# Using the whole cohort in the analysis of countermatched samples

Claudia Rivera
Thomas Lumley
University of Auckland, department of Statistics.
38 Princes Street.
Auckland, New Zealand.
clriverarodriguez@gmail.com, t.lumley@auckland.ac.nz

**Abstract**

In this paper we implement calibration weights for a two-phase sample induced by countermatched designs. Weights can be calibrated by using the estimated influence functions as in Kulich and Lin (2004). This implies that we are able to use all first phase variables to improve efficiency and not just the countermatching variable as in partial likelihood. This is very useful when more than one variable is of interest, for instance, in the presence of confounders. A comparison with simple case cohort studies is made . When we happen to have a surrogate for a rare exposure or when there is a potential partially–known confounder, we obtain large improvement in efficiency using calibrated weights for countermatching. Findings are illustrated by using data from the National Wilms' Tumor Study (D' Angio et al., 1989; Green et al., 1998) and the Welsh nickel refinery workers study(Morabia et al., 1995).
**Keywords:** Calibration, countermatching, Cox model, pseudolikelihood.

## 1 Introduction

In the analysis of cohort studies with few events it may be difficult to obtain reliable estimates and a very large cohort may be needed. Different methods have been proposed to reduce cost and effort in this type of studies. The well-known case cohort study (Prentice, 1986) is one alternative. Other alternatives are methods that involve risk set sampling such as simple or matched nested case control(NCC) (Thomas, 1977) and the recently proposed stratified version of nested case–control called countermatching (CM) (Langholz and Borgan, 1995).
Case–cohort(CC) designs allow us to have more controls for each case, whereas using simple nested case–control only $m-1$ controls are selected for each case. The literature suggests that countermatching can be more efficient in situations such as the study of rare diseases (Langholz and Thomas, 1990), (Langholz and Thomas, 1991). However, it is also desirable to retain some properties of case–cohort designs for nested case–control such as having more controls per case. This was already done for simple nested case–control by Samuelsen (1997) who proposes treating the final sample of cases and controls as a two–phase sample in order to gain efficiency by reusing controls and using inclusion probabilities. He also implements a new estimation approach called pseudolikelihood. One of the advantages of pseudolikelihood is that it does not rely on martingale methods and hence no predictability requirements are needed when using auxiliary information such in calibration (Kulich and Lin, 2004).

## 2 Data and sampling model

We will mostly follow the formulation for countermatching as in Langholz and Borgan (1995) and Andersen and Gill (1982), but using notation in Breslow and Wellner (2007). The framework supposes that a random sample from $(X=\min(T,C), \delta = I_{fT\ \ Cg}, Z)$ is observed. Where $T$ and $C$ denote failure and censoring time respectively. Independence of $T$ and $C$ given $Z$ is also assumed. Moreover, failure time $T$ follows the Cox model,that is, the hazard function is given by

$$\Lambda(t) = \exp(Z^T\beta)\Lambda_0(t) \tag{1}$$

It is of interest to estimate the parameter $\beta$. The sampling that we are interested in is describe in Breslow et al. (2009a), but induced by a countermatched design.

## 3 Inclusion probabilities

In order to carry out a nested case control study, we need to know the complete history

$$\Im = \{(b_i, X_i, \delta_i); i = 1, ..., n\}.$$

where $b_i; i = 1, ...N.$ represents the time at which $i$ enters the study. Therefore, it is possible to find inclusion probabilities for such designs conditioning on $\Im$. Define $\zeta_i$ as the indicator that $i$ is ever selected either as a case or control. Samuelsen (1997) finds the inclusion probabilities induced by a simple nested case–control design. For countermatched studies, inclusion probabilities at each risk set not only depend on the stratum of the case, but also on the number of strata and number of controls. To calculate the inclusion probabilities, we first notice that the inclusion probability for $k$ at time $t$, where $t$ is a failure time, is

$$p_k(t) = P(k \in \tilde{\mathcal{R}}(t)|\Im) = \begin{cases} \frac{m_{A_k(t)}}{n_{A_k(t)}} & \text{if } A_k(t) \neq A_t(t) \\ \frac{m_{A_k(t)} \ \ 1}{n_{A_k(t)} \ \ 1} & \text{if } A_k(t) = A_t(t). \end{cases} \tag{2}$$

There $A_k(t)$ denotes the stratum to which subject $k$ belongs at time $t$ , $b_i$ denotes the time at which subject $i$ enters the study and $n_l(t)$ is the stratum size at time $t$. If $t$ is not a failure time this probability is zero for every $k = 1, ..., N$ because risk sets are only defined for failure times. Therefore, using (2), the probability of $k$ ever being selected either as a case or a control is 1 minus the probability of $k$ not being selected as a control in any of the risk sets, that is

$$\pi_k = 1 - \prod_{b_k} \prod_{X_i} \prod_{X_k} (1 - p_k(X_i)) \tag{3}$$

We also prove consistency of these inclusion probabilities following the lines of the consistency proof for the Kaplan–Meier estimator given in Andersen and Gill (1982). Moreover, as variance estimates are needed, pairwise inclusion probabilities are calculated in an additional document.

## 4 Calibration

We use calibration to improve efficiency of estimators. In the specific situation of fitting proportional hazard models for cohort data, Breslow et al. (2009a) notice that

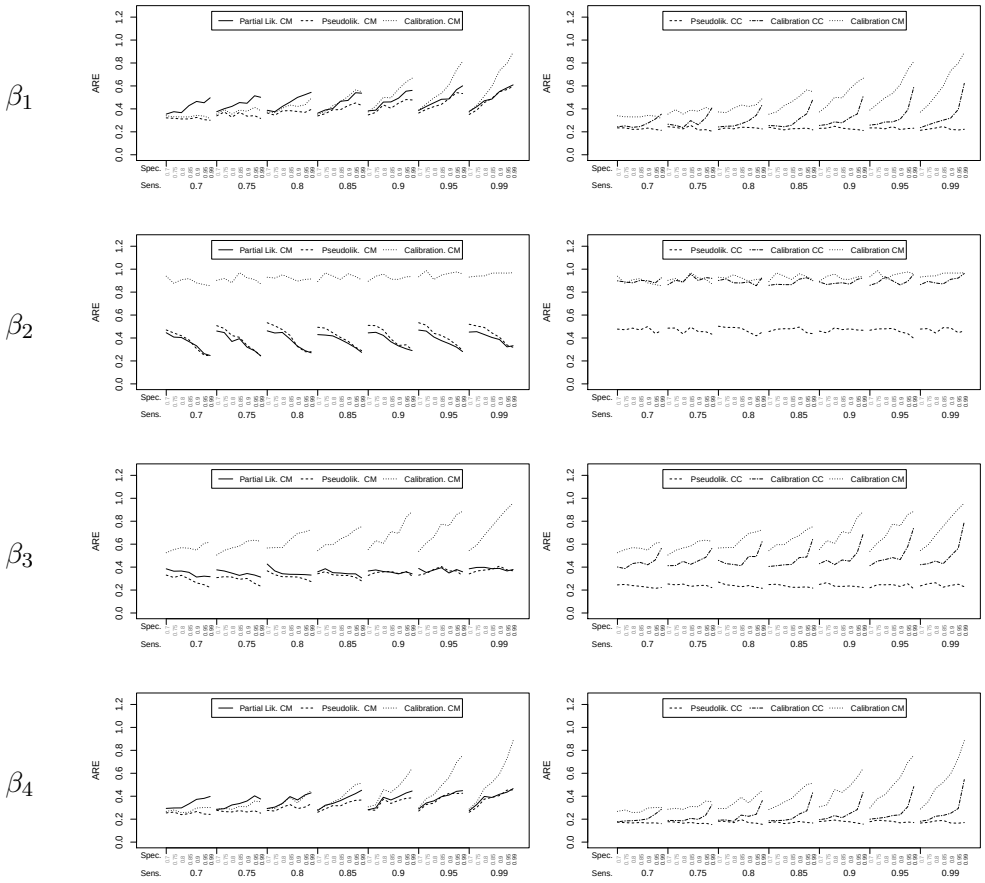
$$\hat{\beta} \approx \beta_0 + I(\beta_0)^{\ 1} \sum_{C} d_i U_i(\beta_0), \tag{4}$$

therefore, if the covariates for the whole cohort were known, the weights could be calibrated by using the influence function ($U(\beta_0)$) as auxiliary information and this may enhance the efficiency of $\hat{\beta}$. Additionally, according to van der Vaart and Wellner (2000), for sampled individuals the influence function contributions may be approximated very precisely from the observed variables by their *dfbeta*'s. So, the calibration method does not require us to find the influence functions; it is sufficient to find *dfbeta*'s, which are already provided by some statistical software such as R (R Core Team, 2013).
Breslow et al. (2009b) demonstrated that using information available for the whole cohort in a case cohort design, the variance of $\hat{\beta}$ can reduced. They follow the procedure proposed by Kulich and Lin (2004). This is also used in this work to calibrate weights.

## 5 Simulation study

The main scenario of interest is when the exposure of interest is expensive to collect but there is an inexpensive surrogate measure. Countermatching on the surrogate and gathering the true exposure for the sample is an attractive alternative (Langholz and Borgan, 1995). The countermatched design is expected to perform well and is more flexible because it allow us to draw conclusion about confounders.

Figure 1: ARE for countermatched and case–cohort simulation for model 5 and simulation 5.




## 5 Countermatching on a surrogate of exposure

A cohort of size 10000 was simulated. It is assumed that there is a binary covariate $Z_1$ with around $10\%$ exposure. That is, $Z_1 \sim Bern(0.1)$. Additionally we simulate a surrogate variable $X_0$ of $Z_1$. This is done for different levels of sensitivity and specificity (from $0.7$ to $0.95$ by jumps of $0.05$ and $0.99$ ).Additionally two continuous covariates, $X_2$ and $X_3$ were included in the model. They are supposed to be known for the entire cohort, so that we have three covariates $X_0$ and $X_2$, $X_3$ to use in calibration. The inclusion probabilities(3) were computed for individuals in each risk set. These are used in the countermatching approach.Time is simulated through an exponentially distributed variable with baseline hazard $h_0$ such that its risk function is

$$r(Z_i) = \exp(\beta_1 z_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i} \times z_{1i}), \tag{5}$$

where $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. The censoring variable is $U[0, 1]$ and is non informative and the baseline $h_0$ is chosen such that the total proportion of failures is around $2.5\%$. The counter–matched sample was composed of 2 subjects at each risk time. These two subjects are a stratified sample of the corresponding risk set, one of them is a case, the other is selected from its stratum by a simple sampling design and both of them form the corresponding sampled risk set. For the pseudolikelihood methods, the same cohort as above is used. Inclusion probabilities (3) are computed for individuals in the final sample for countermatching. For the simple case–cohort approach, simple weights are used (estimator II in Borgan et al. (1998) ).
For calibration, the approach proposed in section 4 was followed. Solid lines represent the partial likelihood method, a dashed line represents pseudolikelihood and a dotted line represents the calibration approach for countermatching. Notice that results for simple case–cohort should not depend on specificity and sensitivity when using pseudolikelihood. However, when applying calibration ARE's can change. ARE's using pseudolikelihood with countermatching are greater than those ones using case–cohort for histology and $age_0$. Using calibration as describe above gives large gains in efficiency, specially for the countermatching approach. Here again all coefficients present higher ARE's using calibrated countermatching than using calibrated case–cohort.

Table 1: Median Estimate and ARE for Wilms' tumour data. PL.: Partial Likelihood, Ps.: Pseudolikelihood, and Cal.:Pseudolikelihood with calibrated weights.

| | Median Est. | | | | | ARE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CC | | CM | | | CC | | CM | | |
| Coefficient | Ps. | Cal. | PL | Ps. | Cal. | Ps. | Cal. | PL. | Ps. | Cal. |
| histol ( 4.04 ) | 4.13 | 4.00 | 4.40 | 4.05 | 4.03 | 0.45 | 0.75 | 0.17 | 0.63 | 0.87 |
| $age_0$ ( -0.66 ) | -0.68 | -0.68 | -0.62 | -0.669 | -0.67 | 0.56 | 0.96 | 0.71 | 0.74 | 0.98 |
| $age_1$ ( 0.1 ) | 0.11 | 0.10 | 0.09 | 0.108 | 0.10 | 0.43 | 0.89 | 0.26 | 0.28 | 0.88 |
| stage ( 1.35 ) | 1.34 | 1.37 | 1.35 | 1.334 | 1.35 | 0.37 | 0.84 | 0.29 | 0.37 | 0.88 |
| Diam. ( 0.07 ) | 0.07 | 0.07 | 0.05 | 0.07 | 0.07 | 0.47 | 0.89 | 0.42 | 0.38 | 0.89 |
| hs:$age_0$ ( -2.64 ) | -2.72 | -2.60 | -3.05 | -2.66 | -2.64 | 0.4 | 0.65 | 0.19 | 0.55 | 0.7 |
| hs:$age_1$ ( -0.06 ) | -0.05 | -0.05 | -0.03 | -0.051 | -0.04 | 0.23 | 0.27 | 0.3 | 0.23 | 0.23 |
| stg:Dm ( -0.08 ) | -0.07 | -0.08 | -0.07 | -0.07 | -0.08 | 0.36 | 0.84 | 0.3 | 0.35 | 0.88 |

For the countermatching approach all effects yield ARE's of almost 1, except for the interaction between central histology and $age_1$. For the interaction term, case–cohort does yield gains in efficiency as expected. Calibration for case–cohort leads to more efficient estimates when specificity and sensitivity increase. However, the gains in efficiency are greater for calibrated countermatching.

## 6 Applications to real data

### Wilms' tumour study

In order to asses the performance of the proposed methods with real data, we use data from the National Wilms' tumor study (NWTS) (Green et al., 1998) D' Angio et al. (1989). The data set consist of $N = 3915$ patients and the objective is to detect which prognostic variables are associated with hazard rates, where the endpoint is taken to be time to relapse or death. This data is available online in the website `http://faculty.washington.edu/norm/` and it is been used previously to evaluate other methods such as weighted case–cohort design and the stratified version with calibration (Kulich and Lin, 2004) (Breslow et al., 2009a). The prognostic variables available are histologic subtype from the central pathology reference laboratory (favourable(FH) or unfavourable(UH)) and there is a similar measure from the institution where the patient was treated. Additionally, we have stage of disease( stage), age at diagnosis (age) and tumour diameter (diameter). Following (Breslow et al., 2009a), the following Cox model is fitted for the entire cohort

$$\lambda(t) = \lambda_0(t) \exp\left(\beta_1 H + \beta_2 ag_0 + \beta_3 ag_1 + \beta_4 stg + \beta_5 dm + \beta_6 H \cdot ag_0 + \beta_7 H \cdot ag_1 + \beta_8 dm \cdot stg\right) \tag{6}$$

where $age_0$ and $age_1$ form a linear spline with separate slopes for less than 1 year age ($age_0$) and older children ($age_1$) and $stag$ is a binary variable indicating whether the subject disease is in low or high stage. This fitting yields a good estimate for the phase one variance in order to compare with the new method. Since we have two histology measures available for the entire cohort, we will implement a countermatching sampling scheme using institutional histology (known for the entire cohort). At each failure time an stratified sample $1:1$ is selected, that is, a control is selected at each case from the opposite stratum of the case. Inclusion probabilities were calculated for all subjects in the cohort and the Cox model is fitted utilizing this probabilities. The countermatching procedure was repeated $1000$ times in order to obtain a second phase variance estimate. We use the procedure describe in section 4. For each countermatched sample, we calibrate the weights by using the previously mentioned approximation for *dfbeta*'s. We assume that all variables are known for the entire cohort except central laboratory histology. We predict it by using a logistic model with the original weights (with the `svyglm` function from R Core Team (2013)). The variables included in the model are institutional histology, stage of disease, tumour diameter and $age_1$. After predicting the variables, we adjust the Cox model for the entire predicted cohort and subsequently obtain an estimation for the corresponding *dfbeta*'s and used raking calibration to adjust the weights. Then, the Cox model is fitted for the sample using calibrated weights.

## 7 Discussion

This article discusses a survey estimation approach for the Cox model under countermatched designs. The main result is the implementation of calibrated weights to fit countermatched data to the Cox model.

After calculating the marginal probabilities induced by the countermatched design, the sample can be seen as a two–phase sample and pseudolikelihood estimating equations are used to find the estimates. Weights can be calibrated by using the estimated influence functions as in Kulich and Lin (2004). This implies that we are able to use all first phase variables to improve efficiency and not just the countermatching variable as in partial likelihood. This is very useful when more than one variable is of interest, for instance, in the presence of confounders. On the other hand, if only one variable is of interest and a surrogate is used for countermatching, partial likelihood turns out to be more efficient.

We make a comparison with simple case cohort studies. Our simulation results show that for some situations of interest, for example, when we happen to have a surrogate for a rare exposure or when there is a potential partially–known confounder, we obtain large improvement in efficiency using calibrated weights. The gains are even bigger using countermatching instead of simple case–cohort. One of the reason why it happens is that countermatching provides a more informative sample. When exposure is rare, the countermatched sample has individuals from all the exposure factors, whereas a case cohort sample may end up having very few of some of them. We implement the methods for two real datasets. The first one is the Wilms' tumour study and the second is the Welsh nickel refinery study. The results are consistent with the simulation results.

## REFERENCES


Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10,** 1100–1120.

Borgan, O., Langholz, B., Samuelsen, S., Goldstein, L., and Pogoda, J. (1998). Exposure stratified case-cohort designs. *Lifetime data analysis* **6,** 39–58.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009a). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences* **1,** 32–49.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* **169,** 1398–405.

Breslow, N. and Wellner, J. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* **34,** 86–102.

Breslow, N. E. and Day, N. (1987). *Statistical Methods in Cancer Research*. IARC Scientific Publications No. 82.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

D' Angio, G., Breslow, N. E., Beckwith, J., Evans, A., Baum, H., de Lorimier, A., Ferbach, D., Hrabovsky, E., Jones, G., and Kelalis, P. (1989). Treatment of Wilms' tumour. results of the third national Wilms tumor study. *Cancer* **64,** 349–360.

Deville, J. and Sarndal, . C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87,** 376–382.

Deville, J., Sarndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88,** 1013–1020.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *The Annals of Statistics* **20,** 1903–1928.

Green, D., Breslow, N., Beckwith, J., Finklestein, J., Grundy, P., Thomas, P., Kim, T.and Shochat, S., Haase, G., Ritchey, M., Kelalis, P., and DAngio, G. (1998). Comparison between single-dose and divided dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the national Wilms' tumor study group. *Journal of clinical oncology* **16,** 237–245.

Kulich, M. and Lin, D. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99,** 832.

Langholz, B. and Borgan, O. (1995). Counter-matching: A stratified nested case-control sampling methods. *Biometrika* **82,** 69–79.

Langholz, B. and Thomas, D. (1990). Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison. *American Journal of Epidemiology* pages 169–76.

Langholz, B. and Thomas, D. (1991). Efficiency of cohort sampling designs: Some surprising results. *Biometrics* **47,** 1563–1571.

Lumley, T., Shaw, P., and Dai, J. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* **79,** 200?220.

Morabia, A., Have, T., and Landis, R. (1995). Empirical evaluation of the influence of control selection schemes on relative risk estimation: the welsh nickel workers study. *Occupational and Environmental Medicine* **52,** 489–493.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Samuelsen, S. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84,** 379–394.

Slavík, A. (2007). Product integration, its history and applications. *History of mathematics* **29,**.

Stoer, N. and Samuelsen, S. (2012). Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Analysis* **18,** 261–283.

Thomas, D. C. (1977). Addendum to *methods of cohort analysis: appraisal by application to asbestos mining* by f.d.k. liddell, j.c. mcdonad and d.c. thomas. *Journal of the Royal Statistical Society. Series A* **140,** 469–491.

van der Vaart, A. and Wellner, J. (2000). Modeling survival data: extending the Cox model. *In High dimensional probability II* pages 115–133.