Empirical Evaluation of Complex Epidemiologic Study Designs: Workplace Exposure and Cancer

David C. Deubner, MD, MPH
H. Daniel Roth, PhD
Paul S. Levy, ScD

Objective: *To test whether a frequently used cohort-nested case-control study design exaggerated exposure-response relationships because of unrecognized study design bias. Our aim was to evaluate empirically the performance of this complex study design.* ***Methods:*** *We applied the design from one such study to a closely related cohort using randomly selected probands as cases. Values for average exposures were assigned to probands equal to, greater than, and less than those assigned to controls (matches).* ***Results:*** *Under certain lag scenarios, the nested study design produced higher average exposure in probands compared with their matches, even when this was clearly not the case.* ***Conclusions:*** *Empirical evaluation demonstrated that the study design produced a biased case-control lagged exposure difference under the null hypothesis and could not distinguish qualitatively between null and alternate hypotheses. Empirical evaluation provided a useful check on results generated from a complex study design. It gave useful insight into the behavior of the index study design that was not otherwise readily deducible.* (J Occup Environ Med. *2007;49:953–959*)

Our objective is to report on an empirical evaluation of a complex study design that has been used repeatedly[1–4] with subtle variations. This study design is so complex that its behavior is generally not correctly predicted a priori by highly qualified persons. We found that empirical evaluation helped us to understand the behavior of the study design and to investigate reasons for the study design behavior that were not initially discernible when considered from a theoretical point of view.

We evaluated the study design using a single "index" study.[3] The index study conducted a cohort-nested case-control examination of the relationship between lung cancer mortality and estimates, unlagged and lagged, of worker beryllium exposure. This study was performed because a previous study[5] found a significantly elevated excess rate of lung cancer among beryllium workers with tenure less than 1 year, a pattern recognized previously.[6] Relatively high intensity exposure in the 1940s was postulated as the explanation for the excess lung cancer in short-tenure beryllium workers. The index study reconstructed beryllium exposure estimates to examine directly the relationship between lung cancer and beryllium exposure, characterized respectively by beryllium industry work tenure and cumulative, average, and maximum (highest average exposure job worked) exposure. The index study found no significantly greater exposure in cases in any exposure category when expo-

From Brush Wellman, Inc (Dr Deubner), Elmore, Ohio; Roth Associates, Inc (Dr Roth), Rockville, Maryland; and RTI International (Dr Levy), Research Triangle Park, North Carolina.

Address correspondence to: David C. Deubner, MD, MPH, Brush Wellman, Inc, 14710 West Portage River South Road, Elmore, OH 43416; E-mail: david_deubner@brushwellman.com.

DOI: 10.1097/JOM.0b013e318145b28d

sure was unlagged. When exposure was lagged 10 or 20 years, the results were strikingly different, with cases now having significantly greater cumulative, average, and maximum exposures. A reanalysis of this study with use of the same data with a different design found no case-control differences, lagged or unlagged.[7]

Here, we provide choice illustrative examples to show how the index study design performs in analyzing exposure-effect relationships. To accomplish this, we were fortunate to have had the National Institute for Occupational Safety and Health (NIOSH) graciously supply us with the index study database on 142 cases of lung cancer occurring through 1992 and 710 controls. NIOSH also provided us with data on the entire cohort of 3569 workers from which the cases and controls were derived, followed for mortality through 1988. This group is hereafter referred to as the "experimental population."

Our illustrations were designed to address two objectives. The first was to ascertain how the study design behaved when "cases" were random picks from the population and there were no differences in intensity of exposure in any subject. For this, we used the experimental population. From this population, we selected "probands" at random and also selected controls, or "matches" for these probands. We assigned all persons identical values for average exposure, and from these data we calculated unlagged and lagged time worked, and cumulative exposure and lagged average exposure as did the index study.

The second was to determine how the study design behaved when there were differences in exposure intensity between probands and their matches. For this, we used the same probands and matches previously sampled but assigned probands both higher and lower average exposure values relative to the matches.

Materials and Methods

Index Study

From the index study paper[3] and conversation with the authors, we acquired detailed information about the index study design, which used one of eight cohorts of beryllium workers analyzed in a previous study.[5] There were 142 persons who died of lung cancer identified in this cohort. Each case was matched with 5 controls (710 total) selected at random from other cohort members who met the following criteria: same sex (almost all male), same race (almost all white), case could not serve as own control, age attained in the cohort (age at death or age at censor) of control greater than or equal to the age at death of the case, and age at hire of the control equal to or less than the age of death of the case.[8]

Using a job exposure matrix developed for the index study,[9] the index study authors calculated a beryllium exposure profile for all subjects. They then calculated exposure variables for each subject: length of employment, cumulative exposure, average exposure, and maximum exposure. With an assumption of lung cancer latency of 10 or 20 years, they then re-calculated the exposure variables with exclusion in subjects in each case-control set of exposure after the age of death of the case minus 10 years, and minus 20 years (10- and 20-year lagging of exposure). When the lagging of exposure produced a zero value, 0.1 was substituted so that geometric means could be calculated.

The authors compared the case and control geometric means of these 12 variables (3 lagging categories [0, 10, and 20 years] for each of the 4 exposure measures [length of employment, average exposure, cumulative exposure, and maximum exposure]). They used conditional logistic regression to estimate the statistical significance of the differences.

The Experiments

The experimental population data set supplied by NIOSH included date of birth, date of hire, date of employment termination, and date of censor (date of death if deceased, and date of follow-up for mortality if not deceased before the date of mortality follow-up). Using these variables for each subject, we calculated age at hire (date of hire minus date of birth), age at termination (date of termination minus date of birth), age at censor (date of censor minus date of birth), and age at death (date of death minus date of birth).

From this cohort, we drew samples as follows: 142 (the number of lung cancer cases in the index study) subjects, labeled "probands" were selected from the experimental population by random sampling without replacement. The probands would constitute a random sample of the experimental population, unbiased with respect to any variable. For each of these probands, we then identified all subjects, except the proband, who met the following criteria: age at censor of subjects greater than or equal to the age at censor of the proband, and age at hire less than the age at censor of the proband. From these pools, we selected five subjects (matches) at random without replacement. We performed this process after replacement 10 times, for a total of 10 sets of 142 probands and their respective 710 matches. The demographic characteristics of the experimental population and the proband and match sets are compared in Table 1.

To examine the behavior of the study design when all subjects had identical exposure intensity, we assigned all probands and matches identical values (20 units) for average exposure, and from this data we calculated length of employment, cumulative exposure, and average exposure, both unlagged and lagged, as did the index study.

We calculated the proband's unlagged length of employment as age at termination minus age at hire. The

TABLE 1
Experimental Population: Demographic Features of Samples From Experimental Population Derived From 10 Trials*

Demographic Variables†	Population 3,569	Probands 1,420	5 Matches 7,100
Age at censor	62.2	62.6 (0.2869)	72.7 (<0.001)
Date of birth	1917.4	1916.8 (0.2383)	1909.4 (<0.001)
Date of hire	1949.2	1949.1 (0.5521)	1946.3 (<0.001)
Date of termination	1954.6	1954.6 (0.9142)	1951.6 (<0.001)
Date of censor	1980.9	1980.6 (0.5177)	1982.0 (<0.001)
Age at hire	31.8	32.3 (0.2621)	37.0 (<0.001)
Age at termination	37.1	37.8 (0.1818)	42.2 (<0.001)
Time from hire to censor	31.6	31.6 (0.8416)	35.8 (<0.001)
Time employed	5.3	5.5 (0.5233)	5.2 (0.3293)

*In each trial, 142 probands were selected with use of random sampling without replacement, and 5 matches were selected for each proband with use of random sampling without replacement from eligible members of the cohort.
†Arithmetic means for demographic variables (*P* value of comparison with column on left).

TABLE 2
Experimental Population: Results of 10 Trials, With Combinations of Unlagged (0) and Lagged (20) Years, With Equal and Unequal Average Exposures Assigned to Probands and Matches*

		Length of Employment		Cumulative Exposure		Average Exposure	
Average Exposure	Lag in Yrs	Probands	Matches	Probands	Matches	Probands	Matches
P = M	0	0.8	0.7 ($P = 0.051$)	16.8	14.8 ($P = 0.051$)	20	20
P = M	20	0.5	0.3 ($P < 0.001$)	5.7	2.6 ($P < 0.001$)	6.9	3.6 ($P < 0.001$)
P > M	0	0.8	0.7 ($P = 0.051$)	33.6	14.8 ($P < 0.001$)	40	20
P > M	20	0.5	0.3 ($P < 0.001$)	9.9	2.6 ($P < 0.001$)	12.0	3.6 ($P < 0.001$)
P < M	0	0.8	0.7 ($P = 0.051$)	16.8	29.6 ($P < 0.001$)	20	40
P < M	20	0.5	0.3 ($P < 0.001$)	5.7	4.1 ($P < 0.001$)	6.9	5.7 ($P = 0.020$)

*Figures are the geometric means with substitution of 0.1 for zero values in the lagged analyses.
P indicates probands; M, matches.

match's length of employment was calculated as the lesser of the match's age at termination minus age at hire or the proband's age at censor minus the match's age at hire. We calculated cumulative exposure as average exposure times the length of employment. We then calculated the natural logarithm of the length of employment, cumulative exposure, and average exposure and used these values to calculate the means for probands and matches (Table 2). We then used conditional logistic regression to determine the statistical significance of the proband-match differences.

For lagging, we calculated proband length of employment as the lesser of the proband's age at termination minus age at hire or (age at censor minus 20 years) minus age at hire. The length of employment for the matches was calculated as the lesser of the match's age at termination minus age at hire or the proband's (age at censor minus 20 years) minus the match's age at hire. When a lagged exposure variable was equal to zero, 0.1 was substituted, becoming −2.302 when the natural logarithm was calculated.

To examine the behavior of the study design when there were exposure intensity differences between probands and their matches, we repeated the analyses above with the probands either assigned a value of average exposure twice the value assigned to the matches (40 and 20 units respectively) or half that of the matches (20 and 40 units, respectively). The results of these analyses, lagged and unlagged, are in Table 2.

To examine interrelationships between variables, we calculated the correlation between age at censor and age at hire and also illustrated the relationship between age at hire and proband-match status as well as its correlation with average exposure, lagged and unlagged (Table 3).

Results

The probands are compared with the experimental population, and the matches with the probands in Table 1. As expected, the probands, selected at random from the experimental population, do not differ significantly from the study population. However, the matches differed significantly from the probands and the experimental population, having a greater age at censor, and age at hire and earlier date of birth, though a similar length of employment. Age at censor correlated with age at hire in the entire experimental population ($r = 0.695$, $P < 0.001$).

When probands and matches were assigned identical exposure values (20 units each), and exposure was not lagged, there were no proband-

TABLE 3
Experimental Population: Correlation Between Age at Censor and Age at Hire, Relationship Between Age at Hire and Proband-Match Status, and Correlation With Average Exposure, Lagged and Unlagged*

Variables	Lag in Yr	Probands	Matches
Age at censor, mean	—	62.6	72.7 ($P < 0.001$)
Age at hire, mean	—	32.3	37.0 ($P < 0.001$)
Correlation of age at hire with logarithm of average exposure	0	0	0
	20	−0.38 ($P < 0.001$)	−0.39 ($P < 0.001$)
% with duration of employment, cumulative exposure, and average exposure = zero	0	0	0
	20	20.1%	32.4% ($P < 0.001$)
Average exposure, geometric mean	0	20	20
	20	6.9	3.6 ($P < 0.001$)

*Figures are the geometric means of average exposure in probands and matches lagged 0 and 20 yr. Calculation of the geometric mean of average exposure lagged 20 yr uses 0.1 substituted for zero values.

match differences in length of employment, cumulative exposure, and average exposure (Table 2). However, when exposure was lagged 20 years, the probands had a significantly greater length of employment, cumulative exposure, and average exposure, despite being assigned the same average exposure as the matches.

When probands were assigned average exposure—either double or half that of the matches—and exposure was not lagged, the probands did not differ by length of employment but did differ in cumulative exposure and average exposure. The difference was, as expected, in the direction of the differentially assigned exposures. When exposure was lagged 20 years, however, the probands exceeded the matches on all three exposure variables, whether the value for average exposure assigned to the probands was double that assigned to the matches (40 vs 20 units) or was half that assigned to the matches (20 vs 40 units).

Significant correlations are apparent between age at hire and the logarithms of lagged average exposure in both the probands and matches (Table 3). This correlation does not exist with the unlagged average exposure. We observed that with 20-year lagging, a higher proportion of matches had zero values for exposures (32.4% for matches compared with 20.1% for probands). This indicates that the age at hire of the matches was more likely to be less than 20 years before the age at censor of their proband than was their proband's age at hire.

Discussion

The empirical evaluation with use of the experimental population gives insight into the bias inherent in the index study design. First, when probands are selected randomly from an industrial cohort population and are given the same value for average exposure as matches, they are found to have a significantly higher lagged exposure, whether measured by length of employment, cumulative exposure, or average exposure. The finding of higher lagged exposure in randomly chosen probands is not interpretable as a causal relationship between exposure and random selection. It also cannot be interpreted as a difference in exposure intensity because all subjects were assigned identical exposures of 20 units. Second, when probands are assigned greater (40 vs 20) or lesser (20 vs 40) average exposure and when exposure is lagged, probands have greater exposures in either case. Therefore, when exposure is lagged, there is no assurance that the study design can distinguish qualitatively whether cases and controls have the same or different exposures. As a result, when this study design is used, no conclusion may be drawn as to whether the case-control differences are an artifact of the study design or are due to true case-control differences.

The explanation for the behavior of this study design may be understood, starting with 5 clues. The first clue is that the study design selects as controls members of the population using the criterion that age at censor must be greater than or equal to that of the cases. The result of this selection is that controls (matches) have a greater age at censor than do cases (probands) (Table 1). Controls are selected with this criterion for comparison of cases with other members of the cohort who have survived at least to the same age (ie, whose age at censor is equal to or greater than that of the case). This is done in order to make the case-control analysis analogous to a proportional hazards cohort analysis.[10]

The second clue is the association in the experimental population between age at censor and age at hire. This association is generated in two ways:

A. Life expectancy increases with advancing age. Therefore, people who are older when they are hired have a better probability of living to a greater age. This is a modest effect and would lead to a weak association between age at hire and age at death.
B. People who are hired at an older age have the potential to reach a greater age within the cohort structure. The potential age a person can reach within a cohort

structure is bounded by the person's age at hire, date of hire, and date of mortality follow-up. To illustrate, if a date of hire is 1960, and the date of follow-up of the cohort is 1990, a person hired at age 20 may reach age 50 within the cohort structure, whereas a person hired at age 50 may reach age 80. This may have a strong effect on creating an association between age at hire and age at censor in an industrial cohort.

The astute reader will realize at this point that combining clue 1, that controls are selected to have greater age at censor, with clue 2, that age of censor is positively correlated with age at hire, will result in controls having a greater average age at hire. This fact comes into play with clue 3, which is harder to recognize.

The third clue is that when exposure is lagged, a greater proportion of controls (matches) acquire zero values for exposure (Table 3). The explanation for this lies in the detail of the exposure lagging procedure. When exposure is lagged for cases and controls, it is lagged from the age of censor of the case. Age at hire is a limiting value for lagged exposure because all exposure variables go to zero when the age at hire of either cases or controls is greater than the case age at censor minus the lag period. To illustrate, if a case dies at age 60, and the lag period is 20 years, exposure occurring after age 40 is not counted in both the cases and the controls. A control for this case hired at age 41 would be assigned a 20-year lagged exposure value of zero for all exposure measures: time of employment, cumulative exposure, and average exposure. Because controls on average have a greater age at hire than their cases, they are more likely to have their exposure "lag to zero." The result is that a larger proportion of controls will have zero values for lagged exposure measures.

The fourth clue to the study design behavior lies in the conversion of the exposure variables to natural logarithms. Because the logarithm of zero is undefined, the study design assigns "a small value," 0.1 as a substitute for zero. The natural logarithm of 0.1 is $-2.3\ldots$, and because after lagging there are proportionately more controls with zero values, there are also proportionately more controls with values of -2.3 when converted to logarithms. When these -2.3 values are averaged in calculating a geometric mean with the logarithms of the other exposure values, large and statistically significant differences are created between cases and controls.

The fifth clue is in the negative association of age at hire with the lagged exposure measures. Since owing to selection, age at hire is negatively correlated with case status, and because after lagging it is negatively associated with lagged exposure, age at hire is a confounder of the relationship between case-control status and lagged exposure. The confounding relationship of age at hire is created by virtue of its association with age at censor and the case-control difference in age at censor in the selection process, and its critical role in the calculation of lagged exposure values. The potential for confounding because of the relationship of demographic variables to disease and exposure in work cohort studies with use of the proportional hazards model is recognized.[11]

We believe that the difficulty in a priori recognition of the bias in the index study design is due to the extreme complexity of the study design. The bias occurs as the result of the combined effects of established analytic practices: nesting a case-control study within an industrial cohort, risk set sampling of controls on the case's age at censor, lagging exposure, and converting exposure variables to natural logarithms. However, once recognized, the mechanism can be, in part, understood as a variation on epidemiologic confounding. The effect is so strong as to overwhelm even substantial differences in assigned exposure. The study design may assign higher lagged exposure to cases whether actual exposure intensity is double or half that of controls.

We had a further concern about the index study design. When the index study design was adjusted to explore latency, consideration was given to time and whether exposure was relevant but not to time and whether persons were at risk. Therefore, persons were included in the analysis even when their time from hire was less than the assumed latency period. Under the latency assumption, these persons are not at risk of work-induced lung cancer during this period. The inclusion of these persons would seem to violate the proportional hazard assumption in the underlying analytic model. When a person not at risk of work-related lung cancer is a case, it is misleading to average in that person's zero exposure with the exposures of cases of people who meet the latency requirement for possibly work-related cancer. In a similar fashion, it is also misleading to average in the zero exposure values for control subjects who are not at risk. The resulting mean values have no intrinsic meaning. For example, the cases' geometric mean 20-year lagged average exposure of 10.2 is a function of the lagged average exposures in two unrelated groups, persons at risk of work-related lung cancer, and persons not at risk, whose latency-adjusted exposure is zero. The magnitude of the number is determined primarily by the proportion of persons not at risk under the latency assumption in each group and is not a level of exposure that can be related to actual risk.

This study had several limitations. Although the index study design is complex, it addresses an even more complicated problem. Neither the index study nor our empirical evaluation addressed all factors potentially important in understanding the relationship between work exposure and disease, such as birth cohort, and details of exposure timing, separate

analysis of shorter-tenure workers from longer-tenure workers, and how to disentangle length of employment from survival. By demonstrating limitations in the index study design, this study raised but did not resolve issues of how to analyze exposure-disease relationships taking latency into account.

Another limitation is that we have used only one industrial cohort data set in this evaluation. It is quite possible that the magnitude of the study design bias found here might be different with cohort databases with different demographic structures and patterns of mortality. For instance, this study design might not have an importantly biased result if applied to a school class cohort initiated at a point in time with little variation in age at matriculation and a common date of follow-up. Other industrial cohorts should be used to confirm our findings.

We also have concerns that empirical evaluation may be an overly precise tool in many instances. In industrial cohort studies, there are so many correlations between demographic factors and exposure and disease that it is possible that no study design will have an expected null hypothesis value of exactly zero exposure difference between cases and comparands. If this is so, then decisions will have to be made between accepting zero as the "close enough" expected value even though this is not precise, or estimating a non-zero difference as the number with which to compare the study result.

Our analyses are based on the results of 10 sets of computer simulations, each one consisting of 142 probands and 5 matches per proband, or 710 matches. In 10 simulations, this totaled 1420 probands with 7100 matches. To check the adequacy of the 10 simulations, we ran all our analyses after 1, 5, 6, 7, 8, 9, and 10 simulations and found that the results varied little after 5 simulations. For example, the geometric average exposure with a 20-year lag of matches was 3.8 based on one simulation, 3.7 based on five simulations, 3.6 based on nine simulations, and 3.6 based on ten simulations. The P values associated with all these simulations were essentially the same, $P < 0.001$. It was clear to us from this exercise that the results stabilized by five simulations and the additional five added little value to our understanding of the study design.

Finally, in several discussions leading up to the completion of this manuscript, we have had feedback that empirical evaluation of this index study design "is not relevant" to judging the behavior of the study design. We are puzzled by this. Because alternate empirical approaches have not been suggested, the implication is that experimentation cannot contradict "solid" theory. We invite further discussion on how to better formulate empirical evaluation of study designs with confidence that the result is relevant.

Conclusions

Empirical evaluation of the performance of a complex industrial cohort-nested case-control study design demonstrated bias when exposure was lagged. It produced the same qualitative result in a range of case-control exposure differences, making it unable to discriminate exposure-response relationships. This means that the study design may not distinguish between the study hypothesis that cases had higher lagged exposure than the controls, the null hypotheses that they had the same exposure, or even an alternate hypothesis that lung cancer cases had less lagged exposure.

The results of the empirical evaluation provided a rationale for expenditure of the effort required to further explicate the behavior of the study design. In doing so, it was deduced that the index study design produced results profoundly distorted by confounding by age at hire. This confounding was created by combined effects of the correlation in the cohort between age at censor and age at hire, the control selection criteria, the inclusion of persons not at risk under the latency assumption, and the conversion of exposure variables to logarithms with a substitution of 0.1 for zero.

It is very important to understand the specificity of the empirical evaluation approach, and that the results may not be extrapolated to even similar but subtly different designs. This specificity is illustrated by the finding with the index study design that though the bias is introduced by other factors, it is the log transformation step that greatly magnifies the bias to the point that it affects the conclusions of the study.

It is noteworthy that highly qualified persons who examine the index study design generally do not a priori deduce the outcome, which we have illustrated above. We believe that the empirical approach to evaluation of study designs offers insights that complement and may contradict those derivable from theory. We cannot answer the question of how many other study designs might benefit from empirical evaluation, but we suspect that other applications of the proportional hazard model might be profitably studied.

Variations on the empirical approaches used herein could be a standard process for evaluating complex study designs for unrecognized effects that invalidate assumptions about the expected result under the null and alternate hypotheses. Such experiments will provide impetus for advances in study design theory and improved guidelines for application.

Acknowledgments

Dr Deubner is an employee of Brush Wellman, Inc. Support for the work of the other two authors was provided by Brush Wellman, Inc, via contract with Roth Associates and also to RTI via contracts between Roth Associates and RTI. Brush Wellman, Inc, is a producer of beryllium materials.

References

1. Langholz B, Thomas D, Xiang A, Stram D. Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. *Am J Ind Med.* 1999;35:246–256.

2. Wild P, Leodolter K, Refregier M, Schmidt H, Zidet T, Haidinger G. A cohort mortality and nested case control study of French and Austrian talc workers. *Occ Environ Med.* 2002;59:98–105.
3. Sanderson WT, Ward EM, Steenland K, Peterson MR. Lung cancer case-control study of beryllium workers. *Am J Indust Med.* 2001;39:133–144.
4. Steenland K, Sanderson W. Lung cancer among industrial sand workers exposed to crystalline silica. *Am J Epidemiol.* 2001;153:695–703.
5. Ward E, Okun A, Ruder A, Fingerhut M, Steenland K. A mortality study of workers at 7 beryllium processing plants. *Am J Indust Med.* 1992;22:885–904.
6. Doll R. Occupational cancer: a hazard for epidemiologists. *Internat J Epidemiol.* 1985;14:22–31.
7. Levy PS, Roth HD, Deubner DC. Exposure to beryllium and occurrence of lung cancer: a re-examination of findings from a nested case-control study. *J Occup Environ Med.* 2007;49:96–101.
8. Beaumont J, Steenland K, Minton A, Meyers S. A computer program for incidence density sampling controls in case-control studies nested within occupational cohort studies. *Am J Epidemiol.* 1989;129:212–219.
9. Sanderson WT, Peterson M, Ward E. Estimating historical exposures of workers in a beryllium manufacturing plant. *Am J Indust Med.* 2001;39:145–157.
10. Lubin JH, Gail MH. Biased selection of controls for case-control analysis of cohort studies. *Biometrics.* 1984;40:63–75.
11. Checkoway H, Pearce N, Kreibel D. *Research Methods in Occupational Epidemiology*, 2nd ed. New York: Oxford University Press; 2004:276.