Medical Research Archives. Volume 5, Issue 1.January 2017.

Flexible Cluster matching methods for Nested Case-Control Studies: Design and Statistical Considerations

Authors:
Robert J. Gallop*,
Wensheng Guo,
J. Richard Landis,
Paul N. Lanken,
Robert A. Lowe,
Jason D. Christie

Authors note:
Landis ,Professor,
e-mail rgallop@wcupa.edu
Lanken ,M.D
Pulmonary and Critical Care Division, Hospital of the University of Pennsylvania.
Lowe ,Professor,
Leonard Davis Institute of Health Economics, University of Pennsylvania and Department of Public Health and Preventive Medicine, Oregon Health and Science University.
Christie M.D.
Pulmonary and Critical Care Division, Hospital of the University of Pennsylvania

Corresponding Author.
Robert J. Gallop*,
Professor, Department of Mathematics, Applied Statistics Program, West Chester University, 25 University Avenue, West Chester, PA 19383, e-mail rgallop@wcupa.edu (610) 436-2419. Guo is Professor, Dept. of Biostatistics and Epidemiology, University of Pennsylvania

.

Abstract
A nested case-control is a case-control study within a cohort. An advantage of the nested case-control design is the number of subjects for whom outcome measures are needed are small compared to the cohort size, which is ideal when acquisition of outcomes is costly and time consuming and prevalence of cases is small. Existing selection schemes such as 1:1, 1:m, and m:n have been implemented in various studies. Upon choice of scheme, implementation requires consistent implementation across all strata, which is usually too restrictive in nested case-control studies. To maximize the usage of the available information, we propose a flexible cluster matching algorithm, in which multiple cases are matched to a group of selected controls. This may lead to unbalanced designs and induce complicated correlation structures. Mixed Effects models provide a flexible modeling framework to account for the clustering structure together with other study designs, such as repeated measures. An illustration from a nested case-control study consisting of longitudinal data is presented.

Key words: Nested Case-Control, clustered structure, stratification, mixed-effects, conditional logistic regression.

1. Introduction:

A nested case-control study is a case-control study performed within a cohort study. The nested case-control design is useful because the number of subjects for whom measurements are required is relatively small to the total size of the cohort. [1-2] This is appropriate when the relative prevalence of cases is small or the cost for measurements is expensive and time consuming.

Difficulty in finding matches is dependent on the matching scheme used. In practical applications it is common to see a 1:1 matching scheme, where every case is matched to one randomly selected control of the available controls per stratum, a 1:n matching scheme, where every case is matched to a randomly selected control of the available controls per stratum, or a m:n matching scheme, where every m cases is matched to n randomly selected controls of the available controls per stratum. With nested case-controls it may be extremely difficult to fit a conventional selection scheme to the observed sample. Consider a stratum which has 2 cases and 3 controls. Fitting a 1:2 matching scheme would require keeping 1 of the 2 cases and 2 of the 3 controls therefore, one case is not selected. With the rareness of each case, each case is valuable and the information provided by each case must be maximized. In most investigations, not utilizing information from all cases is unacceptable. Cases and controls must be identified in some systematic and periodic fashion. Regardless of how the sample nearly fits a conventional scheme, the study can not afford to wait until the sample matches the required specification. Thus, a flexible scheme is desirable. In this paper a flexible cluster matching algorithm is proposed in which multiple cases are matched to a group of selected controls. This flexible cluster matching algorithm will ensure the use of information from all available cases.

Recently, there has been several setting where investigators have adopted variable matching methods.[3-8] These studies discuss applications where the outcome is univariate such as binary response status, summary composite score such as overall mean, or survival time. When the study includes longitudinal data, resulting in multiple outcomes, the flexible matching algorithm may lead to a complex data structure; therefore, choice of a statistical framework must accommodate potential complexities associated with an unbalanced design, within subject correlation, and the correlation due to matching. A Mixed Effects models (MEM) is proposed as a general flexible framework to account for the complicated cluster structure. This provides a definitive reference for application of MEM to such structures; therefore, the flexibility of this approach to these complicated design structures needs to be highlighted.

In MEM, by modeling the random effects due to stratum, the researcher can account for the correlation due to matching, in addition to the within subject correlation, which obtains unbiased estimates. Mixed effects model do not require balanced design; therefore, varying selection of cases and controls per stratum pose no problem to the analytical method. Comparisons between cases and controls are made by the inclusion of a binary indicator of case/control designation in the MEM. Other issues such as within subject clustering and distributional concerns of the outcome measure may be encountered. MEM have been developed for a wide variety of types of outcomes. Each formulation makes assumptions about the distributional form of the outcome measure.

The next section, discusses the motivating example for the proposed algorithm and where the flexible algorithm is defined. Section 3 discusses how to analyze the resultant data using the Linear Mixed Effects Model (LMM) and the Generalized Linear Mixed Effects Model (GLMM). Illustrated for the reader is how the analysis of selected cases and controls identified by the proposed matching algorithm through MEM, is identical to the results acquired by analysis of 1:1 matched data for normal outcomes through a Paired t-test. A similar finding holds for 1:1 matched binary data analyzed through the MEM compared to the standard analytic approach of Conditional Logistic Regression. In section 4, an example

applying MEM for data from selected cases and controls via the proposed flexible algorithm is provided. Some concluding remarks are made in Section 5.

2. Flexible Matching Algorithm

2.1 Motivating Example

What motivated our investigation of these ideas is a nested case-control study coordinated by the University of Pennsylvania Medical Center, on Acute Respiratory Distress Syndrome (ARDS). The goal is to determine biomarkers that serve as predictors of the onset of ARDS among seriously injured trauma patients. Eligibility in the study requires a minimum Injury Severity Score (ISS) and admission to the intensive care unit (ICU). During the five day period post enrollment a patient's status is closely monitored for the onset of ARDS. Various biomarkers are acquired and tracked over this five day period. For inclusion in the analysis sample of the cohort study, two situations may occur during this five day period:

- Patients may develop ARDS.
- Patients may remain ARDS-free and in the ICU the full 5 day observation period

Prevalence of cases is rare and biomarker measurements are quite expensive; therefore, a nested case-control design is used. The number of cases and controls may vary over the nested structure of the design; therefore, care must be observed in choice of a matching algorithm. Choice of matching must guarantee the selection of all cases.

Subsample selection will be conducted for all cases, patients who develop ARDS, and controls, patients who are ARDS-free and remained in ICU the full 5 days. Patients who are ARDS-free and are discharged early from the ICU are excluded from the matching process. Similarly, patients who are ARDS-free but die during the five day observational period are excluded from the matching process. Therefore, the controls will have a minimum of 5 days worth of data. For the cases, these patients are followed for an additional 4 weeks upon development of ARDS. The Injury severity score (ISS) serves as an indicator of a patient's initial severity; therefore, it is believed higher ISS may serve as a potential indicator of the development of ARDS[9]. ISS is stratified into four levels: Low (16 - 19), Mid-Low (20-29), Mid-High (30-39), and High (40 or more). Not only may the ISS score be confounded with determination of cases and controls, but the number of cases to controls may be differential across the four strata of ISS. When employing a 1:1 matching scheme or a 1:n matching scheme, there may not necessarily be a control for each case. Similarly, for lower ISS scores there may be substantially more controls as compared to cases. The following table illustrates the breakdown of 140 eligible patients enrolled during the three year period.

Table 1. Cases and Controls over Levels of ISS and Period of time

Level of ISS	Periods			Total
	Year 1	Year 2	Year 3	
Low	3 (6)	7 (6)	5 (4)	15 (16)
Mid-Low	8 (15)	10 (20)	11 (21)	29 (56)
Mid-High	3 (5)	1 (2)	0 (0)	4 (7)
High	1 (1)	2 (4)	3 (2)	6 (7)
TOTAL	15 (27)	20 (32)	19 (27)	54 (86)

Note: Controls are in parentheses

As indicated in Table 1, the number of cases to control varies over each stratum. For the High ISS strata during Period 1, there is only one case and one control; therefore, implementation of a

conventional 1:2 matching scheme will result in the exclusion of one case due to not enough controls. Similarly during period 3, there are 3 cases and 2 controls in the High ISS level; therefore implementation of a 1:2 conventional matching method will exclude 2 cases. These cases are rare, and therefore, the researcher needs to maximize the use of every case. Thus, a more flexible matching scheme is needed.

2.2 Cluster Matching Algorithm

Proposed is an experimental design that has the flexibility to accommodate the situation where the ratio of cases and controls may vary across strata. Selection is as follows:

(a) Per stratum, take all selected cases. Let n_k denote the number of cases in stratum k.
(b) Assign r as the optimal number of controls desired per case.
(c) Let m_k denote the number of available controls in stratum k.
(d) If $m_k > r \times n_k$ then a subsample if size $r \times n_k$ is randomly selected from the m_k available controls.
(e) If m_k is less than $r \times n_k$ then all m_k available controls are selected.

Under the situation where the desired number of controls per case is available for all strata, this flexible algorithm reduces to the 1:r conventional scheme.

This flexible algorithm is applied for the ARDS study. Let the optimal number of controls per case, r, equal 2. For the k^{th} stratum, if there are n_k cases and m_k controls, we propose the following: (i) if $m_k \le 2n_k$ then take all controls, (ii) if $m_k > 2n_k$ then take a random $2n_k$ controls. Another facet for the ARDS design is the inclusion of time period as a stratification variable.

Adjustment for time period may account for annual effects and hospital staffing issues. Time period is stratified annually for the three year period as follows: the first year (July 1999 - June 2000), the second year (July 2000 – June 2001), and the third year (July 2001 – June 2002).

The major analytical consideration with the ARDS data is the repeated daily measurements. This together with the clustered structure induced by the matching algorithm can be handled by the mixed effects model framework in a unified way, in which random effects are used to model the proper correlations between observations and account for the cluster effect. Thus, this proposed design includes all conventional schemes and MEM provides the means to answer our hypotheses of interest.

3. Analysis using Mixed Effects Models

Mixed effects models are used when outcome responses are clustered. Selection of matched cases and controls are stratified across designated factors. Per stratum, the matched cases and controls form a single cluster; therefore, any analysis must recognize and model the correlation within each cluster, as well as the variability between clusters. MEM provide the flexible analytical framework for the case-control data from implementation of the flexible matching algorithm. Cluster size may vary between clusters. When the study includes longitudinal data, the proposed method can address the within subject correlation together with the correlation due to matching in a unified framework. In addition, outcome measures can be continuous or discrete.

3.1 Linear Mixed Effects Model

Using the Mixed effect methodology, the use of Linear Mixed Effects models (LMM) for continuous outcomes is proposed. The stratification variables are incorporated as random effects. The mixed effects model proposed is as follows:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + ... + \beta_k x_{kij} + b_i + \varepsilon_{ij}, \text{ for i=1,2,...,N and j=1,2,...,n}_i \quad (3.1)$$

In equation (3.1) $x_{1ij}, x_{2ij}, ..., x_{kij}$ are the k predictors for observation j in stratum i, β_0, β_1, …, β_k are the k-fixed effects, b_i is the random effect associated with stratum i, ε_{ij} is the random error for observation j in stratum i, N corresponds to the number of strata, and

$j = 1, 2, ..., n_i$ indicates the subjects in the i^{th} cluster. As described by Laird and Ware[10], the assumption for the random elements are as follows:

$$b_i \sim NIID(0, \sigma_b^2)$$
$$\varepsilon_{ij} \sim NIID(0, \sigma^2) \quad (3.2)$$

The term NIID indicates each term is independent and identically distributed with a normal distribution. An additional assumption is b_i and ε_{ij} are independent. From the above assumptions it can be observed that the marginal response and variance are:

$$E[Y_{ij}] = \beta_0 + \beta_1 x_{1ij} + ... + \beta_k x_{kij}; V[Y_{ij}] = \sigma^2 + \sigma_b^2 \quad (3.3)$$

Observe that the marginal response, $E[Y_{ij}]$, and variance, $V[Y_{ij}]$, do not depend on the random effects, b_i . Thus, for inferences on the outcome measures, Y_{ij}, the random effects have basically ``dropped" from the model, although the inclusion of the random effects properly estimated the variance in outcome. Ignoring the structure of the clustering could result in the underestimation of response variance, which may result in falsely deflated levels of significance. Thus, the LMM provides us a platform to analyze matched cases and controls where outcomes are continuous and the effect of the stratification variable has been removed from the model.

In general, the LMM is represented in vector-matrix form as follows:

$$\mathbf{Y = X\beta + Zb + \varepsilon} \quad (3.4)$$

where **Y** is the vector of outcomes, **X** and **Z** are the design matrices for the fixed effects and random effects respectively, $\boldsymbol{\varepsilon}$ is the vector of random errors, and $\boldsymbol{\beta}$ and **b** are the vectors of fixed effects and random effects respectively[11]. For longitudinal data $\boldsymbol{\varepsilon}$ is a matrix of random errors with Multivariate Normal Distribution (MVN) with mean vector of **0** and covariance matrix represented by **R**. Laird and Ware[10] set down the estimates of the fixed effects and random effects as follows:

$$\boldsymbol{\beta} = \left[\left(\mathbf{X^T V^{-1} X} \right)^{-1} \mathbf{X^T V^{-1}} \right] \mathbf{Y}$$
$$\mathbf{b = GZ^T V^{-1} (Y - X\beta)} \quad (3.5)$$
$$\text{where } \mathbf{V = ZGZ^T + R}$$

The superscript T denotes transpose of the corresponding matrix. Modeling of an appropriate covariance structure is essential for the inferences of the hypotheses to be valid[12].

3.2 Paired t-test and Linear Mixed Effects Model

In the case with 1:1 matching with one outcome measure, LMM is equivalent to a paired t-test, which is commonly used. Consider when there are n strata with two observations per stratum. The LMM we assume is as follows:

$$Y_{ij} = \mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \mathbf{Z} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \varepsilon \quad (3.6)$$

where **X** is the design matrix for the fixed effects, **Z** is the design matrix for the random effects. The vector **Y** is the vector that contains the responses and $\boldsymbol{\varepsilon}$ is the vector that contains the random errors. Under this situation **X** and **Z** are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \end{bmatrix}; \mathbf{Z} = \begin{bmatrix} 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \end{bmatrix} \quad (3.7)$$

The assumed covariance structure, denoted by **V**, which has dimensions $2n \times 2n$, is as follows:

$$\mathbf{V}=\begin{bmatrix}\begin{bmatrix}\sigma_b^2+\sigma^2 & \sigma_b^2\\ \sigma_b^2 & \sigma_b^2+\sigma^2\end{bmatrix} & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \begin{bmatrix}\sigma_b^2+\sigma^2 & \sigma_b^2\\ \sigma_b^2 & \sigma_b^2+\sigma^2\end{bmatrix}\end{bmatrix} \quad (3.8)$$

Let $a=\left(\sigma_b^2+\sigma^2\right)/\left[\sigma^2\left(2\sigma_b^2+\sigma^2\right)\right]$ and $b=\left(-\sigma_b^2\right)/\left[\sigma^2\left(2\sigma_b^2+\sigma^2\right)\right]$, then the inverse of $\mathbf{V}$ denoted by $\mathbf{V^{-1}}$ is as follows:

$$\mathbf{V^{-1}}=\begin{bmatrix}\begin{bmatrix}a & b\\ b & a\end{bmatrix} & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \begin{bmatrix}a & b\\ b & a\end{bmatrix}\end{bmatrix} \quad (3.9)$$

The above vector-matrix equations can be used to derive estimates for the vector of fixed effects and covariance matrix for the estimated vector of fixed effects. Estimates are as follows[12]:

$$\boldsymbol{\beta}=\left[\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\right]\mathbf{Y} \quad (3.10)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}=\mathrm{Var}\widetilde{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

Mclean and Sanders[13] have shown that when estimating a linear combination $\mathbf{h}^T\boldsymbol{\beta}$ of the fixed effects, the distribution of the following tern:

$$\frac{\mathbf{h}^T\boldsymbol{\beta}}{\sqrt{\mathbf{h}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{h}}} \quad (3.11)$$

is approximated by a t-distribution with (n-1) degrees of freedom, where n is the number of stratum. Our main interest is:

$$H0:\hat{\beta}_1=0\ versus\ H1:\hat{\beta}_1\neq 0 \quad (3.12)$$

Thus, setting $\mathbf{h}^T$ equal to $\begin{bmatrix}0 & 1\end{bmatrix}$ will test the hypothesis of interest. By inserting $\mathbf{h}^T$ in equation (3.11), the following are derived:

$$\mathbf{h}^T\boldsymbol{\beta}=\frac{1}{n}\left[(y_{11}-y_{12})+(y_{21}-y_{22})+\cdots+(y_{n1}-y_{n2})\right]=\frac{1}{n}\sum_{i=1}^{n}\left(y_{i1}-y_{i2}\right), \quad (3.13)$$

$$\mathbf{h}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{h}=\frac{2}{n}\sigma^2 \quad (3.14)$$

The test statistic is as follows:

$$t=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_{i1}-y_{i2}\right)}{\sqrt{\frac{2}{n}\sigma^2}}\ \text{where t}\sim t_{n-1} \quad (3.15)$$

If one analyzed this data as paired data, one would define the difference per pair as follows:

$$D_i=(Y_{i1}-Y_{i2}),\ \text{for i = 1,2, ..., n} \quad (3.16)$$

The mean difference, $\overline{D}$, and the variance of $\overline{D}$, are as follows:

$$\overline{D}=\sum_{i=1}^{n}\frac{D_i}{n}=\sum_{i=1}^{n}\frac{\left(Y_{i1}-Y_{i2}\right)}{n}. \quad (3.17)$$

$$Var(\overline{D})=Var\sum_{i=1}^{n}\frac{D_i}{n}=\frac{Var\sum_{i=1}^{n}D_i}{n^2}=\frac{\sum_{i=1}^{n}Var(Y_{i1}-Y_{i2})}{n^2}=$$

$$\frac{n(Var(Y_{i1})+Var(Y_{i2})-2Cov(Y_{i1},Y_{i2}))}{n^2}=\frac{\sigma_b^2+\sigma^2+\sigma_b^2+\sigma^2-2\sigma_b^2}{n}=\frac{2\sigma^2}{n}. \quad (3.18)$$

Thus, the test statistic for the paired t-test is as follows:

$$t=\frac{\overline{D}}{\sqrt{Var\overline{D}}}=\frac{\sum_{i=1}^{n}\frac{y_{i1}-y_{i2}}{n}}{\sqrt{\frac{2\sigma^2}{n}}}, \quad (3.19)$$

where t ~ t_{n-1}. Direct comparison of equation (3.15) and equation (3.19) illustrate that under

the simple structure of a continuous outcome with two measurements per strata, the LMM mode is equivalent to the paired t-test.

4. Application to the ARDS study

For this inspection with the ARDS study, the primary outcome is the carbonyl measurement of blood samples. This outcome measures the amount of protein in the blood. During the five day observation period after enrollment in the Trauma cohort both cases and controls have carbonyl measurements derived from available blood samples. Blood samples per patients were acquired as much as possible. Date and time are recorded on all blood samples. During the derivation process certain blood samples were unable to be processed due to contamination or other mishandling processes. Therefore, the number of carbonyl measurement may be widely varied across all patients. For the ranged from 2 to 24 over the five day period. With this outcome, the focus of interest is on whether the linear change over time is differential between cases and controls. An a priori alpha-level of 0.01 is set for all analyses.

4.1 Application of the Flexible matching scheme:

There are 126 patients from the available 140 who have carbonyl measurements. Of the 126 patients, 54 are cases and 72 are controls after application of the flexible matching scheme. Now applying the MEM to this data, we must recognize the clustering due to the matching, as well as the cluster within patient due to repeated nature of the data. Figure 1 shows the individual spaghetti plots for cases and controls, overlayed with a smoothed profiles over the 5 days for cases versus controls.

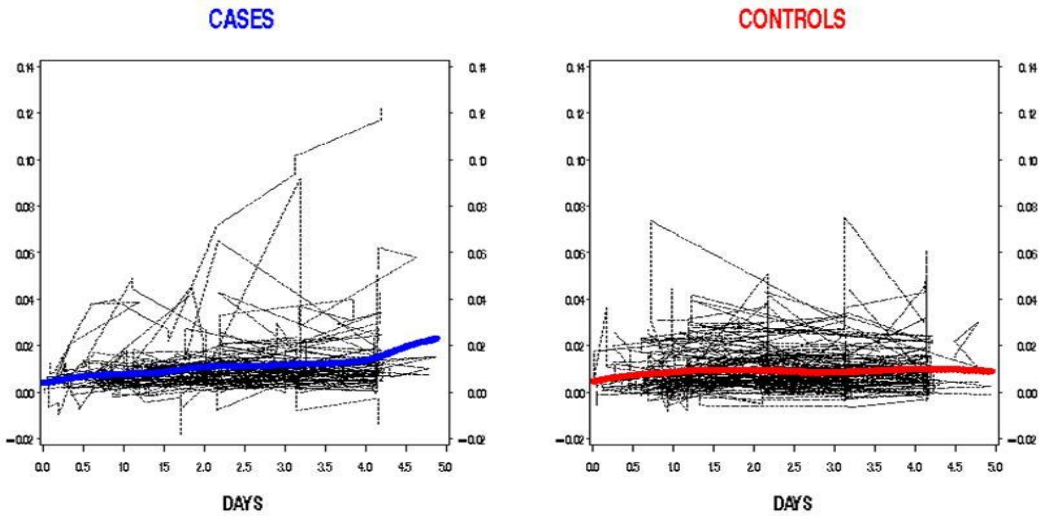

Figure 1. Individual Profiles of carbonyl measurements over the 5 days by CASE versus CONTROL

As seen in Figure 1 there is substantial heterogeneity within Cases and Controls. This heterogeneity may be attributable to the matching with respect to ISS level and the year of the study. The smoothed profiles appear to have a somewhat monotonic pattern, especially over the last three days. This provides the motivation to inspect differences in the rates of change between cases and controls. To properly model this data, the analysis needs to take into

account the within subject correlation associated with the repeated measures, as well as the clustering associated with the matching. Applying the MEM, which accommodates these components, a highly significant difference in the rate of change in the carbonyls over the five days, where carbonyl levels change on-average 2.10×10^{-3} units per day more than the controls ($F(1,161)=9.22$, $p=0.003$) is observed.

Figure 2. Distribution of the Subject-specific rates of change per day

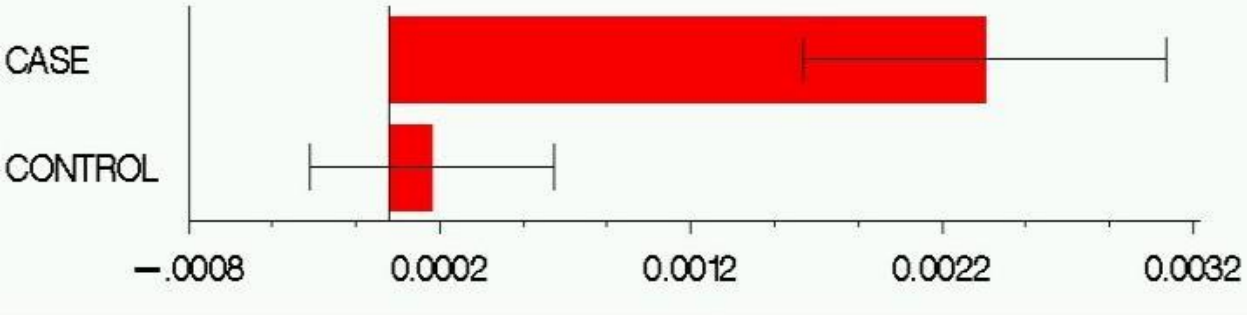

Figure 2 shows the distribution (mean + standard error bars) of the subject-specific slopes by group. Despite the substantial heterogeneity illustrated in Figure 1, the subject specific slope, consistent with the smoothed profiles, shows an on-average 10-fold increase in the rates of change per day for Cases compared to Control. For the controls, we see the standard error bars contain 0; therefore, indicating relatively no statistically significant change in carbonyls over the five day period, whereas for cases, we see a significantly positive rate of change in carbonyls over the five day period. The error bars do not overlap, which is consistent with the statistically significant finding between groups in rate of increase per day, with Cases, on-average, showing positive increase in carbonyl scores per month.

4.2 Ignoring the clustering due to the matching

If the clustering due to matching was ignored, then this source of variability in the outcome would be included in the modeling due to the repeated data within patients. Applying the MEM, a non-significant difference in the rate of change in the carbonyls over the five days, where carbonyl levels change on-average 1.88×10^{-3} units per day more than the controls ($F(1,162)=6.53$, $p=0.012$) is observed. So by failing to account for the variability due to the matching, the differential rate of change between cases and controls is underestimated and no longer have a significant finding.

4.3 Application of a fixed 1:2 matching scheme

By implementing a 1:2 matching scheme, the researcher is forced to exclude certain cases due to insufficient controls. In addition, the matching is subjective to the random assignment of the available cases and control. Applying this 1:2 standard scheme to the available ARDS data, our sample is reduced to 33 cases and 66 controls. Referring to Table 1, the case and control in the High ISS Stratum during year 1 are not included in our analysis. Similarly two cases within the High Stratum during year 3 are not included in the analysis. The researcher is losing critical information due to the lack of flexibility of the standard 1:2 matching scheme with this nested cohort study.

Applying the MEM, a non-significant difference is observed in the rate of change in the carbonyls over the five days, where the on-average carbonyl levels change per day is 2.18×10^{-3} units per more for the cases compared to the controls ($F(1,116)=6.72$, $p=0.011$). So by failing to use all available data as determined by the proposed flexible matching scheme, the researcher no longer has a significant finding at the alpha-level of 0.01. The statistical

significance is right above our alpha-level of 0.01; therefore, we immediately question whether a larger sample size could achieve a higher level of statistical power, to result in the same observed difference in the rate of change being statistically significant. So possibly, by throwing away data at the price of fitting a standard algorithm, may have contributed to the lack of statistical significance.

4.4 Comparison of the results across the three models.

Table 2 contains the rates of change and the variance estimates within the MEM from our three models discussed in the previous sections.

Table 2. MEM results for the ARDS study

Model	Slope Estimate for CASES	Slope Estimate for Controls	Difference in Rate of Change	Variance attributable to matching	Variance of random Slopes	Error Variance
Flexible Matching	0.242 (0.052)	0.032 (0.046)	0.210 (0.069)	0.514 (0.298)	0.106 (0.015)	0.512 (0.021)
Not Accounting for Matching	0.221 (0.055)	0.033 (0.049)	0.188 0.073		0.123 (0.017)	0.537 (0.022)
Forced 1:2 Matching	0.254 (0.067)	0.036 (0.050)	0.217 (0.083)	0.077 (0.048)	0.121 (0.019)	0.422 (0.019)

Note: Standard errors are in parentheses. Estimates are based on a MEM with the outcome increased by a factor of 100.

Notice that the slope estimates are consistent across the three approaches, but the standard errors associated with the slope estimates are higher when the model does not account for the clustering attributable to the matching and when a 1:2 matching algorithm is forced. When the model excludes the source of variability attributable to matching, a portion of this variance is absorbed into the error variance, resulting in increases standard errors. When a 1:2 algorithm is forced, the effective sample size differ between the analyses, where the 1:2 algorithm effective sample size is smaller, which contributes to the increase in the standard errors.

5. Summary

Preliminary investigation for the ARDS study has indicated conventional schemes are not applicable for selection of case-controls nested within cohorts; therefore, flexibility of matching is required. While the flexible matching scheme accommodates the nuances in the allocation of case and controls, a flexible modeling approach was suggested for the analysis of data. The flexible modeling approach recommended was Mixed Effects Models. Mixed effects models adjusts for the confounding effect due to stratification but also offer the flexibility to model a variety of outcomes. In addition, for studies including longitudinal data, the proposed method can handle the within subject correlation together with the correlation due to strata in a unified framework. Linear Mixed Effects models are used for continuous outcomes. With Generalized Linear Mixed Model, a variety of outcomes can be modeled such as binary, count, nominal, and ordinal data. One limitation of this approach is based on the distributional assumption associated with the MEM. While conditional models treat the matching as a nuisance variable, the MEM models the matching through random effects, which accounts for the within correlation within the matching and the between variability across the levels of the matching variables. Therefore, model diagnostics must be implemented to ensure the assumptions associated with the random effects, namely normality of the random effects with mean of zero, are being met.

Conventional matching schemes are special cases of this proposed flexible matching scheme. Likewise, as illustrated in sections 3.2, conventional modeling methods for matched data are special cases of the Mixed Effects Model.

In summary, the flexible matching scheme under the Mixed Effects model framework provides the setting to answer a variety of hypotheses while ensuring the confounding effect of cluster (i.e. matching) is removed for proper interpretation.

Acknowledgment

This research was supported in part by an NHLBI award HL 60290 (Aaron Fisher) and an NCI RO1 CA 84438 (Guo).

References:

1. Beaumont J.J., Steenland K., Minton A, and Meyer S. A computer program for incidence density sampling of controls in case-control studies nested within occupational cohort studies. *American Journal of Epidemiology* 1989; **129**: 212-219.
2. Langholz B. and Thomas, D.C. Nested case-control and case-cohort methods of sampling from a cohort: A Critical Comparison. *American Journal of Epidemiology* 1990; **131**: 169-176.
3. Cepeda M.S., Boston R., Farrar J.T., and Strom, B.L. Optimal matching with a variable number of controls vs. a fixed number of controls for a cohort study: trade-offs. *Journal of Clinical Epidemiology* 2003; **56**: 230-237.
4. Duffy S.W., Rohan T.E., Kandel R., Prevost T.C., Rice K., and Myles J.P. Misclassification in matched case-control study with variable matching ratio: application to a study of c-erbB-2 overexpression and breast cancer. *Statistics in Medicine* 2003; **22**: 2459-2468.
5. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 2000; **56**: 118-124.
6. Saunders C.L. and Barrett J.H. Flexible Matching in Case-Control Studies of Gene-environment interactions. *American Journal of Epidemiology* 2004; **159**: 17-22.
7. Schroder M., Husing I., and Jockel K.H. An implementation of automated individual matching for observational studies. *Methods of Information in Medicine* 2004; **43**: 516-520.
8. Van der Tweel I., Van Noord, P.A.H. Sequential analysis of matched dichotomous data from prospective case-control studies. *Statistics in Medicine* 2000; **19**: 3449-3464.
9. Wolfe R., McKenzie D.P., Black J., Simpson P., Gabbe B.J., and Cameron, P.A. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. *Journal of Clinical Epidemiology* 2006; **59**: 26-35.
10. Laird N.M., and Ware T.H. Random effect models for longitudinal data. *Biometrics* 1982; **38**: 963-974.
11. Verbeke, G., & Molenberghs, G. *Linear Mixed models for Longitudinal Data* 2000, New York: Springer-Verlag.
12. Littell R.C., Milliken G.A., Stroup W.W., and Wolfinger R.D. *SAS System for Mixed Models* 1995, Cary NC: SAS Institute Inc.
13. McLean R.A., and Sanders W.L. Approximating degrees of freedom for standard errors in mixed linear models. *Proceeding of the Statistical computing section, American Statistical Association, New Orleans* 1988; 50-59.
14. Nelder JA, and Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society A* 1972; **135**: 370-384.
15. Diggle P.J., Liang K.Y., and Zeger S.L. *Analysis of Longitudinal Data* 1994; Oxford: Clarendon Press.
16. Wolfinger R.D. Towards Practical Application of Generalized Linear Mixed Models, In: Mark B. and Friedl H, eds. *Proceedings of the 13th International Workshop on Statistical Modeling* 1998; 383-395.
17. Breslow, N.R. and Clayton, D.G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**: 9-25.
18. Wolfinger R.D. and O'Connell N. Generalized linear mixed models: a pseudolikelihood approach. *Journal of Statistical Computation and Simulation* 1993; **48**: 233-243.
19. Littell R.C., Milliken G.A., Stroup W.W., Wolfinger R.D, and Schabenberber, O. *SAS System for Mixed Models*, 2nd Edition2006, Cary NC: SAS Institute Inc.

20. Hedeker D., and Gibbons R.D. A random effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**: 933-944.
21. Stokes M.E., Davis C.S., and Koch G.G. *Categorical Data Analysis Using the SAS System* 1995, Cary, NC: SAS Institute Inc.
22. Breslow N.R., and Day N.E. Cancer Research, Volume 2: *The Design and Analysis in Cohort Studies* 1987; Lyon, France: Lyon/International Agency for Research on Cancer.

