

Counter-matching

BRYAN LANGHOLZ

Volume 2, pp. 1248–1254

In

Encyclopedia of Biostatistics

Second Edition

(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

© John Wiley & Sons, Ltd, Chichester, 2005

Counter-matching

Counter-matching is nested case–control study design (see **Case–Control Study, Nested**) in which a **covariate** is known on all cohort members, and controls are sampled to yield covariate-stratified case–control sets. The design is advantageous when a major analysis variable, or a correlate, is available on all cohort members and additional information is to be collected on a sample. Unbiased estimation requires the numbers of risk set members in each counter-matched sampling stratum.

The Design

Counter-matching was originally proposed as an exposure-stratified, individually matched nested case–control study method in the context of continuous failure-time (cohort) data [13, 15]. Counter-matched sets are characterized by the number of subjects m_l from each of the L sampling strata defined by the counter-matching variable. It is required that the counter-matching variable is known for all risk set members and, in addition to the case, controls are randomly sampled without replacement (see **Sampling With and Without Replacement**) from each of the sampling strata in the risk set to yield the required m_l subjects. As illustrated in Table 1, when the case is from sampling stratum 2, m_l controls are sampled from the n_l in risk set sampling stratum l except for stratum 2, from which $m_2 - 1$ controls are sampled. In the special case of two sampling strata, with one subject from each stratum (the 1:1 design), the control is sampled from the opposite sampling stratum of the case; the opposite of matching and thus motivating the name.

For grouped failure-time or simple **binary data** (multiple cases in the case–control set) counter-matching, the m_l would generally depend on total number of cases $|\mathbf{D}|$ in the study base [17]. (i.e. the design is characterized by $m_1(|\mathbf{D}|), \dots, m_L(|\mathbf{D}|)$.) The actual number of cases that fall into counter-matched stratum l , $|\mathbf{D}_l|$, is random and determines the number of controls to be sampled from stratum l , $m_l - |\mathbf{D}_l|$. This is illustrated in Table 2.

Statistical Analysis

Estimation of Rate (Odds) Ratio Parameters. The analysis of the counter-matched data must take into account the **stratification** of sampled sets. For individual **matching** (continuous time risk sets), the **partial likelihood** is based on the probability that a subject is the case given the counter-matched set and requires the control sampling probabilities. In particular, with l_j indexing the sampling stratum for subject j , the probability of drawing the counter-matched sample if j were the case is given by $\pi_j = n_{l_j}/m_{l_j} \left[\prod_{l=1}^L \binom{n_l}{m_l} \right]^{-1}$. This leads to the **likelihood**

$$\prod_{\text{sets}} \frac{r_{\text{case}}(\beta) \frac{n_{l_{\text{case}}}}{m_{l_{\text{case}}}}}{\sum_{j \in \text{set}} r_j(\beta) \frac{n_{l_j}}{m_{l_j}}}, \quad (1)$$

where $r_j(\beta) = r(Z_j; \beta)$ is the rate ratio associated with Z_j and β is the rate ratio parameter from a **proportional hazards model**. This likelihood can be fitted using standard **conditional logistic regression** software that allows for fixing a regression parameter. For instance, for the standard **loglinear model** $(n_{l_j}/m_{l_j})r_j(\beta) = \exp(Z_j\beta + \log w_j)$ where $w_j = n_{l_j}/m_{l_j}$. So, the log weight can be included in the model with fixed parameter equal to one (an *offset* in the model). Aside from this offset, analysis proceeds as in any standard conditional logistic regression analysis for individually matched case–control studies. The likelihood (1) has the usual likelihood properties so that the standard likelihood inference techniques apply, with no additional modeling assumptions other than appropriate specification of the sampling weights [9, 15]. The full asymptotic theory has been derived [4, 13] and the performance,

Table 1 Individually matched counter-matched study (one case per counter-matched set). In this example, the case is in sampling stratum 2 so m_l controls are sampled from each stratum except stratum 2 for which $m_2 - 1$ controls are sampled.

	Sampling stratum				Total
	1	2	...	L	
Cases	0	1	...	0	1
Controls	m_1	$m_2 - 1$...	m_L	$\sum m_l - 1$
Total in sample	m_1	m_2	...	m_L	$\sum m_l$
Total in risk set	n_1	n_2	...	n_L	$\sum n_l$

2 Counter-matching

Table 2 Unmatched counter-matched study (multiple cases per counter-matched set). The m_l are counter-matching design parameters representing the total number from stratum l . With $|\mathbf{D}_l|$ number of cases in stratum l , $m_l - |\mathbf{D}_l|$ controls are randomly sampled from stratum l to make a total of m_l subjects.

	Sampling stratum				Total
	1	2	...	L	
Cases	$ \mathbf{D}_1 $	$ \mathbf{D}_2 $...	$ \mathbf{D}_L $	$ \mathbf{D} = \sum \mathbf{D}_l $
Controls	$m_1 - \mathbf{D}_1 $	$m_2 - \mathbf{D}_2 $...	$m_L - \mathbf{D}_L $	$\sum m_s - \mathbf{D} $
Total in sample	m_1	m_2	...	m_L	$\sum m_l$
Total in risk set	n_1	n_2	...	n_L	$\sum n_l$

compared with other designs, has been evaluated in a number of situations [1, 7, 10, 11, 16].

For counter-matching with multiple cases per set, the likelihood requires the (control selection) probability of picking a particular counter-matched set if a set of subjects \mathbf{s} (of the same size as the actual set of cases \mathbf{D}) were the set of cases. With \mathbf{s}_l the set of subjects from \mathbf{s} in sampling stratum l , and $|s_l|$ the number of subjects in s_l , the counter-matching control selection probability is given by

$$\pi_{\mathbf{s}} = \left[\prod_{l=1}^L \frac{n_l(n_l - 1) \cdots (n_l - |s_l| + 1)}{m_l(m_l - 1) \cdots (m_l - |s_l| + 1)} \right] \left[\prod_{l=1}^L \binom{n_l}{m_l} \right]^{-1}. \quad (2)$$

This leads to the **likelihood** [17]:

$$\prod_{\text{sets}} \frac{r_{\mathbf{D}}(\beta) \left[\prod_{l=1}^L \frac{n_l(n_l - 1) \cdots (n_l - |\mathbf{D}_l| + 1)}{m_l(m_l - 1) \cdots (m_l - |\mathbf{D}_l| + 1)} \right]}{\sum_{\mathbf{s} \subset \bar{\mathcal{R}}: |\mathbf{s}| = |\mathbf{D}|} r_{\mathbf{s}}(\beta) \left[\prod_{l=1}^L \frac{n_l(n_l - 1) \cdots (n_l - |s_l| + 1)}{m_l(m_l - 1) \cdots (m_l - |s_l| + 1)} \right]},$$

where $r_{\mathbf{s}}(\beta) = \prod_{j \in \mathbf{s}} r(Z_j; \beta)$ is the product of **odds ratios** associated with the Z_j in a **proportional odds (logistic) model** (see **Logistic Regression**). Although, in general, standard software does not accommodate this likelihood, conditional logistic software can be “tricked” to estimate the odds ratio parameters when the odds model is log-linear [17]. Because of the inherently correlated structure, derivation of the asymptotic properties of likelihood (2) poses some theoretical challenges that have not yet been addressed [2]. However, limited derivation and **simulation** studies indicate that the efficiency

performance of (2) for the grouped data counter-matched is similar to that of (1) for individually matched data [17].

Estimation of Other Parameters. Methods for estimation of the cumulative baseline hazard and **absolute risk** from counter-matched data have been described [14] as well as methods for the estimation of regression parameters in the **Aalen linear model** [5]. A weighted unconditional logistic regression can be used to estimate baseline odds parameters from grouped data [17].

Examples

Crystalline Silica Exposure and Silicosis in Gold Miners. In a comparison of nested case–control study design options in an occupational cohort study of 3000 gold miners, counter-matching was compared with **random sampling** of controls [20]. A major cost component in this study was in obtaining silica exposure data from dust samples taken from the mines; a nested case–control study could have avoided much of this expense. Investigators compared random sampling and years-of-employment counter-matching of controls. The correlation between years of employment and cumulative silica dust exposure is about 0.7, and it was found that three counter-matched controls yielded the same statistical efficiency as 15 randomly sampled controls at the same cost. The situation considered in this example is typical of many **cohort studies** in which a “broad” measure (e.g. years of employment) is associated with disease and the nested case–control study is undertaken to identify better the possible causative agents (e.g. cumulative silica dust exposure). Counter-matching incorporates the cohort “broad measure” into the

sampling in order to obtain a sample that is more informative about the specific exposure, compared to random sampling.

Radiation, Hormones, and Breast Cancer in a Cohort of Japanese Atomic Bomb Survivors. A strong association of premenopausal breast cancer risk and radiation dose has been observed in the Radiation Effects Research Foundation's Life Span Study (LSS) of atomic bomb survivors [21] (*see Radiation Epidemiology*). For the Adult Health Study (AHS) cohort, a subgroup of LSS volunteers who participated in biennial clinical examinations, stored blood serum was available for 5724 women, from which estradiol levels could be measured, at some expense. The radiation-dose counter-matched study was undertaken to investigate associations with estradiol (and other hormonal and antioxidant factor) levels and radiation dose jointly on breast cancer risk. For each of the 80 premenopausal breast cancer cases, two controls were sampled with the counter-matching strata defined by radiation dose with a zero dose category and two exposure groups defined by the median of the distribution of the combined cases; that is, a control was randomly sampled from each of the (noncase) sampling strata [11, 19]. Given the actual radiation doses and a likely distribution of estradiol levels, the counter-matching design was compared with random sampling and radiation dose matching of controls. It was found that counter-matching was much more efficient than random sampling and of about equal efficiency to matching for a range of positive multiplicative radiation-estradiol **interactions**. But, unlike matching, counter-matching still allows for estimation of the radiation main effects so that a wider range of questions about the variation of breast cancer risk with radiation dose and estradiol levels can be addressed; in particular, about potential **confounding** [11]. In this study, the counter-matching variable was based on the actual exposure and the goal of the study is to investigate **effect modification** of the exposure-disease risk relationship.

Gene Susceptibility to Radiation Exposure for Second Breast Cancer Risk: The WECARE Study. The main goal of this study is to determine whether the risk of breast cancer after exposure to radiation is higher in women possessing **polymorphisms** of genes involved in double-strand break repair. The cohort consists of 31 243 women diagnosed with

breast cancer identified by five **cancer registries**. There were 801 women with asynchronous bilateral breast cancer who were the cases in this study. A cohort of women with breast cancer is advantageous for addressing the study questions for two main reasons. First, women who have had a breast cancer are likely to have a higher **prevalence** of **genotypes** that cause the disease. Second, a large percentage of the women (about 40%) underwent radiation therapy for their first breast cancer. The "scatter" from the therapeutic radiation can result in significant exposure to the contralateral breast that is often well documented in treatment records. A nested case-control study with two controls per case was dictated by cost considerations. Now, although it may be imperfect, all the cancer registries record whether radiation therapy was part of the treatment regimen (RRT+) or not (RRT-). This was used in an RRT counter-matched design in which two controls were sampled so that the case-control set would possess two RRT+ subjects and one RRT- subject. From each enrolled subject, a blood sample was obtained for the genotyping; medical treatment records were obtained (for all participants) to determine if they had had radiation treatment and, if so, the dose to the contralateral breast was determined; and the women filled out a mailed questionnaire that asked about other treatments and breast cancer risk factors. In this study, the counter-matching variable is correlated to the exposure of interest. Intuitively, there is "more variability" in radiation dose among RRT+ than RRT- subjects suggesting that 2 RRT+, 1 RRT- allocation would be more efficient for assessing radiation dose response and radiation-gene interaction than random sampling two controls. This intuition was confirmed in a simulation study comparison [3]. In this study, the counter-matching variable was based on the a dichotomous correlate of exposure and the goals of the study include characterization of the dose response for exposure and to investigate effect modification of the exposure-disease risk relationship.

Early Asthma Risk Factors Study (EARS) of In Utero and Early Life Exposures and Asthma. In a cohort study of determinants of respiratory health, over 5000 children from 12 communities and three grade levels were surveyed for "baseline" data [18]. Information collected at enrollment to the study included whether the student had ever been diagnosed with asthma, exposed to tobacco smoke *in utero* and during

childhood, and other factors that are potentially related to respiratory health. Using these baseline data, it was found that an asthma diagnosis at age five or younger was associated with maternal smoking during pregnancy (*in utero* smoke exposure) but not with environmental tobacco smoke exposure in early childhood [12]. The EARS follows up on this finding, first, to augment the smoking during pregnancy information (this was just a yes/no question in the baseline questionnaire) to assess dose–response and within-pregnancy timing of exposure and, second, to ascertain the child’s GST-T1 and GST-M1 genotypes and to assess gene susceptibility. For the purpose of this study, the cohort (or study base) consists of subjects enrolled into the longitudinal study, followed from birth to age five. Since *in utero* exposure (yes/no) information is available for the cohort members, children diagnosed with asthma at age less than five years, the cases, were counter-matched on *in utero* smoke exposure, with the number sampled from exposed and unexposed approximately equal to the number of cases within matching strata defined by community, grade, and gender. The additional maternal smoking exposure and other information was obtained in a short interview and genotype status was assessed using standard PCR methods from buccal cells collected from subjects by swabbing the inside of the mouth. In this study, “yes/no” *in utero* smoke exposure information was available on all cohort members, and it is of interest both to obtain more precise maternal smoking information to assess timing and dose-response, as well as joint effects with genetic factors. Because the counter-matching factor is fairly correlated with the number of packs smoked and other smoking information, the study has much more statistical information for inference about such factors than would a comparably sized study with randomly sampled controls [17]. In contrast with the studies described above that are individually matched, this study implements the grouped data version of counter-matching.

Design Considerations

General Considerations. Relative to random sampling, counter-matching enhances statistical efficiency for analyses involving the counter-matching variable or correlates. However, statistical efficiency is reduced for analyses of factors that are not

correlated to the counter-matched variable. Thus, counter-matching is appropriate when the study is focused on questions related to the counter-matching (generally exposure-related) factor. Situations for which there is a large efficiency gain for the counter-matching variable appear to be the situations for which there is a large efficiency loss for factors uncorrelated to the counter-matching variable. In particular, the degree of this gain/loss depends on the rarity of exposure, so that counter-matching on a rare exposure can be very advantageous for exposure-related analyses, to the great detriment of analyses of (main effects) of other factors. Whether this trade off is worthwhile depends on the specific goals of the study.

Counter-matching on an Exposure Correlate. Increased variability of exposure from the exposure-correlate stratified sampling provides some intuition for why counter-matching on an exposure correlate can increase efficiency relative to random sampling, as well as suggest a favorable allocation of subjects across the sampling strata. However, this increase in variability is tempered by the need for a weighted analysis that, in the absence of adequate correlation, works against increased efficiency [9]. Some insight into the relative efficiency of counter-matching to simple random sampling is provided in the dichotomous exposure/correlate situation with 1 : 1 counter-matching on the correlate. Under the “null” situation of no association between exposure and disease, and denoting the **sensitivity** and **specificity** of the correlate for exposure by η and γ , respectively, the asymptotic efficiency of 1 : 1 counter-matching relative to 1 : 1 random sampling is $2[\eta\gamma + (1 - \eta)(1 - \gamma)]$. Counter-matching on the correlate is more efficient when the correlate is both more (or both less) than 50% sensitive and specific. Further, if the “correlate” and exposure are independent, then the counter-matching efficiency is always less than or equal to 1; always worse than random sampling [15, 16]. This illustrates a general principle that the counter-matching factor must be “somewhat correlated” to the exposure in order to realize an efficiency gain.

Allocation of Subjects in Sampling Strata. Although analyses based on (1) or (2) with the appropriate weights are valid for any allocation of subjects in sampling strata, efficiency depends on how the sampling strata are formed and the m_i . As a general guideline, when there are more counter-matched

subjects than strata, an allocation that will yield the greatest exposure variability appears to be most desirable [3]. When the exposure or correlate has more categories than subjects to be sampled, then it is advantageous to create sampling strata that approximately results in equal numbers of cases in each stratum [11, 13, 16, 20]. Determination of the “best” counter-matched design for a given study can be addressed using asymptotic variance calculations and computer simulation.

Counter-matching and Studies of Effect Modification. Although the relative performance of case–control designs for assessing effect modification depends on the distributions of the factors involved and the relationships between these factors and disease risk in a complex way, the increased variability in one or both of the factors in a counter-matched design generally results in enhanced efficiency. An efficiency comparison of random sampling and matching or counter-matching on one of the exposure variables indicated that counter-matching was similar or superior to matched or random sampling over a wide range of situations [10]. A study of feasibility of nested case–control studies for investigation of gene-susceptibility studies compared designs using three controls per case including counter-matched designs with sampling strata defined by exposure only, family history only, and both exposure and family history, and found the latter to be the most efficient in a wide range of circumstances [1]. Other efficiency comparisons have been done in the context of the WECARE and the Radiation, Hormone, and Breast Cancer studies described in the section “Examples” [3, 19].

Other Issues

Marginal Information of the Counter-matching Variable. If the only analysis variable is a function of the counter-matching stratum variable, then counter-matching likelihood is proportional to that of the full cohort. To see this, let $Z(l)$ be a function of the counter-matching stratum l . Then, because there are m_l subjects from stratum l , contributions to (1) become

$$\frac{r_{\text{case}}(\beta) \frac{n_{l_{\text{case}}}}{m_{l_{\text{case}}}}}{\sum_{l=1}^L m_l r(Z(l); \beta) \frac{n_l}{m_l}} \propto \frac{r_{\text{case}}(\beta)}{\sum_{l=1}^L n_l r(Z(l); \beta)},$$

which is the full cohort contribution. This can be similarly shown for grouped time likelihood (2).

Counter-matching and Matching. Counter-matching is essentially the opposite of matching. Matching is a technique to create case–control sets that are *similar* in the matching factor. Counter-matching is a technique to create case–control sets that are *diverse* in the counter-matching factor. The analytic consequences of the two methods are also opposite. In particular, exact matching results in no statistical information for inference about the main effect of the matching factor, while counter-matching brings the full cohort “marginal” information for the counter-matching factor main effect into the sample. In the context of a nested case–control study, the application of the two techniques have a natural orthogonality. Matching is a natural method to incorporate information related to confounding, while counter-matching is a natural method to incorporate information related to exposure. Both methods can be used in a study by counter-matching within matching strata.

Related Designs. Two-phase exposure-stratified sampling (*see Case–Control Study, Two-phase*) differs from counter-matching in that case–control/exposure strata are sampled independently [8]. The design is appropriate for “large strata”, that is, for grouped data with sufficient numbers of cases. A comparison of the two-phase approach and counter-matching is given in [17]. An exposure-stratified **case–cohort study** has been described [6].

References

- [1] Andrieu, N., Goldstein, A.M., Thomas, D.C. & Langholz, B. (2000). Counter-matching in gene-environment interaction studies: efficiency and feasibility, *American Journal of Epidemiology* **153**, 265–274.
- [2] Arratia, R., Goldstein, L. & Langholz, B. (2005). Local central limit theorems, the high order correlations of rejective sampling, and logistic likelihood asymptotics, *Annals of Statistics*; to appear.
- [3] Bernstein, J.L., Langholz, B., Haile, R.W., Bernstein, L., Thomas, D.C., Stovall, M., Malone, K.E., Lynch, C.F., Olsen, J.H., Anton-Culver, H., Shore, R.E., Boice J.D., Jr., Berkowitz, G.S., Gatti, R.A., Teitelbaum, S.L., Smith, S.A., Rosenstein, B.S., Børresen-Dale, A.-L., Concannon, P. & Thompson, W.D. (2004). Study design: evaluating gene-environment interactions

- in the etiology of breast cancer - the WECARE study, *Breast Cancer Research* **6**, R199–R214 .
- [4] Borgan, Ø., Goldstein, L. & Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model, *Annals of Statistics* **23**, 1749–1778.
- [5] Borgan, O. & Langholz, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model, *Biometrics* **53**, 690–697.
- [6] Borgan, O., Langholz, B., Samuelsen, S.O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs, *Lifetime Data Analysis* **6**, 39–58.
- [7] Borgan, O. & Olsen, E.F. (1999). The efficiency of simple and counter-matched nested case-control sampling, *Scandinavian Journal of Statistics* **26**, 493–509.
- [8] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two stage case-control data, *Biometrika* **75**, 11–20.
- [9] Cologne, J. (1997). Counterintuitive matching, *Epidemiology* **8**, 227–229.
- [10] Cologne, J. & Langholz, B. (2003). Selecting controls for assessing interaction in nested case-control studies, *Journal of Epidemiology* **13**, 193–202.
- [11] Cologne, J.B., Sharp, G.B., Neriishi, K., Verkasalo, P.K., Land, C.E., & Nakachi, K. (2004). Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure, *International Journal of Epidemiology* **33**, 485–492.
- [12] Gilliland, F.D., Li, Y.F. & Peters, J.M. (2001). Effects of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children, *American Journal of Respiratory and Critical Care Medicine* **163**, 429–936.
- [13] Langholz, B. & Borgan, O. (1995). Counter-matching: a stratified nested case-control sampling method, *Biometrika* **82**, 69–79.
- [14] Langholz, B. & Borgan, O. (1997). Estimation of absolute risk from nested case-control data, *Biometrics* **53**, 767–774.
- [15] Langholz, B. & Clayton, D. (1994). Sampling strategies in nested case-control studies, *Environmental Health Perspectives* **102**(Suppl. 8), 47–51.
- [16] Langholz, B. & Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies, *Statistical Science* **11**, 35–53.
- [17] Langholz, B. & Goldstein, L. (2001). Conditional logistic analysis of case-control studies with complex sampling, *Biostatistics* **2**, 63–84.
- [18] Peters, J.M., Avol, E., Navidi, W., London, S.J., Gauderman, W.J., Lurmann, F., Linn, W.S., Margolis, H., Rappaport, E., Gong, H. & Thomas, D.C. (1999). A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity, *American Journal of Respiratory and Critical Care Medicine* **159**(3), 760–767.
- [19] Sharp, G.B., Neriishi, K., Hakoda, M., Suzuki, G., Akahoshi, M., Cologne, J.B., Imai, K., Eguchi, H., Nakachi, K., Key, T.J., Stevens, R.G., Kabuto, M. & Land, C.E. A Nested Case-control Study of Breast and Endometrial Cancer in the Cohort of Japanese Atomic Bomb Survivors. Research Protocol RP-6-02, RERF, 2002.
- [20] Steenland, K. & Deddens, J.A. (1997). Increased precision using counter-matching in nested case-control studies, *Epidemiology* **8**, 238–242.
- [21] Tokunaga, M., Land, C.E., Tokuoka, S., Nishimori, I., Soda, M. & Akiba, S. (1994). Incidence of female breast cancer among atomic bomb survivors, 1950–1985, *Radiation Research* **138**, 209–223.

BRYAN LANGHOLZ