multipleNCC: Inverse Probability Weighting of Nested Case-Control Data

by Nathalie C. Støer and Sven Ove Samuelsen

Abstract Reuse of controls from nested case-control designs can increase efficiency in many situations, for instance with competing risks or in other multiple endpoints situations. The matching between cases and controls must be broken when controls are to be used for other endpoints. A weighted analysis can then be performed to take care of the biased sampling from the cohort. We present the R package **multipleNCC** for reuse of controls in nested case-control studies by inverse probability weighting of the partial likelihood. The package handles right censored, left-truncated and additionally matched data, and varying number of sampled controls and the whole analysis is carried out using one simple command. Four weight estimators are presented and variance estimation is explained. The package is illustrated by analyzing health survey data from three counties in Norway for two causes of death: cardiovascular disease and death from alcohol abuse, liver disease, and accidents and violence. The data set is included in the package.

Introduction

The nested case-control (NCC) design (Thomas, 1977) (incidence density sampling/risk set sampling) is popular within epidemiology due to its cost-effectiveness and efficiency. At each event time, m controls are sampled from the subjects at risk. The controls must be event-free at the time their case experienced the event. They might also be matched to the cases on additional factors. Covariates are then obtained for cases and sampled controls.

The analysis of nested case-control data has traditionally been carried out using stratified Cox-regression, where the stratification is with respect to matched case-control sets. That method, however, does not allow for breaking the matching, i.e. analyse the data without directly considering the matched sets. Hence, the sampled controls cannot be used for other cases than they originally were sampled for. In many situations one may want to reuse the controls, for instance when analyzing a subset of the original cases or when there exist more than one endpoint of interest. A few examples of such studies are Hultman et al. (1999); Parsonnet et al. (1991); Floderus et al. (1993); Øyen et al. (1997); Tynes and Haldorsen (1997); Hankinson et al. (1998); Grimsrud et al. (2002); Levine et al. (2004); Clendenen et al. (2011); Meyer et al. (2013).

We assume a competing risks situation in this paper for ease of presentation. Thus there are different types of endpoints, and the subjects can experience at most one of them. A special instance of this is a setting with only one type of endpoint, the single event situation is therefore covered by the competing risks situation. The R package **multipleNCC** is built with competing risks in mind, however it is not limited to such situations. Even though more complex event history settings would often call for more advanced multi-state modelling, reuse of controls may be handled by using **multipleNCC** together with `"coxph"`, see Section 5.

A method for breaking the matching in NCC-designs was introduced by Samuelsen (1997) and further studied by Chen (2001); Samuelsen et al. (2007); Saarela et al. (2008); Salim et al. (2009); Cai and Zheng (2012); Salim et al. (2012); Støer and Samuelsen (2012, 2013); Støer et al. (2014). This method base the estimation on a weighted partial likelihood, thus weighted Cox-regressions are carried out. The weights are inverse sampling probabilities, which must be estimated from the data, and different estimators have been suggested.

Even though it is fairly easy to estimate the weights, it is an extra step in the analysis. Having a more automatic estimation procedure, i.e. a one-line call in R with similar syntax as `"coxph"` will make this way of analyzing NCC-data more generally available.

We present the R package **multipleNCC** in this paper. The function `"wpl"` estimates weights and carries out weighted Cox-regressions. The users can choose between four options for weight estimation. The function handles both right-censored and left-truncated data and has some possibilities for variance estimation, apart from robust variances. Additional matching is incorporated for three of the four weight estimators, and varying number of controls for all four. This is the first statistical software that performs inverse probability weighting (IPW) aimed at NCC-data.

The outline of the paper is as follows; we introduce the general framework of inverse probability weighting for nested case-control data in Section 2. Weight estimation is explained in Section 3 and variance estimation in Section 4. The package is described in Section 5 and illustrated with a data example in Section 6. A comparison between the traditional estimator and the IPW estimators is given in Section 7, followed by a discussion in Section 8.

Inverse probability weighting in NCC

We have a cohort consisting of n subjects, where the i-th individual is followed from the left-truncation time l_i, which may be zero, to time of event $\tilde{t}_i$ or time of censoring c_i. Thus the cohort members are followed from l_i to $t_i = \min(\tilde{t}_i, c_i)$. There are K competing endpoints, and each subject can at most experience one of them.

At each event time, m controls are sampled for the case experiencing the event and $m = m(t)$ may depend on time. Finally, let us introduce what we call sampling-status indicator S_i, which is required input to `"wpl"`. This indicator takes values in $\{0, 1, \ldots, K, K+1\}$, zero indicates a non-sampled subject in the cohort, 1 indicates sampled controls (and not cases) for any of the endpoints in question, while "2,...,K+1" indicate cases of one of K types.

The NCC-design is tightly connected to the Cox proportional hazards model and our model for endpoint k can be written as

$$h_{ki}(t|x_i, z_i) = h_{k0}(t) \exp\left(\beta_k' x_i + \gamma_k' z_i\right). \tag{1}$$

Here $h_{k0}(t)$ is the baseline hazard for endpoint k, x_i are covariates and confounders while z_i are additional matching variables (additional to matching on time), if additional matching has been carried out, otherwise z_i will be zero. The β_k and γ_k are the log-hazard ratios connected to x and z for the k-th endpoint.

Since the matching is broken with IPW, it will generally be important to adjust for the matching variables (Støer and Samuelsen, 2013). They are included as linear functions in Equation 1 for simplicity, however more general models with other types of functions are possible.

The controls in a NCC-design are matched to the cases on at risk status and possibly additional factors. Due to this matching, it is not straightforward to reuse the controls for other endpoints/cases since the matching must be broken. Samuelsen (1997) suggested a weighted partial likelihood which resembles the standard Cox-likelihood

$$L_k(\beta, \gamma) = \prod_j \frac{\exp\left(\beta_k' x_j + \gamma_k' z_j\right)}{\sum_{i \in \mathcal{R}_j} \exp\left(\beta_k' x_i + \gamma_k' z_i\right) w_i}. \tag{2}$$

This likelihood enables the controls to be used for other endpoints and Saarela et al. (2008) and Salim et al. (2009) were the first to discuss this likelihood in connection to competing risks. The product is over all cases of type k, while the sum is over a set $\mathcal{R}_j$, defined as all cases (of all types) and all controls at risk at t_j. The weight, $w_i = 1/p_i$, is the inverse probability that individual i is ever being sampled. This probability will be 1 for cases since all of them are sampled by design, and it must be estimated from the data for the controls. We assume time invariant covariates, although time dependent covariates are in theory possible as long as they are known at all event times at which the subject is at risk. This has, however, not yet been implemented in **multipleNCC**.

The fundamental idea behind inverse probability weighting is to adjust for the biased sample from the cohort. The sample is biased, first and foremost, with respect to the proportion of cases and controls, but with additional matching it can also be biased with respect to matching variables. The idea is then to let each control represent a number of subjects in the cohort by giving them weights larger than 1. The less probable it was for a given subject to be sampled, the more subjects in the cohort it should represent since that "type" of controls likely are under-represented in the NCC-sample. By using inverse sampling probabilities as weights, this is accomplished. The analysis is then carried out "as if the data were from a cohort study", thus by a weighted Cox-regression. This idea was first proposed by Hansen and Hurwitz (1943) with a survey sampling perspective for sampling with replacement, and later generalised to sampling without replacement by Horvitz and Thompson (1952). Inverse probability weighting is also commonly used in the context of missing data (Robins et al., 1994).

Additional matching

To increase efficiency and adjust for confounding, the controls in a nested case-control design are often matched on additional factors than at risk status. This can for instance be year of birth, sex or years since first employment. We divide such matching into two groups; category matching and caliper matching (Cochran and Rubin, 1973).

With category matching, the controls are required to have the same value on the matching variable as the case, and the matching variable is often a categorical covariate. As an example, the controls can be matched to the cases on sex and a male case is then required to have a male control. With caliper matching, the matching variable is typically continuous and the control's value of the matching

variable must lie within a specified interval around the case's value. For instance could the controls be matched on year of birth plus/minus 2 years, then the birth year of a control must be within two years of the case's birth year.

Weight estimation

The weights in Equation 2 must be estimated from the data at hand, and three types of estimators have been considered; Kaplan-Meier (KM) type of weights (Samuelsen, 1997; Salim et al., 2009; Cai and Zheng, 2012), more model based logistic regression type (Mark and Katki, 2006; Samuelsen et al., 2007; Saarela et al., 2008; Støer and Samuelsen, 2013) and local averaging (Chen, 2001), referred to as Chen-weights.

KM weights

The Kaplan-Meier type of estimator without additional matching can be formulated as

$$p_i = 1 - \prod_{l_i < t_j < t_i} \left\{ 1 - \frac{m}{n\left(t_j\right) - 1} \right\}. \tag{3}$$

Here $n\left(t_j\right)$ is the number at risk in the cohort at time t_j and m the number sampled controls per case. The estimator in Equation 3 resembles the Kaplan-Meier estimator. The KM-weights can be generalized to situations with additional matching by taking the product only over subjects that meet the matching criteria, and letting the denominator only consist of subjects at risk that meet the matching criteria. Let $n_j\left(t_j\right)$ count the subjects at risk at time t_j who meet the matching criteria of case j. Then the formula with additional matching can be expressed as

$$p_i = 1 - \prod_j \left\{ 1 - \frac{m}{n_j\left(t_j\right) - 1} I\left(\text{Control } i \text{ could be sampled for case } j\right) \right\}. \tag{4}$$

By replacing m with $m\left(t_j\right)$, where $m\left(t_j\right)$ is the number of sampled controls for the case at t_j, the situation with varying number of controls is covered.

Logistic regression weights

A more model based approach is to use logistic regression models, either traditional logistic regression or the more flexible generalized additive model (GAM) (Hastie and Tibshirani, 2009, Chap. 9) with logit-link. The sampling indicator, O_i is used as outcome and the left-truncation time and censoring time as covariates with ξ being an intercept term:

$$p_i = \mathbb{E}[O_i | t_i, l_i] = \frac{\exp\left(\xi + f\left(t_i, l_i\right)\right)}{1 + \exp\left(\xi + f\left(t_i, l_i\right)\right)}. \tag{5}$$

It is important to note that this regression is carried out on the cohort excluding cases of all types. The reason for this is that all cases are sampled with a known probability of 1, thus including them in the regression would interfere with the estimation of the sampling probabilities for the controls.

With $f\left(t_i, l_i\right) = f_1\left(t_i\right) + f_2\left(l_i\right)$ where $f.\left(.\right)$ are linear functions, the estimator in Equation 5 is the traditional logistic regression model, and the inverse of those probabilities are referred to as GLM-weights. When $f\left(.\right)$ are smooth functions, the result is a GAM-model. In situations with additional matching, the matching variables should also be included in the regression model. Category matched variables are included as categorical factors while caliper matched variables are included as continuous covariates with GLM-weights and as smoothed functions with GAM-weights.

The number of controls for each case is not an explicit part of the estimation procedure for logistic regression weights and therefore no extra care must be taken in situations with time- and endpoint-dependent number of controls.

Local averaging weights

Before we can introduce the Chen-weights (Chen, 2001), which is also known as local averaging, we need some additional notation. Let $0 = l^0 < l^1 < \ldots < l^A$ be a partition of the range of left-truncation

times and $0 = t^0 < t^1 < \ldots < t^B$ a partition of the range of follow-up times where t^A and t^B is the upper limit of the left-truncation times and censoring times respectively. We also define $\mathcal{J}_a = \left(l^{a-1}, l^a\right]$ and $\mathcal{I}_b = \left(t^{b-1}, t^b\right]$. The intervals in each direction are taken to be of the same length in `"wpl"`, however the interval length can in principle vary. The local averaging weights can then be expressed as

$$w_{ab} = \frac{\sum_{i=1}^{n} I\left(l_i \in \mathcal{J}_a, t_i \in \mathcal{I}_b, i \text{ is not a case}\right)}{\sum_{i=1}^{n} I\left(l_i \in \mathcal{J}_a, t_i \in \mathcal{I}_b, i \text{ is a sampled control and not a case}\right)}.$$

All controls included in the study in $\mathcal{J}_a$ with a censoring time in $\mathcal{I}_b$ are given weight w_{ab}. Hence, all subjects sampled within the same combination of intervals will be given the same weight. Samuelsen et al. (2007) noted that this amounts to post-stratifying on grouped left-truncation and censoring times.

Generalizing these weights to handle additional matching would require partitioning of the matching variables in addition to the range of left-truncation and right-censoring times, and this will introduce a large number of combination of intervals. Due to this, the weights will likely be unstable since a small number of subjects would belong to each group. The Chen-weights are therefore only implemented for situations without additional matching.

As with logistic regression weights, the number of controls per case is not an explicit part of the estimation procedure for local averaging. Situations with time- and endpoint-dependent number of controls are therefore handled with no modification of the estimator.

Variance estimation

Since the subjects enter the weighted partial likelihood (Equation 2) whenever they are at risk, the likelihood contributions are not independent and the variance estimation cannot be based on the inverse of the information matrix only. A simple, yet sometimes conservative solution (Cai and Zheng, 2012), is to use robust variances (Lin and Wei, 1989; Barlow, 1994). This is the default option in `"wpl"`. A variance estimator for the KM-weights can be found in Samuelsen (1997) when there is no additional matching. A variance estimator for Chen-weights is given in Chen (2001), but since Samuelsen et al. (2007) argues that this estimator can be considered as a post-stratified case-cohort estimator, one may apply the variance estimator of Borgan et al. (2000). We have implemented the variance estimator of Samuelsen (1997), which exactly account for the procedure used to estimate the weights, and extended it to allow for additional matching, see details below and web-appendix to Cai and Zheng (2012). For the Chen-weights, we have implemented the variance estimator based on post-stratification, but only without additional matching since, as discussed in Section 3.3, the approach will be difficult to extend to additional matching.

Samuelsen (1997) showed that the covariance matrix with KM-weights (without additional matching) can be estimated by

$$\hat{\Gamma} = \hat{I}^{-1} + \hat{I}^{-1}\hat{\Delta}\hat{I}^{-1}. \tag{6}$$

Here $\hat{I}^{-1}$ is the inverse of the information matrix returned from the weighted Cox-regression and

$$\hat{\Delta} = \sum_i U_i U_i^\top \frac{1-p_i}{p_i^2} + \sum_{i \neq j} U_i U_j^\top \frac{\widehat{\text{cov}}_{ij}}{p_{ij} p_i p_j}. \tag{7}$$

The U_i is score contribution for individual i and $U_i^\top$ its transpose, p_{ij} the estimated probability that both individual i and j were sampled and $\widehat{\text{cov}}_{ij}$ the estimated covariance between sampling indicators for these two individuals. Note that the score contributions U_i and the W_i's considered by Samuelsen (1997) are asymptotically equivalent. The indices i and j in the sums run over all non-cases in the study. Let also U be a matrix consisting of all U_i for non-cases. It was argued by Samuelsen (1997) that the term p_{ij} can be replaced by $p_i p_j$. A formula for $\widehat{\text{cov}}_{ij}$ is given in Samuelsen (1997). We have implemented this variance formula both with and without additional matching, although in the latter case the $\widehat{\text{cov}}_{ij}$ needs modification. If two individuals i and j cannot both be sampled for any same case, then these two individuals are sampled independently and the covariance of their sampling indicators equals zero. If they can both be sampled for one or more of the same cases then $\widehat{\text{cov}}_{ij}$ is obtained using equation (3.1) in Samuelsen (1997), replacing the number at risk at time t_j by the number at risk $n_j\left(t_j\right)$ at t_j satisfying the matching criteria, and taking the product only over the cases where both i and j can be sampled as controls.

The covariance matrix estimator in Equation 6 has been considered numerically hard to calculate. However, it can be simplified by noting that Equation 7 can be expressed as a matrix product. Let R

be a matrix with the terms $(1-p_i)/p_i^2$ for all non-cases along the diagonal and terms $\widehat{\text{cov}}_{ij}/\left(p_i p_j\right)^2$ otherwise. Let furthermore $U_{(l)}$ be a vector of the l-th component of U. The the component $\hat{\Delta}_{kl}$ of $\hat{\Delta}$ can be expressed as $U_{(l)}^\top R U_{(k)}$ and taken together this means that with U being the matrix consisting of all U_i for the non-cases we get $\hat{\Delta} = U^\top R U$. Thus the calculation of the second sum in Equation 7 as a double for-loop is avoided. However, with additional matching and many sampled non-cases the $\widehat{\text{cov}}_{ij}$ and the matrix R may still be computationally demanding.

The R package multipleNCC

The **multipleNCC**-package is implemented in one main function 'wpl.r' which carries out the weighted Cox-regressions. It calls one of four hidden functions; `"pKM"`, `"pGAM"`, `"pGLM"` and `"pChen"` for weight estimation and may call two additional functions `"ModelbasedVar"` and `"PoststratVar"` for variance estimation. The functions that estimate the sampling probabilities can be accessed through the wrap-functions `"KMprob"`, `"GAMprob"`, `"GLMprob"` and `"Chenprob"`. The function `wpl(.)` is used for estimation of hazard ratios. It has a number of mandatory arguments and some which are set by default values if not included by the user, see Table 1. An important part of the `wpl`-function is the weighted Cox-regression which is carried out by `"coxph"` in the **survival**-package (Therneau and Grambsch, 2000; Therneau, 2015).

It is important to note that most arguments should have cohort dimension (Table 1). By that we mean that the arguments should have the same length as the number of subjects n in the cohort. Thus, partially known covariate information must be imputed. These imputed values are not included in the estimation, hence any values can be chosen, however `NA` should be avoided.

The GLM-weights are estimated with the `glm`-function in R with '`family=binomial`'. For GAM-weights, the `gam`-function in the **mgcv**-library (Wood, 2015) is used, also with '`family=binomial`'. The KM- and Chen-weights are estimated as explained in Sections 3.1 and 3.3.

For GLM-weights, the matching variables are included in the logistic regression as continuous covariates with caliper matching and as categorical covariates for category matching. For GAM-weights the matching variables are included as smooth functions with caliper matching and as categorical covariates for category matching. If a matching variable has many levels, a large number of parameters must be estimated and some sort of grouping of the levels of the category matching variable(s) could be sensible. Such grouping must be applied to the matching variable before it enters `"wpl"`.

If the method of variance estimation is not specified, the robust option is chosen by default. The `"Modelbased"` option can be chosen for KM-weights. Using the `"Modelbased"` for other weights will result in an error message. Similarly if the option `"Poststrat"` is chosen together with other weights than `"Chen"`, an error message will be displayed.

The code for fitting a wpl-model is

```
wpl(formula, data, samplestat)
```

The formula is a *formula*-object and has the same syntax as the formula in the `coxph`-function. The minimal code for fitting a wpl-model is therefore

```
wpl(Surv(survival_time, status) ~ X, data, samplestat)
```

This is a model with 1 control per case, no additional matching or left-truncation and KM-weights. With left-truncation and GLM-weights it would be

```
wpl(Surv(left_time, survival_time, status) ~ X, data, samplestat, weight.method = "glm")
```

and with additional matching

```
wpl(Surv(left_time, survival_time, status) ~ X + M, data, samplestat,
+ weight.method = "glm", match.var = M, match.int = c(-2, 2))
```

In this model the controls are matched to the cases on only one additional factor `M`, which for instance could be year of birth, plus/minus two years, represented by '`match.int=c(-2,2)`'. Generally, it would be important to adjust for the additional matching variables, since the matching has been broken, thus `M` is also included as an ordinary covariate. If there are more than one matching variable, for instance $M1$ and $M2$, both of them should be included as covariates and the `match.var` should be a matrix consisting of the $M1$ and $M2$. The `match.int` argument is still a vector with the intervals in the same order as in `match.var`. Thus the syntax for a model with one caliper matching criterion and one category criterion could be

Argument	Description	Default value
`Surv` object	A survival object.	Mandatory
`formula`	A formula object, with the response on the left of a ~operator, and the terms on the right. The response must be a survival object. The `status` variable going into `Surv` should have 1 for cases and zero for controls and non-sampled subjects. Cohort dimension.	Mandatory
`data`	A `data.frame` in which to interpret the variables named in the `formula`. Cohort dimension.	Mandatory
`samplestat`	A vector containing sampling and status information: 0 represents non-sampled subjects in the cohort, 1: sampled controls, 2,3,... indicate different events. Cohort dimension.	Mandatory
`m`	Number of sampled controls. A scalar if equal number of controls for all case. If unequal number of controls per case: A vector of length number of cases. The vector must be in the same order as the cases in the `samplestat`-vector	1
`weight.method`	One of four weights; `KM`, `gam`, `glm`, `Chen`.	`KM`
`no.intervals`	Number of intervals for censoring times for Chen-weights with only right censoring.	10
`variance`	`Robust`, or `Modelbased` for KM-weights and `Poststrat` for Chen-weights.	Robust
`left.time`	Entry time if the survival times are left-truncated. Cohort dimension.	Zero
`no.intervals.left`	Number of intervals for Chen-weights with left-truncation. A vector on the form [number of intervals for left truncated time, number of intervals for survival time].	[3,4]
`match.var`	If the controls are matched to the cases (on other variables than time), `match.var` is the vector or matrix of matching variables. Cohort dimension.	Zero
`match.int`	A vector of length 2*number of matching variables. For caliper matching (matched on value pluss/minus epsilon) `match.int` should consist of `c(-epsilon,epsilon)`. For exact matching `match.int` should consist of `c(0,0)`.	Zero

Table 1: Arguments to `wpl`.

```
wpl(Surv(left_time, survival_time, status) ~ X + M1 + M2, data, samplestat,
+ weight.method = "gam", match.var = cbind(M1, M2), match.int = c(-2, 2, 0, 0))
```

Note that the interval should be 'c(0,0)' for category matching. As an example, if the controls were matched on year of birth plus/minus 2 years, county of residence and years since first employment plus/minus 6 months, with year of birth and year since first employment measured in years, 'match.int = c(-2,2,0,0,-0.5,0.5)'.

In a traditional analysis of NCC-data, subjects appear in the data set as many times as they are sampled. For instance if a subject is first sampled as a control and later itself becomes a case, it appears in the data set first as a control and then a second time as a case. With IPW, the subjects should only appear in the data set once, and it is important to let all subjects who at some point became a case, be a case in the data set.

The event indicator included in the Surv-function could have a one for cases and a zero for non-cases. However, this event indicator is not actually used by the program. With only right-censored data an event indicator is not required by 'Surv' and it can thus be omitted. When the data is left-truncated the event indicator must be included and we therefore suggest to always include it even though it is not used by the program. All information regarding events, possibly of different types, controls and non-sampled subjects in the cohort is included through the sampling-status indicator 'samplestat'. Non-sampled subjects should be given value zero, sampled controls (who are not cases of any type) should have 1, while cases of the first type should be given 2, cases of the second type, 3 etc. By default all subjects with non-zero values (except for the cases in question) are used as controls. If this is not desirable, the sampling-status indicator can be modified, see Section 5.1.

The estimated weights can be extracted directly from the wpl-object by $weights, it is, however important to note that the data is sorted by inclusion time inside wpl and the ordering of the weights is therefore not the same as the ordering of the original data. The sampling probabilities can also be estimated directly with KMprob, GAMprob, GLMprob and Chenprob which also return them in the same order as the input data. These four functions have similar syntax, although with some differences in required arguments.

```
KMprob(survtime, samplestat, m, left.time = 0, match.var = 0, match.int = 0)

GLMprob(survtime, samplestat, left.time = 0, match.var = 0, match.int = 0)

GAMprob(survtime, samplestat, left.time = 0, match.var = 0, match.int = 0)

Chenprob(survtime, samplestat, no.intervals = 10, left.time = 0,
+ no.intervals.left = c(3, 4))
```

Some arguments to these functions are mandatory, while some arguments should only be supplied in given situations. Those arguments are 'left.time' and 'no.intervals.left' which should only be included with left-truncated data, and 'match.var' and 'match.int' only with additionally matched data. All arguments to the functions above can be found in Table 1, except for 'survtime' which is the follow-up time. It is important to note that 'survtime', 'samplestat', 'left.time' and 'match.var' should have length or number of rows equal to the cohort size n.

When the controls are matched on additional factors it may happen that there are none, or too few, eligible controls for some cases. Normal practice is then to widen the matching criteria somewhat for those particular cases. But by doing this there will be fewer subjects at risk that meet the original matching criterium than there are sampled controls. "wpl" will therefore print out a warning telling the user for which case this happen and that the controls for that particular case are given weight = 1. This is reasonable since those controls are sampled with probability close or equal to 1.

multipleNCC does not carry out a traditional analysis using a stratified Cox-regression. The main reason for this is that each subject should only appear once in the data set provided to **multipleNCC**. Without explicit information about the original case-control sets it is impossible to reconstruct the original nested case-control data where each subject appear as many times as they are sampled. Additionally, it is so simple to carry out a traditional nested case-control analysis with a stratified Cox-regression that there is no reason to include it in **multipleNCC**.

Sub-endpoints and complex multiple endpoint situations

A main endpoint can often be divided into sub-endpoints. For instance a cancer endpoint can be divided into metastatic and non-metastatic cancer. Analyzes of sub-endpoints using all sampled controls can be carried out with "wpl" by making a new sampling-status indicator with unique values for each sub-type. So instead of having a samplestat indicator from 0-2 (0=non-sampled subjects,

1=sampled controls, 2=cancer cases), it will have values 0-3, where 2 correspond to metastatic cancer and 3 to non-metastatic cancer.

Many multiple endpoint applications do not fit into a competing risks framework. However, reuse of controls can still be of interest. The solution can then be to estimate the sampling probabilities with `KMprob`, `GAMprob`, `GLMprob` or `Chenprob` and use the inverse of those estimated probabilities as weights in the ordinary `coxph`-function with robust variances. An example of a multiple endpoint situations which do not fit into the competing risks framework is subsequent events (Støer et al., 2014). This is a situation with two types of events and the second type may only occur after the first, e.g. incidence of prostate cancer and death from prostate cancer. Using all sampled controls when analyzing the subsequent endpoint may increase the efficiency substantially.

Other models and estimators

Inverse probability weighting is a standard approach for handling missing data and case-control data can be considered as data missing by design. Thus the functions for estimating inclusion probabilities described above can be used more generally for addressing other models than proportional hazards. For instance Samuelsen (1997) discuss the possibility of fitting parametric survival models, Suissa et al. (1998) considered estimation of standardized mortality ratios and Cai and Zheng (2012) discuss estimation from estimating equations in general with IPW from NCC studies.

Often software allows for weighting and then it is straightforward to obtain the weighted estimators. Examples of this are Nelson-Aalen and Breslow estimators for cumulative hazards or additive hazards models. With respect to variance estimation of weighted estimators we believe that robust variances often will give results that closely matches the model based, although theoretically they are conservative (Samuelsen, 1997; Cai and Zheng, 2012). In general, however, theory has not been carefully developed for such estimators and implementation of robust procedures has not generally been validated, thus care should be taken when interpreting the variance estimates.

Furthermore, we believe that the IPW weights for NCC are in particular useful when considering time to event data with the original cohort follow-up scheme. We are less convinced that the IPW approach is the best choice when for instance estimating the population means of the exposures obtained in the NCC or considering regression models for how such exposures depend on other cohort information.

Other packages

There exists a number of packages for weighted analysis for different purposes. Some are aimed at causal inferences such as **ipw** (van der Wal and Geskus, 2011) and **MatchIt** (Ho et al., 2011). Other have a missing data or survey sampling perspective. Two such packages are **NestedCohort** (Mark and Katki, 2006; Katki and Mark, 2008) and **survey** (Lumley, 2004, 2014). The **survey**-package was developed for analyses of survey data, but include functions for two-phase designs. A nested case-control design can be seen as a two-phase design where the cohort is collected at Phase 1 and the Phase 2 data is the cases and sampled controls with additional covariate information. The Phase 2 sampling scheme is of course somewhat more complex than what is usually seen in survey sampling. However, when Chen-weights are used for the nested case-control design, a stratified Phase 2 sampling is implicitly assumed, which is a well-known survey sampling design. Hence, a weighted analysis using Chen-weights can be carried out by specifying a stratified two-phase design with the function `twophase` and carry out the Cox-regression using this design with `svycoxph`.

The **NestedCohort** is a general package for cohort analyses with a missing data/two-phase design perspective. The theory behind it is based to an extent on Robins et al. (1994) and the variance estimators also builds on this paper. Weighted Cox-regressions are carried out with `nested.coxph` where a working model is assumed for the sampling probabilities in Phase 2. This working model is a logistic regression model for the sampling indicator with all variables contributing to the missingness as covariates. This is identical to specifying `weights=glm` in `wpl`. However, the variance estimator in `nested.coxph` may be somewhat better in special situations where the robust variances are conservative. The **NestedCohort** can however only be used for glm-weights so that the more commonly used KM-weights are not applicable with this package. The authors of the package also state that fine matching is problematic with **NestedCohort**. In the data example below `nested.coxph` gave identical estimates and standard errors as using glm-weights with `wpl`.

Analysis of example data set

The data set CVD_Accidents

We use a collection of cardiovascular health screenings as our cohort, also used in Aalen et al. (2008). All men and women aged 35-49 from three Norwegian counties: Oppland, Sogn og Fjordane and Finnmark were invited to participate in health screenings from 1974-1978. The screening consisted of a health examination including measurement of height and weight. The participants were also asked to respond to a questionnaire which among other things contained questions regarding smoking status.

More than 90% of the invited subjects chose to participate which resulted in a cohort of about 50 000 subjects. The cohort was linked to the Causes of Death Registry kept by Statistics Norway and followed up for deaths until the end of year 2000. Since there were few subjects at risk younger than 40 years, the survival times were left-truncated at age 40 or age at health screening. The left-truncation time is named `agestart` in the data. The survival time is named `agestop` and is either the age at death or censoring.

The data set included in **multipleNCC** and used as illustration here, is a random sample of 3933 subjects from the cohort. It is a cohort study, but we have carried out a synthetic nested case-control study within this smaller cohort. Having full cohort information enables comparison between `"wpl"` and corresponding analyses carried out on the full cohort. The synthetic data generation was performed in two steps. In the first step a nested case-control sampling was carried out by sampling one control per case matched sex and BMI $\pm$ 2. This was done in a for-loop over event times sampling one subject at random from those still at live at that time in addition to fit the matching criterium. When the controls are not additionally matched or matched only on a categorical variable the `ccwc`-function from the **Epi**-package (Carstensen et al., 2016) offers a simple way to create a nested case-control design from a cohort. However, if some of the matching variables are continuous it still has to be done as described above. The second step is necessary only for the weighted Cox-regression and it involves removing duplicate subjects, due to that some controls are sampled multiple times and some controls later become cases, from the sampled data. For those controls who later become cases it is important to keep the "case-entry", while for those controls sampled multiple times, the entries will be identical and one of them are kept at random.

Four types of deaths are recorded in the data; (1) cancer, (2) cardiovascular disease, (3) alcohol abuse, liver disease, accidents and violence, and (4) other medical causes. For simplicity we only use cardiovascular disease (n=236) and alcohol abuse, liver disease, accidents and violence, henceforth referred to as ALAV (n=60). The data set is included in **multipleNCC** and can be loaded by `data("CVD_Accidents")`.

When we analyze this data set with IPW, we will use all sampled controls for both endpoints, hence supplement the controls for ALAV deaths with the cardiovascular disease controls and cases. And vice versa, supplement controls for cardiovascular disease cases with cases and controls for ALAV cases. For ALAV deaths, the number of controls increase substantially when including the controls for cardiovascular deaths.

The matching variables are BMI (M_1 = `bmi`) and sex (M_2 = `sex`). The matching variables are adjusted for in the model to remove confounding due to breaking the matching. We included BMI as continuous variable and sex is a categorical variable. Smoking (X = `smoking3gr`) is our explanatory variable and is categorized into never smoked, former smoker and current smoker.

Fitting models

We consider the two models

$$h_{1i}\left(t_i|X, M_1, M_2\right) = h_{10}\left(t_i\right)\exp\left(\beta_1' X_i + \gamma_{11} M_{1i} + \gamma_{12} M_{2i}\right)$$
$$h_{2i}\left(t_i|X, M_1, M_2\right) = h_{20}\left(t_i\right)\exp\left(\beta_2' X_i + \gamma_{21} M_{1i} + \gamma_{22} M_{2i}\right).$$

Here $h_1\left(.\right)$ and $h_2\left(.\right)$ are the hazard for death from CVD and ALAV, respectively. To fit the weighted partial likelihoods corresponding to those models, the command below can be used

```
fit = wpl(Surv(agestart, agestop, dead24) ~ factor(X) + bmi + sex,
+ data = CVD_Accidents, m = 1, samplestat, weight.method =
+ "glm", match.var = cbind(CVD_Accidents$BMI, CVD_Accidents$sex),
+ match.int = c(-2, 2, 0, 0))
```

The 'status' argument to the 'Surv' function (in this example 'dead24') can in practice contain anything since the information regarding who are cases and controls, and also which type of endpoint the cases experienced are provided through `samplestat`. The '`match.int`' argument will in this

situation be a vector of two elements, 'match.int=c(-2,2,0,0)', since the controls are matched to the cases on BMI ±2, and sex.

The 'summary' and 'print' commands can be used to display the results of the analysis. The output resembles output from traditional Cox-regression (see next page), and the results for the different endpoints are printed below each other. Both the naive and robust/estimated standard errors are printed, although only the robust/estimated should be reported.

The $-operator is usually used when specific parts of an object is to be extracted (i.e. 'fit$coefficients'). When there is more than one endpoint this will only give you the corresponding element for the first endpoint, i.e. 'fit$coefficients' will give you (0.47107, 1.34245, 0.08051, -1.22307), thus only the estimates for the cardiovascular disease endpoint. To extract any element from the object, [[]] can be used, i.e. 'fit[[23]]' will give you the estimates from the second analysis of death from ALAV. For a full list of elements in the "wpl"-object, the names-function can be used. Each element will occure the same number of times as there are different endpoints, and the first occurrence correspond to the first endpoint, the second occurrence to the second endpoint, etc.

```
summary(fit)
Endpoint 1 :
Call:
wpl.formula(formula = Surv(agestart, agestop, dead24) ~ factor(smoking3gr) +
    bmi + factor(sex), data = CVD_Accidents,
    samplestat = CVD_Accidents$samplestat,
    match.var = cbind(CVD_Accidents$bmi, CVD_Accidents$sex),
    match.int = c(-2, 2, 0, 0), weight.method = "glm")

  n= 566, number of events= 236

                       coef exp(coef) se(coef) robust se      z Pr(>|z|)
factor(smoking3gr)2  0.47107   1.60171  0.22099   0.26057  1.808  0.07062 .
factor(smoking3gr)3  1.34245   3.82842  0.19400   0.23424  5.731 9.98e-09 ***
bmi                  0.08051   1.08384  0.01799   0.02562  3.143  0.00167 **
factor(sex)2        -1.22307   0.29433  0.15980   0.22475 -5.442 5.27e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
factor(smoking3gr)2    1.6017     0.6243    0.9611    2.6692
factor(smoking3gr)3    3.8284     0.2612    2.4190    6.0591
bmi                    1.0838     0.9226    1.0308    1.1396
factor(sex)2           0.2943     3.3976    0.1895    0.4572


Endpoint 2 :
Call:
coxph(formula = Surv(left.time.ncc, survtime.ncc, status.ncc ==
    i) ~ x + cluster(ind.no.ncc), weights = 1/p)

  n= 566, number of events= 60

                       coef exp(coef) se(coef) robust se      z Pr(>|z|)
factor(smoking3gr)2 -0.61629   0.53994  0.45484   0.48347 -1.275 0.202413
factor(smoking3gr)3  0.92343   2.51792  0.32202   0.34402  2.684 0.007270 **
bmi                  0.08383   1.08744  0.03813   0.04587  1.828 0.067619 .
factor(sex)2        -1.42549   0.24039  0.32702   0.36888 -3.864 0.000111 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
factor(smoking3gr)2    0.5399     1.8520    0.2093    1.3928
factor(smoking3gr)3    2.5179     0.3972    1.2829    4.9417
bmi                    1.0874     0.9196    0.9939    1.1897
factor(sex)2           0.2404     4.1599    0.1167    0.4954
```

The sampling probabilities can be directly estimated using one of the functions: `KMprob`, `GAMprob`, `GLMprob` and `Chenprob`, for instance

```
glmp = GLMprob(CVD_Accidents$agestop, CVD_Accidents$samplestat,
         left.time = CVD_Accidents$agestart,
         match.var = cbind(CVD_Accidents$sex, CVD_Accidents$bmi),
         match.int = c(0, 0, -2, 2))

kmp = KMprob(CVD_Accidents$agestop, CVD_Accidents$samplestat, 1,
             left.time = CVD_Accidents$agestart,
             match.var = cbind(CVD_Accidents$sex, CVD_Accidents$bmi),
             match.int = c(0, 0, -2, 2))
```

It is of interest to examine the weights for the sampled controls only, as the cases have weight 1 and the non-sampled subjects will not affect estimates. We can for instance inspect summary statistics

```
summary(1/glmp[CVD_Accidents$samplestat == 1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.575   6.545   9.201  13.240  15.080  73.180

summary(1/kmp[CVD_Accidents$samplestat == 1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.481   6.839   8.772  13.150  15.020  95.810
```

The two type of weights correspond well, although KM-weights have a somewhat heavier tail. It is worth noting that the subjects with the largest weights are those with the shortest follow-up time and therefore those subjects do not have a too large impact on the analysis even though their weight is large, since they are included in few risk sets. GAM-weights were similar to GLM-weights, but with a somewhat heavier tail (not shown) and Chen-weights are not applicable here since they are not implemented for additional matching.

Comparison between stratified coxph and wpl

Table 2 displays a comparison between the full cohort analysis, the traditional estimator for nested case-control data and the IPW-estimator using `"wpl"` with GLM- and KM-weights. Being a smoker significantly increases the risk of death from cardiovascular disease. Being a former smoker also increases the risk of death from cardiovascular disease, although only significantly in the cohort analysis. Being a former smoker has a non-significant protective effect on death from ALAV, while being a smoker significantly increases the risk of dying from ALAV.

The hazard ratios estimated with the traditional estimator and with `"wpl"` are fairly similar to the cohort estimates taking into account the size of the standard errors. The most pronounced difference is between the cohort hazard rate and the hazard rate for the traditional estimator for being a smoker on death from ALAV, 2.05 vs. 2.90. However, considering the size of the standard errors this is not a large difference.

For the cardiovascular disease endpoint, the standard errors of the IPW analyses are somewhat smaller than the standard errors of the traditional estimator, resulting in a little bit higher efficiency for the IPW-estimators. For the ALAV endpoint, the standard errors are substantial lower and a large efficiency gain is obtained with the IPW-estimators. The reason for this is that the IPW-estimators make use of a number of extra controls as all cases and controls from the cardiovascular disease endpoint are used as additional controls. On average the number of controls per case increase from 1 to more than 8. We have chosen to include cardiovascular disease cases as additional controls for the cases who died from ALAV, and also the cases who died from ALAV as additional controls for the cases who died from cardiovascular disease. However, the cases who are used as additional controls will contribute little to the analysis since they are non-cases (for the particular endpoint) with weight equal to 1. We have reported robust standard errors for KM-weights in Table 2, however the estimated standard errors are very similar, for CVD endpoint 0.27 and 0.24 and for ALAV 0.48 and 0.35 for former and current smoker respectively.

Discussion

We have demonstrated the **multipleNCC**-package in R which allows for breaking the matching and reusing controls in NCC-designs. The main function`"wpl"` estimates sampling probabilities and perform weighted Cox-regressions. It handles right censored, left-truncated and additionally

	CVD					
	Former smoker			Current smoker		
	HR(95 % CI)	SE(β)	Eff.	HR(95 % CI)	SE(β)	Eff.
Cohort	1.72(1.11 - 2.66)	0.22	1.00	3.25(2.21 - 4.79)	0.20	1.00
Trad.	1.66(0.94 - 2.93)	0.29	0.59	3.52(2.09 - 5.94)	0.27	0.55
IPW-GLM	1.60(0.96 - 2.67)	0.26	0.73	3.83(2.42 - 6.06)	0.23	0.72
IPW-KM	1.65(0.98 - 2.78)	0.26	0.66	3.97(2.48 - 6.35)	0.24	0.69
	ALAV					
	Former smoker			Current smoker		
	HR(95 % CI)	SE(β)	Eff.	HR(95 % CI)	SE(β)	Eff.
Cohort	0.56(0.23 - 1.38)	0.46	1.00	2.05(1.07 - 3.90)	0.33	1.00
Trad.	0.48(0.13 - 1.69)	0.65	0.50	2.90(1.11 - 7.61)	0.49	0.45
IPW-GLM	0.54(0.21 - 1.39)	0.48	0.90	2.52(1.28 - 4.94)	0.34	0.92
IPW-KM	0.55(0.21 - 1.40)	0.49	0.90	2.56(1.28 - 5.02)	0.35	0.89

Table 2: Results from analyses on entire cohort, traditional nested case-control analyses and IPW analyses with GLM-weights and robust variances, and KM-weights and estimated variances. Never smoked used as reference, Eff. - variance for cohort estimator divided by variance for corresponding estimator. SE(β) - standard error (robust SE for IPW estimators with estimated SE in parentheses for KM-weights.

matched data, and varying number of sampled controls. We have also explained how the variance can be estimated without additional matching (KM- and Chen-weights) and with additional matching (KM-weights).

The package is particularly useful in situations with multiple outcomes. It has a competing risks perspective, in the sense that with more than one type of endpoint, each endpoint is estimated separately and all controls and cases of other types are used as additional controls. In many situations the competing risks framework may not be suitable, although reusing controls can still be of interest. The solution can then be to estimate the sampling probabilities for all cohort members, using one of the four functions `KMprob`, `GAMprob`, `GLMprob` or `Chenprob`, and carry out weighted analysis that fit the situation at hand.

The `"gam"`-function in the **mgcv**-package is used for estimation of the GAM-weights. We could alternatively have used the `gam`-function in the **gam**-package (Hastie, 2015) or even a different form of smoothing. It has however become evident that the exact value of the weights are not too important as long as they are fairly reasonable, thus the choice of smoothing does probably not affect final hazard ratios and standard errors. For the same reason there is usually only minor differences with regards to final hazard ratios and standard errors between the four weight estimators discussed in Section 3 (Støer and Samuelsen, 2012, 2013).

Sometimes the cases and controls are matched closer together than is strictly necessary. Very close matching can lead to small KM-weights since the probability of being sampled will be large for the controls that were sampled (and very small for most of the non-sampled subjects) and this could lead to biased estimates (Støer and Samuelsen, 2013). A solution could be to increase the length of '`match.int`'. For example in the data example above, we could replace '`match.int=c(-2,2)`' with '`match.int=c(-4,4)`'. Widening the matching interval could introduce bias, thus it is important to carry this out with caution. The equivalence of this for category matching is to reduce the number of levels of the matching variable by some sort of grouping.

Bibliography

O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer, first edition, 2008. [p9]

W. E. Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, 50(4):1064–1072, 1994. [p4]

Ø. Borgan, B. Langholz, S. O. Samuelsen, L. Goldstein, and J. Pogoda. Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6(1):39–58, 2000. [p4]

T. Cai and Y. Zheng. Evaluating prognostic accuray of biomarkers in nested case-control studies. *Biostatistics*, 13(1):89–100, 2012. [p1, 3, 4, 8]

B. Carstensen, M. Plummer, E. Laara, and M. Hills. *Epi: A package for Statistical Analyis in Epidemiology.*, 2016. URL http://CRAN.R-project.org/package=Epi. R package version 2.0. [p9]

K. N. Chen. Generalized case-cohort sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 63(4):791–809, 2001. [p1, 3, 4]

T. V. Clendenen, E. Lundin, A. Zeleniuch-Jacquotte, K. L. Koenig, F. Berrino, A. Lukanova, A. E. Lokshin, A. Idahl, N. Ohlson, G. Hallmans, V. Krogh, S. Sieri, P. Muti, A. Marrangoni, B. M. Nolen, M. L. Liu, R. E. Shore, and A. A. Arslan. Circulation inflammation markers and risk of Epithelial Ovarian Cancer. *Cancer Epidemiology, Biomarkers and Prevention*, 20(5):799–810, 2011. [p1]

W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(4):417–446, 1973. [p2]

B. Floderus, T. Persson, C. Stenlund, A. Wennberg, Å. Öst, and B. Knave. Occupational exposure to electromagnetic fields in relation to leukemia and brain tumors: A case-control study in Sweden. *Cancer Causes and Control*, 4(5):465–476, 1993. [p1]

T. K. Grimsrud, S. R. Berge, T. Haldorsen, and A. Andersen. Exposure to different forms of nickel and risk of lung cancer. *American Journal of Epidemiology*, 156(12):1123–1132, 2002. [p1]

S. E. Hankinson, W. C. Willett, G. A. Colditz, D. J. Hunter, D. S. Michaud, B. Deroo, B. Rosner, F. E. Speizer, and M. Pollak. Circulating concentrations of insulin-like growth factor-I and risk of breast cancer. *The Lancet*, 351(9113):1393–1396, 1998. [p1]

M. H. Hansen and W. N. Hurwitz. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:333–362, 1943. [p2]

T. Hastie. *gam: Generalized Additive Models*, 2015. URL http://CRAN.R-project.org/package=gam. R package version 1.12. [p12]

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, London, 2009. [p3]

D. E. Ho, K. Imai, G. King, and E. A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011. URL http://www.jstatsoft.org/v42/i08/. [p8]

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. [p2]

C. M. Hultman, P. Sparén, N. Takei, R. M. Murray, and S. Cnattingius. Prenatal and perinatal risk factors for schizophrenia and affective psychosis, and reactive psychosis of early onset: Case-control study. *British Medical Journal*, 318(7181):421–426, 1999. [p1]

H. A. Katki and S. D. Mark. Survival analysis for cohorts with missing covariate information. *The R Journal*, 8(1):14–19, 2008. [p8]

R. J. Levine, S. E. Maynard, C. Qian, K. H. Lim, L. J. England, K. F. Yu, E. F. Schisterman, R. Thadhani, B. P. Sachs, F. H. Epstein, B. M. Sibai, V. P. Sukhatme, and S. A. Karumanchi. Circulating angiogenic factors and the risk of preeclampsia. *New England Journal of Medicine*, 350(7):672–683, 2004. [p1]

D. Y. Lin and L. J. Wei. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989. [p4]

T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004. [p8]

T. Lumley. *survey: analysis of complex survey samples*, 2014. URL http://CRAN.R-project.org/package=survey. R package version 3.30. [p8]

S. D. Mark and H. A. Katki. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (sampled) cohort studies with missing case data. *Journal of the American Statistical Association*, 101:460–471, 2006. [p3, 8]

H. E. Meyer, T. E. Robsahm, T. Bjørge, M. Brunstad, and R. Blomhoff. Vitamin D, season and the risk of prostate cancer: A nested case-control study within Norwegian health studies. *American Journal of Clinical Nutrition*, 97(1):147–154, 2013. [p1]

N. Øyen, T. Markestad, R. Skjærven, L. M. Irgens, K. Helweg-Larsen, B. Alm, G. Norvenius, and G. Wennergren. Combined effects of sleeping position and prenatal risk factors in Sudden Infant Death Syndrome: The Nordic Epidemiological SIDS Study. *Pediatrics*, 100(4):613–621, 1997. [p1]

J. Parsonnet, G. D. Friedman, D. P. Vansdersteen, Y. Chang, J. H. Vogelman, N. Orentreich, and R. K. Sibley. Helicobacter Pylori infection and the risk of Gastric Carcinoma. *New England Journal of Medicine*, 325(16):1127–1131, 1991. [p1]

J. Robins, A. Rotnitzky, and L. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. [p2, 8]

O. Saarela, S. Kulathinal, E. Arjas, and E. Läärä. Nested case-control data utilized for multiple outcomes: A likelihood approach and alternatives. *Statistics in Medicine*, 27(28):5991–6008, 2008. [p1, 2, 3]

A. Salim, C. Hultman, P. Sparén, and M. Reilly. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. *Biostatistics*, 10(1):70–79, 2009. [p1, 2, 3]

A. Salim, Q. Yang, and M. Reilly. The value of reusing prior nested case-control data in new studies with different outcome. *Statistics in Medicine*, 31(11-12):1291–1302, 2012. [p1]

S. O. Samuelsen. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394, 1997. [p1, 2, 3, 4, 8]

S. O. Samuelsen, H. Ånestad, and A. Skrondal. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*, 34(1):103–119, 2007. [p1, 3, 4]

N. C. Støer and S. O. Samuelsen. Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Analysis*, 18(3):261–283, 2012. [p1, 12]

N. C. Støer and S. O. Samuelsen. Inverse probability weighting in nested case-control studies with additional matching - a simulation study. *Statistics in Medicine*, 32(30):5328–5339, 2013. [p1, 2, 3, 12]

N. C. Støer, H. E. Meyer, and S. O. Samuelsen. Reuse of controls in nested case-control studies. *Epidemiology*, 25(2):315–317, 2014. [p1, 8]

S. Suissa, M. Edwardes, and J. Boivin. External comparisons from nested case-control designs. *Epidemiology*, 9(1):72–78, 1998. [p8]

T. M. Therneau. *survival: A Package for Survival Analysis in S*, 2015. URL http://CRAN.R-project.org/package=survival. R package version 2.38-3. [p5]

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000. [p5]

D. C. Thomas. Addendum to: "Methods of cohort analysis: Appraisal by application to asbestos mining" by Liddell FDK, McDonald JC and Thomas DC. *Journal of the Royal Statistical Society: Series A (General)*, 140(4):469–491, 1977. [p1]

T. Tynes and T. Haldorsen. Electromagnetic fields and cancer in children residing near Norwegian high-voltage power lines. *American Journal of Epidemiolgy*, 145(3):219–226, 1997. [p1]

W. M. van der Wal and R. B. Geskus. ipw: An R package for inverse probability weighting. *Journal of Statistical Software*, 43(13):1–23, 2011. URL http://www.jstatsoft.org/v43/i13/. [p8]

S. Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*, 2015. URL http://CRAN.R-project.org/package=mgcv. R package version 1.8-6. [p5]

Nathalie C. Støer
Department of Mathematics
University of Oslo
Norway
and
Medical Epidemiology and Biostatistics
Karolinska Institutet
Sweden
nathalie.stoer@ki.se

Sven Ove Samuelsen
Department of Mathematics
University of Oslo
Norway
osamuels@math.uio.no