from https://academic.oup.com/ije/article-abstract/41/4/1174/687507 by guest on 22 February 2019
Introduction

A recent explosion in molecular genetic methods facilitates detailed investigation of mechanisms of radiation-related cancer, which is one of the most well-documented outcomes of radiation exposure in the Japanese atomic bomb survivors. Using stored leukocytes obtained from survivors who attended clinical examinations in the Adult Health Study conducted at the Radiation Effects Research Foundation, studies are being conducted to investigate various components of the immunogenome and cancers at several sites. The standard approach to assessing risk is through the effect of radiation exposure on rates of disease or death using cohort follow-up. Analysis may be based on the proportional hazards model fit via the partial likelihood (PL). Extensions to the proportional hazards model involving more general risk functions other than proportional hazards, such as excess relative risk (ERR) or excess absolute rates, are often used.

Cohort studies involving costly or time-consuming methods to obtain covariate information, such as interviews, use of precious biological specimens, collecting hospital or employment records, or genotyping, can be carried out efficiently using sampling from the cohort non-cases because the cases are more influential on risk estimates and typically represent a relatively small fraction of the total cohort. Selecting subjects from the cohort using the counter-matched nested case–control study design can be effective, especially when studying interactions. However, with separate studies for multiple outcomes (e.g. multiple sites of cancer to be studied individually) selecting control subjects separately for each outcome wastes resources and can be inefficient. An appealing alternative is the case–cohort design, in which one selects a single subcohort from the initial cohort at its inception with a pre-specified sampling fraction, either randomly or using stratified random sampling, and adds in all cases that occur in the cohort outside the subcohort. Several discussions of the relative merits of case–cohort and nested case–control designs are available and both designs have been shown to be cost-effective for gene–disease association studies. In this review, we address case–cohort study design and analysis focusing, in particular, on exposure-based stratified random sampling from the cohort as a means to increase statistical efficiency relative to simple random sampling.

Because the subcohort is a random or stratified random sample from the cohort, it can be analysed by itself as a cohort study using the ordinary PL, but this is clearly inefficient because of the loss of cases that occur outside the subcohort. Thus, the key feature of the case–cohort design is the inclusion of all cases that occur in the cohort regardless of whether they are in the selected subcohort. Because of such biased sampling with regard to case status, risk estimation using the ordinary PL is not appropriate, one reason being that asymptotic distribution theory for the PL estimators breaks down because cases outside the subcohort induce non-nesting of the so-called sigma fields, or sets on which the counting process probability measure is defined. However, that problem can be overcome using straightforward adjustments. Most importantly, naive estimates of parameter standard errors output by ordinary proportional hazards regression programmes will be incorrect because of the statistical dependencies between elements of the estimating equations that inflate the variance, but those can also be handled in a straightforward manner using influence residuals available in most regression programmes. The upshot of this is that, although the synthetic nature of a case–cohort design may seem baffling to researchers, there is no impediment to obtaining appropriate estimates of risk and standard errors as long as appropriate methods of analysis are applied, and there are methods for performing such analyses using standard software.

Conventional options for analysing randomly sampled case–cohort data include the exact pseudo-likelihood (PsL) and approximate PsL, with various adjustments to parameter standard error estimates. In addition, more recent approaches based on considering the case–cohort sample as a cohort study with missing data or with alternative weighting or as the second phase in a two-phase design have been proposed. A generalized approach to analysing various designs involving sampling from a cohort has also been proposed. Although the conventional approaches have been available for some time and have been studied in some detail, the more recent approaches still require further evaluation and comparison among themselves and with the more conventional methods. Therefore, we believe it is possible to draw generalizations and conclusions regarding choice of design and method among conventional approaches, and that is the focus of this review.

The exact PsL is score unbiased and has been shown to have the best small-sample properties in simulation studies. However, there was little difference among the various PsL estimators when subcohort size was not small (>15% sampling fraction relative to the full cohort). Procedures for adjusting the standard errors of estimated parameters include the asymptotic method (based on the pseudo-score, analogous to likelihood-based methods) and the robust (empirical or ‘sandwich’) types of estimators. Both methods use delta–beta statistics (influence residuals) that can be obtained from most Cox regression software, and they are asymptotically equivalent. However, the robust variance estimates can be conservative (too large) in finite samples.

If there are risk factors whose values are known in the entire cohort, an alternative approach is to construct the subcohort using stratified sampling. In that case, the methods appropriate for randomly selected subcohorts require further modification to achieve...
unbiased parameter estimates and correct variance estimates. Borgan et al. 29 described three estimation methods, analogous to those for randomly sampled subcohorts. Their method III, known as the ‘swapper’ method because a case not in the subcohort swaps places with a member of the subcohort in the same sampling stratum, is the analogue of the exact PsL and is the only score-unbiased procedure among the three. Score unbiased means that the estimating equations have expected value zero; otherwise they can result in finite-sample bias even though they are statistically consistent (unbiased asymptotically). Although the computational complexity involved in analysing data from a stratified case-cohort study is not great, 18 it nevertheless requires some effort, and it is not clear what magnitude of efficiency gains may be achieved by stratification to guide investigators in choosing between simple and stratified subcohort selection. In addition, the score-unbiased swapper method for stratified data does not seem to be used, despite mention of a stratified design in some published works, 27 and it is not implemented in the case-cohort function ‘cch’ 30 in the R package ‘survival’. 31

The present report describes the conventional simple and stratified random-sampling approaches to case-cohort study design and compares them using actual data from a study of epithelial growth factor receptor (EGFR) gene repeat length polymorphism and radiation joint effects on lung cancer in atomic bomb survivors. 3 We provide simple recommendations as to which methods of analysis should be used, and we present examples of how to implement the methods using Epicure and SPSS, which are popular statistical packages among epidemiologists. We also provide S-plus code for implementing the swapper method for stratified case-cohort data, which makes critical use of a built-in function not available in R.

**Review of conventional case-cohort methodology**

Early work focused on estimation of relative risk based on the analogy of a case-control comparison. 32–34 Subsequent analytical methods capitalized on the analogy with cohort follow-up and PL, although the unique sampling approach (over-selecting cases) leads to a PsL rather than the usual PL. The PsL may be used to fit the analogue of a PL analysis (e.g. the proportional hazards model). Prentice noted that parametric functions other than log-linear risk functions could be used, and the PsL can be extended to accommodate absolute rate differences and other non-proportional hazard models for estimating the hazard rate ratios for covariates. 11 However, with the exception of Epicure, 6 such models are not generally available in case-cohort analysis software.

Analyses of the case-cohort sample must adjust for bias introduced in the distributions of covariates used in calculating the denominator of the PsL, as the case-cohort sample is not population-based because of over-sampling of cases. Bias incurred by including cases outside the subcohort is corrected by not allowing those cases to contribute to risk sets other than their own (risk sets comprise cohort subjects still under observation and at risk at the times cases occur). Two approaches exist based on whether cases outside the subcohort are allowed to enter the denominators of the PsL for their own risk sets. With Prentice’s original method, 11 cases outside the subcohort appear in their own risk set denominators; this method is called the exact PsL. The method on which the asymptotic variance estimate was originally based 17 excludes cases not in the subcohort even from the denominators of their own risk sets; this method is called the approximate PsL. These inclusions or exclusions can be carried out by weighting the cases. With the exact PsL, cases outside the subcohort are given Weight 1 in their own risk set and Weight 0 otherwise; with the approximate PsL, cases outside the subcohort are given Weight 0 in all risk sets, including their own, although those cases still contribute to the numerators of the PsL in their own risk sets by virtue of being cases (defining the risk sets). In practice, this means that weighting must be accomplished by the use of offsets rather than prior weights because the latter will exclude non-subcohort cases altogether. The exact PsL is score unbiased, whereas the approximate PsL is not score unbiased and can result in some small-sample bias, particularly if the subcohort risk sets are small. 18

In addition to case weighting, all subjects can be weighted using risk set weights based on follow-up time and cohort attrition: who among the subcohort subjects are still at risk as time passes. This type of weighting, proposed by Barlow, uses the inverse of the sampled fraction as a weight. 35 The original proposal updates the weights at each failure time (case occurrence or risk set) using time-dependent weights according to the surviving fraction in the sampled subcohort vis-à-vis the surviving fraction in the full cohort. This approximates the likelihood function denominator contribution that would pertain in the full cohort analysis and is intuitive based on counting process theory. 36

Barlow’s robust variance estimate, a jackknife estimate related to the influence function, was shown to have good performance and said to be easier to calculate than either the bootstrap estimate 37 or the asymptotic variance estimate, the latter requiring calculation of the correlations among components of the scores (estimating equations). An advantage of the robust variance vis-à-vis the asymptotic variance is that the latter is difficult to compute in open cohorts where there is staggered entry. 15 Practical computation was via SAS using the PHREG procedure.
with non-subcohort cases defined to enter the study just before their failure times and variances derived from the delta–beta residuals \(^\text{15}\) (originally using martingale residuals \(^\text{13}\)). Similar computations may be done in S-plus \(^\text{36}\) or R. If the data are partitioned into risk sets (if there are time-dependent covariates, for example), the subject-specific residuals must be summed up separately for each subject over all risk sets in which the subject appears.

A simplification of the Barlow weighting approach is to use the single, original subcohort sampling fraction at all failure times (i.e. weights are fixed, not depending on risk set). This has the practical advantage that multiple records (one record for each subject for each risk set in which she or he is at risk) do not have to be constructed, as the number of records in the analysis data set can significantly increase if there are many risk sets (many cases of disease or death). As with the exact PsL estimator, the Barlow method weights each case with Weight 1 at its failure time whether or not it is in the subcohort. Therneau and Li suggested that cases in the subcohort be weighted by the appropriate risk set fraction just as the non-cases, rather than converting their weight to 1 at their time of failure, to reflect the fact that they are legitimate risk set members. \(^\text{19}\)

Details of variance estimation are described in a number of publications. \(^\text{17,19,35,38}\) The asymptotic variance consists of within-risk set variances and between-risk set covariances, the latter arising from correlations among score contributions (components of the estimating equations). Briefly, the problem is solved by adding to the usual PL parameter covariance matrix a weighted product of influence residuals for the subcohort data. One approach uses weights based on the subcohort sampling fraction. \(^\text{29}\) Another uses as weights the proportion of cases actually sampled in the subcohort. \(^\text{19}\) With a large amount of data, the two types of weights will be similar because the expected fraction of cases in the subcohort approaches the subcohort sampling fraction as the subcohort size increases. However, due to random sampling variation, the actual proportion of cases in the subcohort may deviate from the expected value. Essentially all of the so-called ‘robust’ variance estimators that have been proposed can be expressed as special cases of a general influence function variance estimator that is robust to misspecification of the model relating the risk parameters to the underlying distribution function. \(^\text{38}\)

All of the aforementioned methods consistently estimate the risk and variance asymptotically (with large studies), but they differ in their efficiency (estimates of the parameter variances) and small- or finite-sample bias (in the case of the approximate PsL). Therefore, a major concern is how well they perform in studies of the size typically encountered in epidemiological practice. According to Onland-Moret et al., \(^\text{26}\) the ‘most simple approach’—Prentice’s original, exact PsL method—provides estimates and standard errors closest to the full-cohort results (i.e. using all subjects with no case–cohort sampling) when the cohort is small (about 1000 subjects or less) or when the subcohort is small (sampling fraction 0.15 or less with a cohort size of 1000 subjects). Those authors also demonstrated that the approximate PsL is inferior to the exact PsL method. However, they did not implement Barlow’s method as originally proposed: they used the initial sampling fraction rather than time-dependent weights. Langholz and Jiao \(^\text{18}\) reported that the approximate PsL estimator does not perform as well as the exact estimator when risk set sizes are small (e.g. when there are large numbers of cases and a small subcohort sampling fraction). The approximate PsL estimator is also not score unbiased in finite samples, \(^\text{29}\) which means that it can result in small-sample bias.

From a practical standpoint, either the exact or approximate PsL involves about the same amount of data manipulation, and both require constructing duplicate data records for subcohort cases to accommodate the case weighting described earlier. The approximate method requires one set of records for all subcohort members (including subcohort cases) having Weights 1, and a second set of records for all cases (including both subcohort and non-subcohort cases) having Weights 0. Weight 0 is achieved by including a large, negative offset, \(-100\) say, which when included in the proportional hazards model becomes \(\exp\{-100\}\) or a number essentially equal to zero to the level of machine precision. A weight of one is achieved by using an offset of zero (\(\exp(0) = 1\)). This means that cases not in the subcohort will define risk sets but will not enter into the denominators of the PsL calculations, whereas subcohort cases are included in all risk set denominators up to (in time) and including their own. This data setup was illustrated by Therneau and Li, who also described how to calculate the asymptotic variance estimates, using SAS and S-plus. \(^\text{19}\) Their S-plus routine calculates the variances directly, whereas their SAS routine requires the IML procedure and an additional calculation by hand.

With the exact PsL, non-subcohort cases should be included in the denominators of their own risk sets. Instead of calculating offsets to down weight the cases, it is easier to specify entry and exit times in such a way as to prevent non-subcohort cases from contributing to the overall subcohort follow-up while allowing them to contribute to their own risk sets. One record is constructed for each case with entry time set to a small increment before its onset time. The increment should be much smaller than the differences between successive failure times so as to avoid overlap of risk sets. Another record is constructed for each subcohort subject (including subcohort cases) with exit time set to a small increment before its onset or censoring time. This ensures that
subcohort cases contribute to all risk sets before their time of onset. Langholz and Jiao\textsuperscript{18} described SAS commands to set up the data for the exact PsL. They also described fitting the proportional hazards model and computing asymptotic and robust variance estimators in SAS using the same data setup, although (as with the Therneau and Li\textsuperscript{19} SAS variance estimate) the asymptotic variance estimate requires an additional calculation by hand.

The exact PsL with asymptotic variance estimate is also available in the Epicure survival analysis module ‘Peanuts’.\textsuperscript{6} In Epicure, the user only need specify, in addition to the usual PL variables (entry time, exit time and failure indicator), an indicator of which cases are outside the subcohort (CSCHRT) and the subcohort sampling fraction; thus, substantially less setup is required than with SAS or S-plus. Furthermore, Epicure allows fitting general risk models, including absolute rates; therefore, it is our method of choice for unstratified case–cohort analyses.

Up to this point, we have focused on case–cohort studies involving simple random selection of the subcohort. Stratification (stratified random sampling) can be used to increase efficiency (‘exposure stratification’) and to deal with confounding (‘confounder stratification’).\textsuperscript{18} In the present work, we are concerned with efficiency of risk estimation when exposure is known in the entire cohort and interest lies in making inference about interactions with exposure (exposure risk modification). Thus, we focus on exposure stratification. Stratified sampling in case–cohort studies merely involves randomly selecting subcohort subjects separately within exposure strata. As with all stratified sampling scenarios, the optimal number of strata and optimal numbers of persons to sample from each stratum are issues worth consideration,\textsuperscript{29} but we do not pursue them here as they are difficult to generalize.

Borgan et al.\textsuperscript{29} defined three methods for handling stratified case–cohort data. The first of their estimators (method I) is analogous to the approximate PsL estimator in the unstratified situation in that a case outside the subcohort does not enter into the denominator, even in its own risk set. The second estimator (method II) is somewhat analogous to the exact PsL, but non-cases are weighted according to their respective stratum proportions. The third estimator (method III) weights all subjects in the subcohort according to their respective stratum proportions and further deals with cases not in the subcohort by randomly selecting one member of the subcohort in the same stratum to be removed or ‘swapped’. The reason for this approach is that the exact PsL for stratified data remains score unbiased if non-subcohort cases are excluded from the denominators of their likelihood terms, but such exclusion is inefficient.\textsuperscript{29} Swappers may be chosen on a risk set basis if the analysis includes time-dependent covariates; otherwise, the swapper is selected from the initial subcohort strata, and if the randomly chosen swapper has been censored as of the time of the current risk set, swapping for the corresponding non-subcohort case is simply ignored.

The swapper method (method III) is our recommended method for stratified case–cohort data, as it is the only score-unbiased method. Unfortunately, it is not included among the procedures for stratified case–cohort data analysis in the cch function of the R survival package. We therefore adapted to S-plus the SAS macros of Langholz and Jiao.\textsuperscript{18} The script is provided in the Supplementary Appendix A1 available at IJE online, but cannot be directly implemented in R due to the lack of a necessary function—match.data.frame—that is included with S-plus, but not with R.

Mark and Katki\textsuperscript{38} stated that variance estimation with the stratified case–cohort design might be handled using their influence function variance estimator starting with a stratified proportional hazards model, but did not provide details except to suggest that one approach is to treat cases with missing information as if they occurred outside the subcohort. Additive models may also be entertained in the stratified case–cohort design\textsuperscript{39} as may frailty models for cluster data.\textsuperscript{40} Many other options and proposals exist in the literature, but they are too numerous to describe in detail. Table 1 presents an overview of the evolution of conventional methods for case–cohort study design and analysis. We conclude that score-unbiased methods with asymptotic variance are preferable based on both theoretical and small-sample properties. In particular, the robust variance cannot be implemented with stratified data, and risk-set (time-dependent) weights do not seem to be necessary with PsL estimators (although they may be related to efficiency with other types of estimators\textsuperscript{22}). We therefore recommend the use of the exact PsL (using its analogue, the swapper method, in the stratified case) and the asymptotic variance estimates. We restrict our attention to that approach in the remainder of this work.

**Empirical investigations**

The initial design of the immunogenome and cancer case–cohort study included two subcohorts: one using simple random sampling and one using stratified random sampling with stratification on radiation dose. Five strata were defined approximately as quintiles, but with round numbers as the cutpoints (Table 2). Both subcohorts were constructed using a sampling fraction of 0.5 to reduce genotyping effort by about one-half while retaining most of the full-cohort power.\textsuperscript{3} To minimize the effort involved in molecular analyses and to facilitate direct comparison between the two sampling procedures, the two subcohorts were selected with the greatest amount of overlap possible between them. This was done by modifying the simple random subcohort, randomly adding to or excluding from each stratum the
<table>
<thead>
<tr>
<th>Reference</th>
<th>General approach</th>
<th>Special characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prentice^{11}</td>
<td>Initial proposal of case-cohort design based on time-to-event analysis^{2}</td>
<td>Defined exact pseudolikelihood (PsL) analogous to partial likelihood (PL)</td>
</tr>
<tr>
<td>Self and Prentice^{17}</td>
<td>Asymptotic distribution of approximate PsL</td>
<td>Asymptotic normality of pseudolikelihood estimate and consistency of Prentice’s variance</td>
</tr>
<tr>
<td>Wacholder et al^{17}</td>
<td>Variance estimators derived under superpopulation (sampling without replacement) model or using the bootstrap</td>
<td>Superpopulation variance valid under $\beta = 0$, bootstrap method non-standard due to lack of knowledge of covariate information for all cohort members</td>
</tr>
<tr>
<td>Barlow^{35}</td>
<td>Robust variance estimator based on influence functions. Weights based on subcohort membership within risk sets</td>
<td>Analogous to Lin and Wei^{41} robust variance for the full-cohort proportional hazards (PH) model. Preserves correct expectation of PsL denominator in each risk set.</td>
</tr>
<tr>
<td>Lin and Ying^{20}</td>
<td>Missing-data approach to estimation with proportional hazards model</td>
<td>Variance estimators more readily derived with large data than those of Self/Prentice or Wacholder et al.</td>
</tr>
<tr>
<td>Barlow et al^{15}</td>
<td>Jackknife variance estimator computed via delta-beta residuals</td>
<td>SAS macros available through statlib</td>
</tr>
<tr>
<td>Therneau and Li^{19}</td>
<td>Intuitive description of fitting PH model with approximate pseudolikelihood and asymptotic variance estimation</td>
<td>SAS and S-plus examples provided</td>
</tr>
<tr>
<td>Chen and Lo^{21}</td>
<td>More efficient use of case and cohort information in the estimating functions</td>
<td>Derived three methods depending on extent of knowledge of case proportion in the population</td>
</tr>
<tr>
<td>Borgan et al^{29}</td>
<td>PsL estimators for stratified case-cohort data</td>
<td>Score-unbiased ‘swapper’ method analogous to unstratified exact PsL</td>
</tr>
<tr>
<td>Mark and Katki^{38}</td>
<td>Generalized influence function approach to variance computation</td>
<td>Dealing with missing data among subcohort or case subjects</td>
</tr>
<tr>
<td>Scheike and Martinussen^{23}</td>
<td>Maximum likelihood via EM algorithm (including non-sampled cohort members who do not fail)</td>
<td>Efficiency gain related to disease intensity. Variance estimated by profile likelihood.</td>
</tr>
<tr>
<td>Kulich and Lin^{22}</td>
<td>Doubly weighted estimation utilizing covariate information available on the full cohort</td>
<td>Efficiency gain greatest for continuous covariates, less for binary covariates</td>
</tr>
<tr>
<td>Nan^{42}</td>
<td>Estimation via efficient scores</td>
<td>Relaxes some assumptions underlying the proportional hazards model</td>
</tr>
<tr>
<td>Langholz and Jiao^{18}</td>
<td>Computational methods for random and stratified designs using exact PsL and asymptotic variance</td>
<td>SAS code provided for data construction, proportional hazards model fitting and variance estimation</td>
</tr>
<tr>
<td>Samuelsen et al^{28}</td>
<td>Intuitive methods analogous to those of Therneau and Li^{19} for the asymptotic variance</td>
<td>Script provided for asymptotic variance estimation using S-plus/R Compared asymptotic and robust variance estimators</td>
</tr>
<tr>
<td>Kulathinal et al^{27}</td>
<td>Design options and overview of analysis methods</td>
<td>SAS and R command examples</td>
</tr>
<tr>
<td>Moger et al^{40}</td>
<td>Case-cohort methods for cluster-based sampling</td>
<td>Extended pseudolikelihood approach to gamma frailty and copula failure models</td>
</tr>
</tbody>
</table>

---

^{2}Prentice^{11} attributes the earliest ideas similar to the case-cohort design to Kupper et al^{43} and Miettinen^{32}, but seems to be the first to describe the link to partial likelihood and use of continuous explanatory factors.
necessary number of individuals to obtain a stratified sample.

The first study performed using the immunogenome and cancer case–cohort design, a study of lung cancer and EGFR gene, was originally conducted using the simple random case–cohort sample.4 However, the EGFR gene was ultimately assessed on the full cohort (this would not generally be the case with a case–cohort study). Hence, the two designs (simple and stratified) can be directly compared and various sampling fractions (i.e. <50%) can be evaluated via Monte Carlo simulation drawing from the full cohort. We performed such simulations analysing the data using the exact PsL (or its analogue with stratified data, the swapper method) and asymptotic variance adjustment, the methods of choice noted earlier.

Relative efficiency (efficiency of the case–cohort sample relative to that of the full cohort) was calculated as the inverse ratio of standard errors. Analyses were conducted using S-plus (version 6.2 for Windows) or R (version 2.10.0 for Windows). Analyses of simple case–cohort data were made using the cch function in the R survival package with the default ‘Prentice’ option (exact PsL). Analyses of stratified case–cohort data were performed using method I or II in the R survival package cch function. For method III, we adapted the SAS programs described by Langholz and Jiao18 in S-plus using the coxph function (Supplementary Appendix A1 available at IJE online). Asymptotic standard errors were based on Cox regression delta–beta statistics [S function residuals.coxph(type = “dfbeta’’)]. Residuals for the non-failures were used and summed over risk sets within individuals using the aggregate function in S-plus. Age was used as the primary time scale in proportional hazards regression with left truncation at entry age.44 Covariates were as follows: city of residence, gender, year of birth (centred at the median, 1927, and divided by 5 years), smoking frequency (number of cigarettes smoked per day divided by 10), whole-body radiation dose (wGy: weighted Gray skin dose, using weights 10 for neutron dose and 1 for gamma dose), an indicator of EGFR gene CA repeat length < 38 (short-repeat genotype) and the product of radiation dose and EGFR CA repeat length indicator. Confidence intervals (95%) and tests were based on Wald statistics. Results obtained using SPSS and Epicure are provided along with relevant commands in the Supplementary Appendix A1 available at IJE online.

Results

Table 3 shows results of analyses of the actually selected simple and stratified 50% case–cohort designs. There was little practical difference between the two designs in terms of estimated parameters and standard errors, consistent with the expectation that a large case–cohort sample should provide close to full-cohort efficiency. Thus, the original analysis,4 which was based on the simple random subcohort with 50% sampling fraction, should be reasonably efficient. The exact PsL method with asymptotic variance for the random case–cohort sample was programmed in S-plus and gave results identical (to three decimal places) to those obtained with the R cch function (data not shown). The three methods for a stratified case–cohort sample produced nearly identical results. How the methods compare with smaller sample sizes is addressed below.

Proportional hazards imply multiplicative effects; therefore, the negative interaction suggests a submultiplicative joint effect of radiation and CA repeat length polymorphism. What this implies about biological interaction, however, cannot be assessed without examining other scales of the joint effect, such as a purely additive statistical model.45 An additive ERR model for the joint effects of radiation and repeat length polymorphism (which can only be fitted using Epicure with the simple case–cohort sample; see Supplementary Appendix A1 available at IJE online) produced virtually the same result as the multiplicative
proportional hazards model, except that the radiation dose response is linear rather than log linear. The ERR for radiation was 1.00/wGy (RR = 2.00 at 1 wGy) compared with the log-linear relative risk model RR = exp(0.396) = 1.49 at 1 wGy. With the multiplicative relative risk or the additive ERR model, the effect of the short CA repeat length polymorphism is to approximately double the cancer risk: the ERR for the short repeat length polymorphism was 1.06 (RR = 2.06) compared with the log-linear model RR = exp(0.601) = 1.82. With both models, statistical interaction between radiation dose and repeat length polymorphism essentially negated the radiation dose–response main effect. Interaction on the additive ERR scale was \(-0.40 (0.19)\) compared with the radiation main effect (Table 3). Thus, there is apparently no practical radiation effect among individuals with the short CA repeat length polymorphism.

Results of Monte Carlo sampling from the full cohort with sampling fractions varying from 0.5 down to 0.01 are shown in Figure 1 and Table 4. Relative efficiency decreased with decreasing sampling fraction. Relative efficiency of the simple case–cohort sample declined more dramatically than that of the exposure-stratified sample—that is, the stratified design becomes relatively more efficient with smaller subcohort fractions (Figure 1).

To compare by simulation the three methods for stratified data, we ran the S-plus swapper function in R by importing without modification the S-plus match.data.frame function. All three methods produced similar results with subcohort fractions down to \(~10\%\), but with sampling fraction of 5% or 1% method I had greater bias and larger standard error than methods II and III (Table 4). Overall, method III (the swapper method) performed the best, having the least bias and smallest standard errors, although differences between methods II and III were trivial.

### Table 3

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Case–cohort sampling method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random (exact method)</td>
</tr>
<tr>
<td>Year of birth (5-year difference; centred at 1927)</td>
<td>0.61 (0.12)</td>
</tr>
<tr>
<td>Gender</td>
<td>0.001</td>
</tr>
<tr>
<td>City (Nagasaki: Hiroshima)</td>
<td>0.01 (0.21)</td>
</tr>
<tr>
<td>Gender (female: male)</td>
<td>0.96</td>
</tr>
<tr>
<td>Gender</td>
<td>0.008</td>
</tr>
<tr>
<td>Number of cigarettes smoked per day (increment of 10)</td>
<td>0.45 (0.08)</td>
</tr>
<tr>
<td>Radiation dose (Gray)</td>
<td>0.40 (0.10)</td>
</tr>
<tr>
<td>EGFR repeat polymorphism (indicator of length &lt;38)</td>
<td>0.60 (0.24)</td>
</tr>
<tr>
<td>Interaction between radiation and EGFR polymorphism</td>
<td>0.012</td>
</tr>
<tr>
<td>P-value</td>
<td>0.036</td>
</tr>
</tbody>
</table>

*Although not statistically significant, city of residence was included to avoid possible confounding of the gene–radiation interaction.*
cohort data were stratified case–cohort data, using simulations with a fraction of the population. Borgan et al. recommended that the exact PsL be used. Alternatively, we recommend that the approximate PsL be used.

Little difference in the data preparation required for the two methods is asymptotically equivalent. Because there is an approximate PsL in small samples, even though the PsL is not yet available. It has been noted that choice of appropriate weights might be affected when specimen storage duration or analytical batch effects are a concern (i.e. for cases outside the subcohort in a prospectively conducted case–cohort study). This would not be a problem with retrospectively conducted case–cohort studies. In fact, although we are not aware that it has been proposed before, it should be possible to perform risk set subsampling of nested control subjects from the subcohort, if the timing of case ascertainment could lead to biases due to storage or batch effects.

Missing data are an important consideration in observational studies. Regardless of design, measurements made on biological specimens can be missing for a number of reasons. Mark and Katki suggested that, as long as missingness is not related to covariate values, case–cohort analyses without time-varying weights should remain valid if cases with missing covariate information are simply ignored.

One advantage of the nested case–control and case–cohort designs as compared with the traditional case–control study design is their ability to estimate absolute risk difference (as opposed to relative risk) and survival probability. However, when there is sampling of cases or missing case data, estimators of those quantities require modification to avoid bias. Another advantage of the cohort-based sampling designs is that the reference group can be used to study population characteristics, such as allele frequencies or Hardy–Weinberg equilibrium; the control subjects

### Table 4 Mean and average standard error (SE) of estimates of the radiation–gene interaction with stratified case–cohort data using simulation from the full cohort

<table>
<thead>
<tr>
<th>Subcohort fraction (%)</th>
<th>Method I Mean (SE)</th>
<th>Method II Mean (SE)</th>
<th>Method III Mean (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>−0.43 (0.18)</td>
<td>−0.43 (0.18)</td>
<td>−0.43 (0.18)</td>
</tr>
<tr>
<td>25</td>
<td>−0.43 (0.19)</td>
<td>−0.43 (0.19)</td>
<td>−0.43 (0.19)</td>
</tr>
<tr>
<td>10</td>
<td>−0.45 (0.22)</td>
<td>−0.45 (0.21)</td>
<td>−0.45 (0.21)</td>
</tr>
<tr>
<td>5</td>
<td>−0.47 (0.26)</td>
<td>−0.46 (0.25)</td>
<td>−0.45 (0.25)</td>
</tr>
<tr>
<td>1</td>
<td>−0.72 (0.64)</td>
<td>−0.48 (0.44)</td>
<td>−0.43 (0.41)</td>
</tr>
</tbody>
</table>

*Results are based on 2000 simulations at each subcohort fraction.

*For comparison, the estimate and standard error with the full cohort data were −0.42 and 0.17, respectively.

### Discussion and conclusions

The case–cohort design is not used as frequently as it might be given the advantages it offers epidemiologists, perhaps because the design is misunderstood given its synthetic nature. It is hoped that the present review will serve to heighten investigators’ awareness of the definition and advantages of the case–cohort design and provide guidance to facilitate its implementation and analysis. Issues that can be resolved in a fairly general manner include: whether to stratify, whether to perform exact or approximate PsL analysis, whether to use asymptotic or robust variance estimates and whether/how to weight the observations in the analysis.

Our results suggest that selecting the subcohort using stratified, rather than simple, random sampling can be beneficial in terms of statistical efficiency, especially for studying interaction, when the sampling fraction is substantially less than one-half. Although relatively small subcohort fractions are typical, subcohorts as large as 50% (such as in the immunogenome and cancer study described here) are not beyond imagination. Even such a large study reduces by almost half the amount of covariate collection effort yet retains efficiency close to that of the full cohort. If an investigator is able to conduct a study with large subcohort fraction, there may be little benefit from the stratified sampling. However, there is little additional effort involved in selecting and analysing a stratified subcohort; therefore, if there are factors for which it makes sense to stratify, we recommend doing so, regardless of the sampling fraction.

Simulation studies conducted by others have demonstrated that the exact PsL is preferable to the approximate PsL in small samples, even though the two are asymptotically equivalent. Because there is little difference in the data preparation required for analysis, we recommend that the exact PsL be used. Borgan et al. compared methods I, II and III for stratified case–cohort data, using simulations with a 10% subcohort fraction and did not note any major differences among them. However, from our Table 4, it is apparent that method I can incur more small-sample bias and lose efficiency more rapidly than the other methods with decreasing subcohort sampling fractions <10%. Method III produced the best results overall. These new findings support our conclusion that the swapper method (method III) is the method of choice for stratified case–cohort data.

Other authors have noted that the robust variance estimate can perform poorly in small samples, and the robust variance estimate cannot be used with stratified data. We therefore recommend use of the asymptotic variance estimate with both the simple (unstratified) and stratified designs.

Because risk set weighting is related to the counting-process theory underlying PL analysis, it is expected to perform well in theory. The results of Onland-Moret et al. did not resolve its performance vis-à-vis other methods because they used the simplified version with constant weight. In the simulations of Borgan et al., time-dependent weights only slightly improved efficiency of the stratified design. We conclude that gains from time-dependent weighting probably are not sufficient to justify the additional data construction effort with unstratified data. With stratified data, the data construction facilitates time-dependent weighting, but a variance estimator is not yet available. It has been noted that choice of appropriate weights might be affected when specimen storage duration or analytical batch effects are a concern (i.e. for cases outside the subcohort in a prospectively conducted case–cohort study); this would not be a problem with retrospectively conducted case–cohort studies. In fact, although we are not aware that it has been proposed before, it should be possible to perform risk set subsampling of nested control subjects from the subcohort, if the timing of case ascertainment could lead to biases due to storage or batch effects.

Missing data are an important consideration in observational studies. Regardless of design, measurements made on biological specimens can be missing for a number of reasons. Mark and Katki suggested that, as long as missingness is not related to covariate values, case–cohort analyses without time-varying weights should remain valid if cases with missing covariate information are simply ignored.

One advantage of the nested case–control and case–cohort designs as compared with the traditional case–control study design is their ability to estimate absolute risk difference (as opposed to relative risk) and survival probability. However, when there is sampling of cases or missing case data, estimators of those quantities require modification to avoid bias. Another advantage of the cohort-based sampling designs is that the reference group can be used to study population characteristics, such as allele frequencies or Hardy–Weinberg equilibrium; the control subjects
for a traditional case–control study are not population based, being left over after cases have been removed from the population.

The ability to study multiple outcomes individually without having to obtain separate sets of comparison subjects is often cited as one of the benefits of the case–cohort design, but this aspect is seldom illustrated or examined in the literature. In particular, the use of the same comparison subjects induces correlations among results for multiple outcomes. Sørensen and Andersen investigated this via asymptotic theory and simulations and, using competing risks and martingale theory, derived a consistent estimate of the asymptotic covariance matrix. They found that correlations increase with smaller subcohort sampling fractions or, equivalently, with larger cohorts and a fixed subcohort size. For example, with a cohort of size 500 and subcohort sampling fraction of 10%, correlations were substantially larger than 0.5 for the scenarios they studied. With a subcohort sampling fraction of ~50%, they estimated correlations of 0.1–0.2 for their scenarios. In the immunogenome and cancer study described here, where the subcohort sampling fraction was 50%, correlations between effect estimates for multiple cancer outcomes should therefore not be a problem (studies involving cancers at sites other than the lung are ongoing). However, such correlations could be a problem with more typical designs (25 or 10%), which may be used if more expensive procedures, such as larger numbers of loci or whole-genome sequencing, are used.

Another interesting aspect of the case–cohort design is that outcomes occurring more frequently have smaller effective control:case ratios given that the number of comparison subjects is fixed. Kulathinal et al. noted that choosing a subcohort size suitable for more common outcomes would provide flexibility for including additional, less common outcomes after the fact of subcohort selection. With common outcomes, sampling of cases may be preferable to using all cases in terms of reducing effort without seriously affecting statistical efficiency.49

Even though ascertainment of some covariates, such as genotypes, may be conducted only on the case–cohort sample, efficiency may be lost because analysis based on the case–cohort sample alone wastes other information available in the entire cohort. For example, if modelling adjustment is made for important factors or potential confounders that are known in all cohort subjects, that adjustment is inefficient if restricted only to the case–cohort sample. Further efficiency gains may be obtained by considering the case–cohort sample as the second phase in a two-phase study design by weighting the analysis of the case–cohort sample through post-stratification or calibration. A similar strategy can be applied to nested case–control studies and more general multi-phase methods can be considered. Alternatively, efficiency might be improved by considering the case–cohort study as a full cohort with missing data using, for example, inverse probability of sampling weighted methods. We do not pursue those methods in this work, as they are beyond the scope of traditional case–cohort analysis, and we believe that more comparative work needs to be done to fully understand the relative merits of these new approaches vis-à-vis the conventional design and analysis of case–cohort studies. We are currently exploring the two-phase approach in comparison with the conventional methods recommended in the present review.

The typical case–cohort analysis is based on a restricted model for the hazard rate (incidence or mortality rate) as a log-linear function of explanatory covariates. Kong and Cai described how to fit accelerated failure time models to case–cohort data. The accelerated-failure-time model does not require the assumption of proportional hazards, although that assumption can also be relaxed in standard proportional hazards models by including interactions of covariates with the basic time variable. Apart from Langholz and Jiao describing how to estimate absolute risks, a frustrating aspect of many current software implementations of case–cohort analysis methods is the restriction to standard proportional hazards models. In radiation risk assessment, more general models are routinely used. A Bayesian full-likelihood approach may allow more general risk models. At present, Epicure only allows fitting general risk models with unstratified case–cohort data, but a procedure to accommodate stratified data should become available in the near future.

Because we have focused on gene–environment interaction, a few words about such tests are appropriate. Failure to adequately model the main effect of an environmental risk factor can lead to inflated type I errors (false positives) in tests of gene–environment interaction involving that factor. This result emphasizes the need for use of general risk models, such as the ERR model used in our example. In our opinion, models for main effects of genotype have often not been given due consideration. Perhaps this is due to the popularity of the Cochran–Armitage trend test, which is applied under the assumption of a linear (co-dominant) genomic model. It is unfortunate that one of the major software packages for genomic analysis (R haplo.stats) does not include the arbitrary two-parameter genomic model which we deem an important starting point (e.g. for testing homogeneity with biallelic loci). The case-only design can be powerful when gene and environmental factor are independent, and hybrid methods address the fact that such independence often cannot be verified. However, the case-only design is limited to a multiplicative test of statistical interaction, whose use has been over-emphasized. Furthermore, the subcohort in a
case–cohort design is suitable for cross-sectional analyses, such as assessing allele frequencies and Hardy–Weinberg equilibrium (an important but not essential assumption for association studies).

Finally, we note that if the only impact of a gene is to modify the effect of environmental exposure, a main effect of genotypes may not be required in the model. Alternatively, it is conceivable that a gene evidencing both a direct effect on outcome and modification of risk of an environmental factor might not necessarily exert both effects according to the same genomic model if they involve separate mechanisms. A number of further issues are discussed in detail in two recent publications.

In conclusion, for studying interactions with an exposure variable, we recommend that case–cohort studies be conducted based on exposure-stratified random sampling when exposure can be ascertained in the entire cohort. When stratification is not necessary, the exact PSL with asymptotic variance estimates may be used and the most flexible implementation is to be found in the Epicure software. With stratified case–cohort data, we recommend the score-unbiased method III (the swapper method) of Borgan et al., with variances estimated using the asymptotic method. Stratified data may be analysed using the SAS macros of Langholz and Jiao or the S-plus function provided here (Supplementary Appendix A1 available at IJE online). Further work is needed to extend existing software for stratified case–cohort data to allow fitting of risk models that are more general than the standard proportional hazards model.

Supplementary Data

Supplementary Data are available at IJE online.

Funding

This work was supported by the Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan, a private, non-profit foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the U.S. Department of Energy (DOE), the latter in part through the National Academy of Sciences [RERF Research Protocol RP 4-04]; the Ministry of Education, Culture, Sports, Science and Technology of Japan [Grant-in-Aid for Scientific Research (B) 21390199]; and the Ministry of Health, Labour and Welfare of Japan [Grant-in-Aid for Cancer Research 22090501].

Acknowledgements

The authors express sincere gratitude to Professor Bryan Langholz for assistance with the swapper method for stratified case–cohort data and to Dr Robert Abbott for guidance on the logistic regression approximation to Cox regression (both have provided their consent to be so acknowledged).

Conflict of interest: D.L.P. was the principal developer of the Epicure software, which is now owned by Risk Sciences International (RSI). Although he no longer owns the software, there is a possibility he might receive royalties on some future sales of the software.

KEY MESSAGE

- We summarize the literature on case–cohort design and analysis and make recommendations regarding its implementation for studies of gene–environment interaction. Stratified, rather than simple, random sampling should be preferred when exposure information exists for the entire cohort. The current state of relevant software is reviewed, and needs are identified for further development of software tools that allow fitting general, flexible risk models. Given that the subcohort can be used for assessment of population parameters, such as allele frequencies and Hardy–Weinberg equilibrium, the case–cohort design is well suited to molecular epidemiological research.

References


3184 INTERNATIONAL JOURNAL OF EPIDEMIOLOGY

Downloaded from https://academic.oup.com/ije/article-abstract/41/4/1174/687507 by guest on 22 February 2019


