Soz Präventivmed 1993; *38*: 217–221

0303-8408/93/040217-05 $1.50 + 0.20/0

A multivariate example of case-cohort analysis: Neonatal mortality in Switzerland, 1979–81

David E. Matthews [1], Si-Chang Fan [2]

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada
[2] Department of Medical Statistics, Fourth Military Medical University, Xi'an, Shaanxi, China

Recent advances in statistical methods have opened up new possibilities with respect to the design and analysis of epidemiologic studies. In particular, the past decade has witnessed a virtual revolution in the methods which can now be used by epidemiologists and biostatisticians to wring interesting features, and conclusions, from well-defined and well-executed study designs. At the same time, advances in study design have progressed to such a degree in recent years that practitioners are hard pressed to keep abreast of recent developments. Matthews and Farewell [1] provide a qualitative summary of some of these possibilities. For a comprehensive survey reviewing developments in the 1980s, see Gail.[2]

In 1986, Prentice [3] described a major improvement in epidemiologic study methodology which he called the *Case-Cohort* design. The essence of this approach lies in the use of a random sample of cohort subjects to furnish vital covariate information which is used in the subsequent analysis and evaluation of relative risk differences between the population represented by the cases, and the corresponding cohort. The primary purpose of this paper is to describe the case-cohort design, and to discuss the merits of this relatively new method for analyzing cohort data. Kupper, McMichael and Spirtas [4] previously suggested an approach, called the hybrid retrospective design, which is similar to the case-cohort design. However, the estimation procedures which they describe do not apply to inference on population parameters, since the only variation which they considered is that attributable to sampling from the cohort. Likewise, the case-base design proposed by Miettinen [5] is not applicable because the focus of estimation is the population risk ratio, i.e., the ratio of disease probabilities at different values of some binary covariate z, say.

To illustrate the use of the case-cohort methodology, we describe a Swiss example of a case-cohort study, and discuss the findings of a multivariate case-cohort analysis of the study data. These data were derived from the 1979–81 birth cohort for Switzerland, which has been investigated from various perspectives by a number of authors.[6–9]

In paragraph 2 we describe the data set which was originally compiled, and indicate how it was subsequently organized to facilitate analysis according to the methodology for case-cohort designs [3]. The results or our analysis form the basis of paragraph 3. The paper concludes with a discussion of the study findings, and some summary remarks concerning the case-cohort methodology.

Methods

The data on which this study is based consist of a linked file of all Swiss births and neonatal deaths during the years 1979–81. During that time, a total of 220 540 births were recorded. From the linked file, a case-file consisting of all deaths and a simple random sample from the cohort of survivors was prepared, using a fixed sampling rate of three survivors to each neonatal death. One of the principal advantages of the case-cohort design [3] is the reduction in overall study cost which results because covariate information only needs to be obtained, entered and stored for the cases and for the random sample from the cohort. In the present instance, the cost savings are realized, primarily, in the ease with which the data can be manipulated, and analyzed. Use of the case-cohort design in this study resulted in at least a 50-fold reduction of the data. In other situations, significant savings could be achieved when use of the case-cohort design eliminates the necessity to obtain expensive laboratory test results or sample survey responses on all but the cases and a random sample of the noncases.

Minder et al. [9] indicates that a case-file was prepared based on the number of deaths (cases) observed, during the three-year study period, within four hours of birth. A total of 335 infants died during this time-frame, and with a sampling rate of 3:1 this collection of cases generated a case-file containing information on 1340 infants, the 335 deaths and a simple random sample of 1005 infants who survived for at least four hours following birth. In addition to this case-file, several others were also assembled, based on neonatal deaths occurring after the initial four hours had elapsed. The majority of deaths observed in neonatal infants during 1979–81 occurred between four and twenty-four hours following birth, and the corresponding case-file contains 5688 records, consisting of 1422

deaths and a random sample of 4266 infants from the cohort of twenty-four hour survivors. A third case-file corresponds to all neonatal deaths occurring within the first week of life among infants who survived for at least 24 hours and contains a total of 1580 records, consisting of 395 deaths and a random sample of 1185 one-week survivors. The final case-file is defined by the occurrence of death within the first month of life, among all infants who survived for at least one week. Since an analysis of the first case-file provides ample illustration of the various features of the case-cohort methodology, we have chosen not to report any findings from our analysis of the latter three sets of data.

The case-files which we used were those prepared by Minder et al.,[9] who provide a complete list of the explanatory variables which were available for analysis. We included in our multivariate analyses those variables on the original case-files which were individually associated with neonatal mortality during the time period in question or which, from the literature, have previously been shown to be associated with neonatal mortality. These include birthweight,[10,11] maternal age,[12] the sex of the child,[12,13] marital status of the mother,[12,14] body length,[15] birth order,[12] the urban or rural nature of the mother's residence[12] and the type of birth.[16] A list of the variables studied, and the corresponding coded values used in our analyses, may be found in Table 1.

In describing the case-cohort design, Prentice[3] discussed two separate cases. The first of these concerns a binary response variable, such as the occurrence of death within a fixed number of hours following birth. The second case involves time to response data, and is not applicable in this particular study of the cohort of all Swiss births in the years 1979–81. Prentice[3] demonstrated that in analysing a case-cohort study design involving all observed cases and a random sample of noncases from the cohort, i.e., neonatal survivors, asymptotic inference on the odds ratio can be carried out by fitting the binary logistic failure probability model directly to the subjects for whom covariate data has been assembled. Since the case-files previously described contain a binary response, namely neonatal death, and various explanatory variables (those listed in Table 1), the results of Prentice[3] ensure that an analysis of the relationship between neonatal death and the explanatory variables can be based on a binary logistic regression model such as the one that can be fitted using the software package GLIM.[17] For a description of the binary logistic probability model and the corresponding regression model, see chapter 11 of Matthews and Farewell.[1]

Because of missing values for various covariates in some of the records on the case-file, the actual sample sizes for the cases and noncases used in the binary logistic regression models that were fitted were less than the nominal sizes described above. Records with missing values were excluded from analysis only for models involving the explanatory variables whose values had not been recorded. In view of the known importance of birth weight as a predictor of neonatal mortality[10], all analyses described in paragraph 3 included weight at birth (in g.) as an explanatory variable. Thus, the estimated odds ratios, and corresponding 95% confidence intervals, are adjusted for the predictive effect of birth weight.

Tab. 1. Variable names and the corresponding coded values used in the case-cohort analysis of Swiss neonatal mortality, 1979–81.

Variable name	Coding
Weight at birth	In grams
Sex	0 = Female 1 = Male
Mother's marital status	0 = Married 1 = Otherwise
Birth rank	0 = Number of siblings ≤ 2 1 = Number of siblings > 2
Urban	0 = City 1 = Otherwise
Birth place	0 = Hospital 1 = Otherwise
Birth type	0 = Singleton 1 = Multiple
Body length	In cm
Maternal age	Years −28

Results

The results of our analysis of the four-hour neonatal mortality case-file described in paragraph 2 is presented in Table 2. Since birth weight was the single most important predictor of neonatal mortality in all of the time periods which we investigated,we first display a preliminary model involving only birth weight in order to provide a basis against which to compare the results of the final multivariate logistic regression model obtained. For each model fitted we show the variables included in the model, the sample sizes for cases (deaths) and noncases (survivors), the estimated regression coefficients and their standard errors, the estimated odds ratio of neonatal death and the corresponding 95% confidence interval. The approximate significance level of a test that the adjusted regression coefficient is zero, i.e., that the corresponding explanatory variable is not associated with neonatal mortality during the first four hours of life, is indicated in the final column of the table. Since birth weight, body length and the interaction of birth weight and birth type were modelled as continuous covariates, the corresponding estimated odds ratios and 95%

Tab. 2. Results of the case-cohort analysis of neonatal mortality during the first four hours of life.

Explanatory variable	Sample sizes (cases, noncases)	Estimated regression coefficient (standard error)	Estimated odds ratio (95% CI)	Significance level
Birth weight only	325,1003	−0.0027 (0.00015)	0.76 [a] (0.74, 0.79)	< 0.00001
Birth weight		−0.0022 (0.00032)	0.80 [a] (0.75, 0.86)	< 0.0001
Birth rank		1.522 (0.464)	4.58 (1.84, 11.4)	0.0001
Birth type	279,943	−3.70 (1.231)	0.02 (0.002, 0.28)	0.003
Birth type × Birth weight		0.0016 (0.0005)	1.17 [a] (1.05, 1.30)	0.004
Body length		−0.179 (0.053)	0.84 [b] (0.75, 0.93)	0.0007

[a] Change in the estimated odds ratio of neonatal death per 100 g. increase in birth weight.
[b] Reduction in the estimated odds ratio of neonatal death per 1 cm. increase in body length.

confidence intervals indicate the change in the odds ratio which results from a 100 g. increase in birth weight or a one cm. increase in an infant's length.

In our investigations of the data represented by these 1340 births, we considered all the variables specified in Table 1 and various interactions, as well as higher powers of birth weight, maternal age and body length. A quadratic function of the mother's age would be one way in which to model the unadjusted relationship between maternal age and neonatal mortality which has been reported elsewhere.[12] However, as Table 2 indicates, after adjusting for the predictive effect of weight at birth, we found that only the individual regression coefficients corresponding to birth rank, birth type, the interaction between birth weight and birth type and body length were significantly different from zero. After adjusting for the predictive effect of these variables, the regression coefficients corresponding to sex, birth place, urban or rural residence and maternal age were all not significantly different from zero.

Minder et al.[9] noted that information on the birth rank of a child was only available if the mother was married. Consequently, in the reduced case-file which resulted when records for which the birth rank information was missing were eliminated, the strength of the associations between neonatal mortality and each of the explanatory variables was reduced. This is not surprising since investigation of the 93 records for which the birth rank information was missing showed that more than one-third were cases (35) and the remaining 58 were noncases. Relative to the fixed sampling rate of three survivors for each case, this constitutes a disproportionate representation for the cases. However, after adjusting for the predictive effect of the variables specified in the model summarized in Table 2 (excluding birth rank), the regression coefficient corresponding to the marital status of the mother was not significantly different from zero.

Discussion

A careful scrutiny of the results described in the previous section, and summarized in Table 2, shows that the explanatory variable weight at birth is the single best predictor of neonatal mortality. Moreover, the inverse nature of this relationship is one which has been documented repeatedly in many different settings.[10,11,15] Based on the final model summarized in Table 2, the ratio of the odds of neonatal death per 100 g. of weight at birth is estimated to be 0.80 (95% confidence interval (0.75, 0.86)) during the first four hours of an infant's life. This estimated odds ratio indicates that for two infants resulting from singleton births which differ only with respect to their weight at birth, i.e., all other influential factors are identical, the ratio of the odds of neonatal death for the heavier infant relative to the lighter child is approximately $\exp(-0.0022\,\Delta\omega)$, where $\Delta\omega$ represents the difference in the weights of the two infants at birth. Thus, if the heavier child weighs 3500 g. and the lighter child weighs only 2500 g. at birth, the ratio of the odds of neonatal death for the heavier infant, relative to the lighter child, is approximately $\exp(-2.18) = 0.11$. Because of the significant interaction between birth weight and birth type identified in Table 2, a more extensive calculation, which is described below, is required to evaluate the effect of birth weight on the odds of neonatal mortality in the case of an infant resulting from a multiple pregnancy.

Although the regression coefficient associated with the marital status of the mother was significantly

different from zero in a univariate analysis of neonatal mortality, this effect did not persist in the multivariate model summarized in Table 2. Minder et al.[9] reported that marital status of the mother was associated with birth weight, but only weakly related to death during the first four hours of life. The disappearance, in the multivariate model, of the univariate association between neonatal mortality and marital status of the mother suggests that the effects related to being unmarried are mediated through a change in the distribution of birth weight. Similar findings have been reported in Canada.[12]

Table 2 indicates that during the first four hours of life, the risk of death may be similar for single and multiple births. To evaluate the ratio of the odds of neonatal death for a multiple pregnancy we need to multiply the estimated odds ratio for a multiple birth and the corresponding estimated odds ratio based on the birth weight of the child concerned (birth type × birth weight interaction; the estimated odds ratio of neonatal death for a multiple pregnancy is approximately 1.17 per 100 g. of weight at birth). If the birth weight is 2500 g., the estimated odds ratio for a multiple birth is only 1.19, and the corresponding 95% confidence interval, which is (0.04, 35.3), includes 1. In other studies reported in the literature, it has been reported that the excess mortality in twins is due almost entirely to a higher incidence of low birth weight in twin pregnancies.[16] The inclusion of the covariates birth type and the interaction between birth type and weight at birth suggests that a similar effect may be evident in the 1979–81 Swiss birth cohort as well.

The fitted logistic regression model also reveals that the odds of neonatal death are inversely related to the infant's body length; the estimated ratio is 0.84 per cm., with a corresponding 95% confidence interval of (0.75, 0.93). A similar linear dependence between the odds of neonatal mortality and body length, after adjustment for the predictive effect of weight at birth, was also reported by Herman and Hastie.[15]

A natural question that arises with respect to the case-cohort analysis discussed above concerns the relationship between case-cohort estimation of an odds ratio, and estimation of the same odds ratio using the full cohort information. In the case of a binary response, such as neonatal mortality, and a binary covariate such as sex, it can be shown that the estimated variance of an odds ratio estimate based on the full cohort, and on the case-cohort sample, should be virtually identical when the numbers of noncases, i.e., survivors, is large relative to the smaller number of cases in the two categories defined by the levels of the covariate. It follows that, in these circumstances, the case-cohort odds ratio estimator will have good efficiency properties with high probability[3], and confidence intervals for odds ratios based on a case-cohort analysis would be very similar to corresponding confidence intervals derived from the full cohort since the dominant terms in the estimated variance of the log odds ratio are the same in both analyses. In comparable circumstances, similar results would also be expected to hold in the case of a multivariate vector of covariates, e.g. sex, birth weight, maternal age, etc. However, detailed comparisons for the case of odds ratio estimation do not seem to have appeared in the statistical literature. Instead, attention has focussed on analysis of the efficiency and power of the case-cohort design for time to response data. In this situation, the comparisons are invariably complicated by the necessity to adopt various assumptions concerning the underlying failure time distribution, accrual rates, censoring mechanisms, etc. Although definitive conclusions with regard to statistical efficiency and power in the case-cohort design for time to response data have not yet been reached, the results published to date seem to indicate that the asymptotic relative efficiency improves as the number of noncases increases, and also as the relative risk increases.[18,19] Irrespective of the final outcome of statistical research concerning efficiency and power considerations, the case-cohort sampling design for time to response data facilitates estimation of (possibly stratified) cumulative baseline failure rates; such estimation appears to be unavailable in the alternative approach known as the synthetic case-control design.[20]

A particular feature of the regression-based approach which is worth noting is the fact that logistic regression models of the type used in the example discussed above are readily generalized to stratified versions. Thus, the epidemiologist or biostatistician is offered a choice for the control of confounding factors – stratification, regression modelling or both. An example of the use of stratification, in the context of a logistic regression model, to control for confounding is discussed in chapter 11 of Matthews and Farewell.[1]

Epidemiologic cohort studies frequently involve the periodic acquisition of raw data which are used to construct individual covariate histories. Such data may include the results of biochemical analysis of blood samples or cells, detailed occupational exposure records or dietary intake records to name only a few possibilities. In these circumstances, the cost of full cohort analysis can easily be dominated by the expense of acquiring covariate information, and may seriously jeopardize the initiation, or successful completion, of an otherwise well-designed study. The use of case-cohort sampling, which requires the assembling of covariate information only for cases and a random sample of noncases, may enable researchers to effectively bridge the gap between what is desirable and what is practical due to study resource limitations.

In conclusion, the case-cohort design for an epidemiologic study represents a notable advance in epidemiologic methodology. In the particular example which we have described, namely neonatal mortality during the first four hours of life in the 1979–81 birth cohort in Switzerland, we have shown that case-cohort methods can be used to estimate, efficiently, the association between various explanatory variables such as weight at birth, sex, body length, etc. and the odds of neonatal death. Use of this design offers researchers the potential for a significant reduction in study costs while preserving all the best features of more traditional epidemiologic methods, such as the cohort study.

Summary

A study of all births in Switzerland during the years 1979–81 is used to illustrate the advantages of the case-cohort design for this epidemiologic analysis of neonatal mortality. The example shows that familiar associations between infant mortality and explanatory variables such as sex and weight at birth can be precisely estimated using only a sample from the full cohort.

Résumé

Un exemple d'analyse multivariée cas-témoin: La mortalité néonatale en Suisse de 1979 à 1981

Une analyse épidémiologique de la mortalité néonatale utilise toutes les naissances en Suisse au cours des années 1979 à 1981 pour illustrer les avantages d'une analyse cas-témoin. L'exemple montre que les relations connues entre mortalité infantile et les variables explicatives telles le sexe et le poids de naissance peuvent être précisément estimées en utilisant seulement un échantillon de la cohorte complète.

Zusammenfassung

Ein mehrdimensionales Beispiel einer Fall-Kohorten-Analyse: Sterblichkeitsziffer Neugeborener in der Schweiz, 1979–1981

Eine Studie aller Geburten in der Schweiz in den Jahren 1979–1981 wird benutzt, um die Vorteile des Fall-Kohorten-Plans für diese epidemiologische Analyse der Sterblichkeitsziffer Neugeborener zu demonstrieren. Das Beispiel zeigt, daß eine Stichprobe aus der Kohorte ausreicht, um bekannte Zusammenhänge zwischen Kindersterblichkeit und abhängigen Kovariablen, wie z. B. Geschlecht und Geburtsgewicht, präzise zu schätzen.

References

1 *Matthews DE, Farewell VT.* Using and Understanding Medical Statistics. 2nd ed. Basel: S. Karger AG, 1988.

2 *Gail M.* A bibliography and comments on the use of statistical models in epidemiology in the 1980s. Stat med 1991; *10*: 1819–85.

3 *Prentice RL.* A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986; *73*: 1–11.

4 *Kupper LL, McMichael AJ, Spirtas R.* A hybrid epidemiologic study design useful in estimating relative risk. J Am Stat Assoc 1975; *70*: 524–8.

5 *Miettinen OS.* Design options in epidemiologic research: an update. Scand J Work Environ Health 1982; *8, Suppl 1*: 7–14.

6 *Bundesamt für Statistik.* Geburtsgewicht und Säuglingssterblichkeit in der Schweiz: 1979–81. Bern: Bundesamt für Statistik, 1985. (Beiträge zur schweizerischen Statistik, 126).

7 *Paccaud F, Martin-Béran B, Gutzwiller F.* Hour of birth as a prognostic factor for perinatal death. Lancet 1988; *i*: 340–3.

8 *Ackermann-Liebrich U, Romanens M, Paccaud F.* Epidemiologie der letalen Missbildungen in der Schweiz. Soz Präventivmed 1985; *30*: 9–12.

9 *Minder ChE, Ackermann-Liebrich U, Paccaud F.* Die Säuglingssterblichkeit in der Schweiz: multivariate Betrachtung. Soz Präventivmed 1985; *30*: 258–9.

10 *McCormick MC.* The contribution of low birth weight to infant mortality and childhood morbidity. N Engl J Med 1985; *312*: 82–90.

11 *Hogue CJ, Strauss LT, Buehler JW, Smith JC.* Overview of the National Infant Mortality Surveillance (NIMS) project. MMWR CDC Surveill Summ 1989; *38*: 1–46.

12 *Silins J, Semenciw RM, Morrison HJ. et al.* Risk factors for perinatal mortality in Canada. Can Med Assoc J 1985; *133*: 1214–9.

13 *Berman SM, Binkin NJ, Hogue CJR.* Assessing sex differences in neonatal survival: a study of discordant twins. Int J Epidemiol 1987; *16*: 436–440.

14 *Hein HA, Burmeister LF, Papke KR.* The relationship of unwed status to infant mortality. Obstet Gynecol 1990; *76*: 763–8.

15 *Herman AA, Hastie TJ.* An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression. J Clin Epidemiol 1990; *43*: 1179–90.

16 *Kiely JL.* The epidemiology of perinatal mortality in multiple births. Bull NY Acad Med 1990; *66*: 618–37.

17 *Baker RJ, Nelder JA.* The GLIM System. Release 3. Oxford: Numerical Algorithms Group, 1978.

18 *Self SG, Prentice RL.* Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Stat 1988; *16*: 64–81.

19 *Prentice RL, Self SG, Mason MW.* Design options for sampling within a cohort. In: Moolgavkar S. H., Prentice R. L. eds. Modern statistical methods in chronic disease epidemiology. New York: J Wiley and Sons, 1986: 50–62.

20 *Prentice RL, Farewell VT, Moolgavkar SH.* Biostatistical issues and concepts in epidemiologic research. J Chron Dis 1986; *39*: 1169–83.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and was carried out while S. C. Fan was a Visiting Associate Professor at the University of Waterloo. We are grateful to Dr. Ch. E. Minder for providing access to the data on which the analysis described in the paper is based, and to the editor and referees for suggestions and advice which led to significant improvements in a previous version of this paper.

Address for correspondence:
David E. Matthews
Statistics & Actuarial Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1