Epidemiologic Perspectives & Innovations

Methodology

Causal analysis of case-control data

Stephen C Newman*

Address: Department of Psychiatry, Mackenzie Health Sciences Centre, University of Alberta, Edmonton, Alberta, T6G 2B7, Canada

Email: Stephen C Newman* - stephen.newman@ualberta.ca

* Corresponding author

Published: 27 January 2006

Epidemiologic Perspectives & Innovations 2006, **3**:2 doi:10.1186/1742-5573-3-2

Received: 20 July 2005

Accepted: 27 January 2006

This article is available from: http://www.epi-perspectives.com/content/3/1/2

Abstract

In a series of papers, Robins and colleagues describe inverse probability of treatment weighted (IPTW) estimation in marginal structural models (MSMs), a method of causal analysis of longitudinal data based on counterfactual principles. This family of statistical techniques is similar in concept to weighting of survey data, except that the weights are estimated using study data rather than defined so as to reflect sampling design and post-stratification to an external population. Several decades ago Miettinen described an elementary method of causal analysis of case-control data based on indirect standardization. In this paper we extend the Miettinen approach using ideas closely related to IPTW estimation in MSMs. The technique is illustrated using data from a case-control study of oral contraceptives and myocardial infarction.

Introduction

In a series of papers, Robins and colleagues describe inverse probability of treatment weighted (IPTW) estimation in marginal structural models (MSMs) [1-7], a method of causal analysis of longitudinal data based on counterfactual principles. This family of statistical techniques is similar in concept to weighting of survey data, except that weights are estimated using study data rather than defined so as to reflect sampling design and post-stratification to an external population. Several decades ago Miettinen [8] described an elementary method of causal analysis of case-control data based on indirect standardization. In this paper we extend the Miettinen approach using ideas closely related to IPTW estimation in MSMs. For simplicity we ignore random error until the illustrative example.

Population-based incidence case-control study

Consider a population-based case-control study having an incidence design, that is, one in which only incident cases are eligible for recruitment. Let E be a dichotomous variable (0: absent, 1: present) representing the exposure of interest, and let F be a polychotomous variable ($i = 0,1, \ldots, I$), which we later treat as a confounder. At any time point we may think of the population as being comprised of exposed and unexposed (sub)populations. Suppose that recruitment of cases and controls takes place over a period of T years. We assume that during the period of recruitment the exposed and unexposed populations are stationary (i.e., independent of time) with respect to population size and incidence rate (of disease) in each of the strata of F [9]. Provided that T is not too large, say no more than two or three years, this assumption is likely to be approximately satisfied in practice.

Let N_{1i} be the number of people in the ith stratum of the exposed population who are free of disease (at any time during the period of recruitment), and let N_{0i} be the corresponding number in the ith stratum of the unexposed population. Let $N_1 = \sum_i N_{1i}$ and $N_0 = \sum_i N_{0i}$. There-

Table 1: Number of cases and controls in *i*th stratum of *F* under simple random sampling

E	Case	Control
1	$a_{1i} = \gamma R_{1i} N_{1i} T$	$b_{1i} = \lambda N_{1i}$
0	$a_{0i} = \gamma R_{0i} N_{0i} T$	$b_{0i} = \lambda N_{0i}$

fore at any time during the period of recruitment, there are N_1 exposed and N_0 unexposed people in the population "at risk" of disease, hence eligible to be controls. Since the population is stationary, we may assume that controls are selected at the end of the period of recruitment. This avoids the inconvenience of having a control selected early in the study become a case later on. In practice, controls are usually sampled throughout the period of recruitment, with one or more controls enrolled as each case enters the study. The case triggering this activity and the associated controls can be thought of as a matched set, where the matching variable is "time." This method of subject recruitment is a type of risk set sampling and, in theory, should be followed by a conditional statistical analysis [10]. Generally, matching on time is ignored in the analysis of case-control data, which in practical terms is not that different from making the stationary population assumption.

Let R_{1i} and R_{0i} be the incidence rates (of disease) in the *i*th stratum of the exposed and unexposed populations, respectively. The crude incidence rates are

$$R_1 = \frac{\sum_i R_{1i} N_{1i}}{N_1} \qquad (1)$$

and

$$R_0 = \frac{\sum_i R_{0i} N_{0i}}{N_0}.$$

The impact of exposure can be measured using the standardized morbidity ratio, which has different forms depending on the choice of standard population [11]. Taking the standard population to be, in turn, the exposed, unexposed, and total (exposed plus unexposed) populations, the corresponding standardized morbidity ratios are

$$SMR_{\mathrm{E}} = \frac{\sum_i R_{1i} N_{1i}}{\sum_i R_{0i} N_{1i}} \qquad (2)$$

$$SMR_{\mathrm{U}} = \frac{\sum_i R_{1i} N_{0i}}{\sum_i R_{0i} N_{0i}}$$

and

$$SMR_{\mathrm{T}} = \frac{\sum_i R_{1i} \left(N_{1i} + N_{0i} \right)}{\sum_i R_{0i} \left(N_{1i} + N_{0i} \right)}.$$

We now view the population as an open (dynamic) cohort that is followed over the period of recruitment, with onset of disease as the endpoint of interest [12]. Entry into the cohort occurs, for example, as a result of birth and in-migration, and censoring takes place when, for instance, there is out-migration and death from a cause other than the disease of interest.

Simple random sampling

Assume that cases and controls are sampled using simple random sampling. Let γ and λ be the sampling probabilities for cases and controls, respectively; that is, γ is the proportion of eligible cases enrolled in the study during the period of recruitment, and λ is the corresponding proportion of controls. We assume that these are also the sampling probabilities within each of the strata of $E \times F$, the cross-classification of E and F. It follows from the stationary population assumption that over the period of recruitment the number of person-years experienced by individuals in the *i*th stratum who are exposed and at risk of disease is $N_{1i}T$. The corresponding number of (incident) cases is $R_{1i}N_{1i}T$, with $a_{1i} = \gamma R_{1i}N_{1i}T$ of them recruited into the study. Likewise, the number of cases recruited into the study among individuals in the *i*th stratum who are unexposed and at risk of disease is $a_{0i} = \gamma R_{0i}N_{0i}T$. In view of remarks made above, $b_{1i} = \lambda N_{1i}$ exposed and $b_{0i} = \lambda N_{0i}$ unexposed controls will be recruited into the study from the *ith* stratum. Table 1 summarizes these observations.

It follows from Table 1 that

$$SMR_{\mathrm{E}} = \frac{\sum_i a_{1i}}{\sum_i \left(\frac{b_{1i}}{b_{0i}} \right) a_{0i}} \qquad (3)$$

$$SMR_{\mathrm{U}} = \frac{\sum_i \left(\frac{b_{0i}}{b_{1i}} \right) a_{1i}}{\sum_i a_{0i}}$$

and

$$SMR_{\mathrm{T}} = \frac{\sum_i \left(1 + \frac{b_{0i}}{b_{1i}} \right) a_{1i}}{\sum_i \left(1 + \frac{b_{1i}}{b_{0i}} \right) a_{0i}}$$

Table 2: Weighted number of cases and controls under simple random sampling

E	Case	Control
I	$\sum_i a_{1i} = \gamma \sum_i R_{1i} N_{1i} T$	$\sum_i b_{1i} = \lambda N_1$
0	$\sum_i \left(\frac{b_{1i}}{b_{0i}} \right) a_{0i} = \gamma \sum_i R_{0i} N_{1i} T$	$\sum_i \left(\frac{b_{1i}}{b_{0i}} \right) b_{0i} = \lambda N_1$

which shows that SMR_E, SMR_U and SMR_T can be estimated from incidence case-control data [13-15]. Note that nowhere have we made the rare disease assumption.

We are interested in measuring the causal effect of exposure on the exposed cohort using counterfactual methods [16-21]. To accomplish this we imagine the group of individuals in the exposed cohort *prior to exposure* and consider two scenarios: in the first, exposure subsequently occurs (as it does in reality); in the second, exposure does not occur. The second scenario is counterfactual because it rests on the hypothetical condition that exposure does not take place, when in fact it does. By contrasting outcomes arising out of the two scenarios we are able to define parameters having a causal interpretation. This is because we are (in theory) comparing two groups of individuals that are identical except for exposure status. The crude incidence rate corresponding to the first scenario is R_1. Denote the crude incidence rate for the second scenario by R_1^*. Even though the second scenario is counterfactual, it is possible, provided certain assumptions are satisfied, to estimate R_1^*, as discussed below.

In practice, the unexposed cohort, not the exposed cohort under the counterfactual condition, is used for comparative purposes. To the extent that the two associated incidence rates, R_0 and R_1^*, differ, we say that there is confounding. More precisely, the counterfactual definition of confounding states that confounding is present if and only if $R_0 \neq R_1^*$ [16-21].

We now make two fundamental assumptions: (1) E does not "affect" F (in particular, F is not on a causal pathway between E and the disease), and (2) there is no confounding (according to the counterfactual definition) in the strata of F. Using arguments analogous to those in [21] and [22], we have

$$R_1^* = \frac{\sum_i R_{0i} N_{1i}}{N_1}. \qquad (4)$$

Since there is no confounding in the strata of F, when *confounding* is present, that is, $R_0 \neq R_1^*$, we attribute it to F and say that F is a *confounder*. It follows from (1), (2) and (4) that

$$SMR_E = \frac{R_1}{R_1^*} \qquad (5)$$

which shows that under the above two assumptions, SMR_E has a causal interpretation.

Following the approach of Sato and Matsuyama [11], we assign each exposed subject in the ith stratum the weight 1, and each unexposed subject the weight b_{1i}/b_{0i}. We refer to these weights as the empirical weights. Note that b_{1i}/b_{0i} is the odds that a control in *ith* stratum is exposed. From Table 2, which gives case-control counts after applying these weights, we see that SMR_E can be interpreted as a weighted odds ratio. Accordingly, in the case-control setting we denote SMR_E by sOR and refer to it as the standardized odds ratio.

Let

$$OR_i = \frac{a_{1i} b_{0i}}{a_{0i} b_{1i}}$$

and $n_i = a_{1i} + a_{0i} + b_{1i} + b_{0i}$. It is readily demonstrated that sOR as given by (3) and the Mantel-Haenszel odds ratio estimate OR_{MH} [23] can be expressed as weighted sums of the OR_i:

$$\text{s}OR = \frac{\sum_i \left(\frac{a_{0i} b_{1i}}{b_{0i}} \right) OR_i}{\sum_i \frac{a_{0i} b_{1i}}{b_{0i}}}$$

$$OR_{MH} = \frac{\sum_i \left(\frac{a_{0i} b_{1i}}{n_i} \right) OR_i}{\sum_i \frac{a_{0i} b_{1i}}{n_i}}. \qquad (6)$$

These expressions differ only to the extent that the relative magnitudes of the b_{0i} and n_i vary across strata. For case-control studies in which unexposed controls constitute the majority of subjects, sOR and OR_{MH} will be close in value.

It was pointed out by Greenland [15] that OR_{MH} does not have an epidemiologic interpretation when there is effect modification. This is because the stratum-specific weights in (6) do not reflect a recognizable target population. With sOR the target population is clearly specified (namely, the exposed population), and so sOR has a causal interpretation even in the presence of effect modi-

Table 3: Number of cases and controls in *ij*th stratum of *F* × *G* under stratified random sampling

E	Case	Control
1	$a_{1ij} = \gamma_j R_{1ij} N_{1ij} T$	$b_{1ij} = \lambda_j N_{1ij}$
0	$a_{0ij} = \gamma_j R_{0ij} N_{0ij} T$	$b_{0ij} = \lambda_j N_{0ij}$

fication. This is advantageous in a number of settings. Consider the familiar situation in which, after stratification by one or more confounders, the stratum-specific odds ratio estimates do not exhibit a meaningful pattern, or the differences in these estimates can be distinguished on statistical grounds but are of no practical importance. When this occurs it is desirable to have recourse to a summary odds ratio estimate, even though effect modification may be present.

Stratified random sampling

Let G be a polychotomous variable ($j = 0, 1, ..., J$) and suppose that cases and controls are sampled using stratified random sampling based on the strata of G. Let γ_j and λ_j be the sampling probabilities for cases and controls in the jth stratum, respectively. We assume that these are also the sampling probabilities for the exposed and unexposed populations in the jth stratum. Corresponding to Tables 1 and 2 we have Tables 3 and 4, from which it follows that

$$SMR_{\mathrm{E}} = \frac{\sum_{ij}\left(\frac{1}{\gamma_j}\right)a_{1ij}}{\sum_{ij}\left(\frac{b_{1ij}}{\gamma_j b_{0ij}}\right)a_{0ij}}.$$

Under stratified random sampling, we assign each exposed subject in the ijth stratum the (empirical) weight $1/\gamma_j$, and each unexposed subject the weight $b_{1ij}/\gamma_j b_{0ij}$. As before, in the case-control context we denote SMR_{E} by sOR.

MSM-IPTW approach

When there are multiple confounders, the data can be stratified according to their cross-classification and the above method used. However, this may lead to cells with small or zero entries, resulting in instability of estimates. A statistically more efficient alternative is to adopt the MSM-IPTW approach and obtain the weights (for controls) from a logistic regression analysis of control data, where E is the dependent variable and the confounders (of the E-disease association) are the independent variables. We refer to these weights as regression weights.

Under simple random sampling, the weight for each exposed subject is set equal to 1, and the weight for each unexposed subject is taken to be the fitted odds for that individual. For stratified random sampling, the logistic regression analysis of control data must include the stratifying variable. In the jth stratum, the weight for each exposed subject is set equal to the reciprocal of the sampling probability, and the weight for each unexposed subject is taken to be the fitted odds for that individual multiplied by the reciprocal of the sampling probability.

Once the regression weights have been calculated, the odds ratio for the exposure-disease association is estimated from a weighted logistic regression analysis using generalized estimating equations (GEE) [24], where E is the sole independent variable. As remarked by Hernán et al. [6], it has been shown by Robins [1,2] that for longitudinal data where there are no unmeasured confounders and where a certain positivity assumption is met, the weighted GEE approach produces an asymptotically unbiased estimate of the causal parameter. Depending on the software used for the GEE analysis, it may be necessary to scale the weights such that their sum across all cases equals the actual number of cases, and likewise for controls.

Example

Table 5 presents data from an incidence case-control study of oral contraceptives (OC) and myocardial infarction (MI) [25]. We are interested in measuring the causal effect of oral contraceptive use on myocardial infarction in women taking this medication; that is, the target population is women taking oral contraceptives. For the purposes of illustration, we assume that age (AGE) and

Table 4: Weighted number of cases and controls under stratified random sampling

E	Case	Control
1	$\sum_{ij}\left(\frac{1}{\gamma_j}\right)a_{1ij} = \sum_{ij} R_{1ij} N_{1ij} T$	$\sum_{ij}\left(\frac{1}{\lambda_j}\right)b_{1ij} = N_1$
0	$\sum_{ij}\left(\frac{b_{1ij}}{\gamma_j b_{0ij}}\right)a_{0ij} = \sum_{ij} R_{0ij} N_{1ij} T$	$\sum_{ij}\left(\frac{b_{1ij}}{\lambda_j b_{0ij}}\right)b_{0ij} = N_1$

Table 5: Case-control study of oral contraceptives and myocardial infarction [25]

CIG		AGE						Total	
		25–34		35–44		45+			
	OC	Case	Control	Case	Control	Case	Control	Case	Control
none	1	0	38	1	12	3	2	4	52
	0	1	281	13	318	20	155	34	754
		$\widehat{OR}$ = 2.44		$\widehat{OR}$ = 2.03		$\widehat{OR}$ = 11.63		$\widehat{OR}$ = 1.71	
1–24	1	2	35	1	15	0	1	3	51
	0	5	221	32	249	42	96	79	566
		$\widehat{OR}$ = 2.53		$\widehat{OR}$ = 0.52		$\widehat{OR}$ = 0.76		$\widehat{OR}$ = 0.42	
25+	1	11	22	8	8	3	2	22	32
	0	8	112	53	125	31	50	92	287
		$\widehat{OR}$ = 7.00		$\widehat{OR}$ = 2.36		$\widehat{OR}$ = 2.42		$\widehat{OR}$ = 2.14	
Total	1	13	95	10	35	6	5	29	135
	0	14	614	98	692	93	301	205	1607
		$\widehat{OR}$ = 6.00		$\widehat{OR}$ = 2.02		$\widehat{OR}$ = 3.88		$\widehat{OR}$ = 1.68	

OC: oral contraceptives
CIG: cigarettes
AGE: age

cigarettes (CIG) are sufficient to control confounding and that there is no misclassification or other source of bias.

We first performed a standard logistic regression analysis, with MI as the dependent variable and OC, AGE and CIG as the independent variables. As pointed out by Greenland and Maldonado [26], there are problems identifying the target population when using standard logistic regression analysis. Models were fit using EGRET [27]: statistical significance of individual terms was determined using the likelihood ratio test, and the goodness-of-fit statistic G^2 was based on the deviance. On purely statistical grounds the best-fitting model had main effects for OC, AGE and CIG, along with the interaction term AGE × CIG (G^2 = 12.0, df = 8, p = .15). The odds ratio estimate for the OC-MI association was 2.82 (95% confidence interval [CI]: 1.70,4.68). Of note, the Mantel-Haenszel odds ratio estimate, OR_{MH}= 2.82 (95% CI: 1.70,4.69), was virtually identical to the logistic regression estimate. The OR_{MH} confidence interval was based on the variance estimate described by Robins, Breslow and Greenland [28,29]. The model with main effects for OC, AGE and CIG, along with the interaction term OC × CIG also fit the data quite well (G^2 = 17.4, df = 10, p = .068). Given that oral contraceptive use is the exposure of interest, it is reasonable – on substantive grounds – to consider this as the "final" model. If so, because of the OC × CIG interaction, the model no longer provides a summary estimate of the odds ratio for the OC-MI association.

Next, we conducted an analysis using the MSM-IPTW approach. To obtain regression weights, a standard logistic regression analysis of control data was performed, with OC as the dependent variable, and with AGE and CIG as the independent variables. The best-fitting model had only a main effect for AGE (G^2= 5.06, df = 6, p = .54). We then conducted a weighted logistic regression analysis using generalized estimating equations, with MI as the dependent variable and OC as the sole independent variable. Following Hernán et al. [4] and Sato and Matsuyama [11], calculations were performed using the SAS procedure PROC GENMOD [30]. The odds ratio estimate for the OC-MI association was 3.34 (95% CI: 2.15, 5.21). Interestingly, when empirical weights were used instead of regression weights, the odds ratio estimate (which equals s*OR*) was 2.83 (95% CI: 1.82,4.41). This is very close to the odds ratio and confidence interval estimates based on the standard logistic regression and Mantel-Haenszel analyses.

Discussion

The counterfactual definition of confounding represents an important conceptual advance over earlier formulations of confounding. Working within the counterfactual

framework, Robins and colleagues developed inverse probability of treatment weighted estimation in marginal structural models for the analysis of longitudinal data [1-7]. Although primarily aimed at the problem of time-dependent confounding, this method is valid when confounders are independent of time.

Extending the work of Miettinen [8], in this paper we present a method of causal analysis of case-control data that is closely related to IPTW estimation in MSMs. We consider only case-control studies conducted in a stationary population. Provided the time period during which the study is conducted is not too long, it may be reasonable to regard the population as at least approximately stationary. Whether strictly valid or not, the stationary population assumption appears to be made routinely – usually implicitly – when case-control studies are conducted. An alternative is to match controls to cases on time of recruitment using risk set sampling [10] and perform a conditional data analysis. Under the rare disease assumption, approximate parameter estimates can then be obtained using the MSM-IPTW approach [7].

Declaration of competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

The author thanks Dr. James Robins for helpful discussions.

References

1. Robins JM: **Marginal structural models.** In *1997 Proceedings of the Section on Bayesian Statistical Science* Alexandria, VA, American Statistical Association; 1998:1-10.
2. Robins JM: **Marginal structural models versus structural nested models as tools for causal inference.** In *Statistical Models in Epidemiology: the Environment and Clinical Trials* Edited by: Halloran ME, Berry D. New York, Springer-Verlag; 1999:95-134.
3. Robins JM: **Association, causation, and marginal structural models.** *Synthese* 1999, **121:**151-179.
4. Hernán MA, Brumback B, Robins JM: **Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men.** *Epidemiology* 2000, **11:**561-570.
5. Robins JM, Hernán MA, Brumback B: **Marginal structural models and causal inference in epidemiology.** *Epidemiology* 2000, **11:**550-560.
6. Hernán MA, Brumback B, Robins JM: **Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures.** *Statistics in Medicine* 2002, **21:**1689-1709.
7. Robins JM: **Comment on " Covariance adjustment in randomized experiments and observational studies" by Rosenbaum PR.** *Statistical Science* 2002, **17:**286-327.
8. Miettinen OS, Cook EF: **Confounding: essence and detection.** *American Journal of Epidemiology* 1981, **114:**593-603.
9. Keyfitz N: **Introduction to mathematical demography. With revisions.** Reading, MA, Addison-Wesley; 1977.
10. Langholz B, Goldstein L: **Risk set sampling in epidemiologic cohort studies.** *Statistical Science* 1996, **11:**35-53.
11. Sato T, Matsuyama Y: **Marginal structural models as a tool for standardization.** *Epidemiology* 2003, **14:**680-686.
12. Rothman KJ, Greenland S: **Modern epidemiology.** Second edition. Philadelphia, Lippincott-Raven; 1998:94.
13. Miettinen OS: **Components of the crude risk ratio.** *American Journal of Epidemiology* 1972, **96:**168-172.
14. Miettinen OS: **Estimability and estimation in case-referent studies.** *American Journal of Epidemiology* 1976, **103:**226-235.
15. Greenland S: **Interpretation and estimation of summary ratios under heterogeneity.** *Statistics in Medicine* 1982, **1:**217-227.
16. Greenland S, Robins JM: **Identifiability, exchangeability, and epidemiologic confounding.** *International Journal of Epidemiology* 1986, **15:**412-418.
17. Robins JM, Morgenstern H: **The foundations of confounding in epidemiology.** *Computers and Mathematics with Applications* 1987, **14:**869-916.
18. Greenland S, Robins JM, Pearl J: **Confounding and collapsibility in causal inference.** *Statistical Science* 1999, **14:**29-46.
19. Greenland S, Morgenstern H: **Confounding in health research.** *Annual Revue of Public Health* 2001, **22:**189-212.
20. Maldonado G, Greenland S: **Estimating causal effects (with commentary).** *International Journal of Epidemiology* 2002, **31:**422-438.
21. Newman S: **Commonalities in the classical, collapsibility and counterfactual concepts of confounding.** *Journal of Clinical Epidemiology* 2004, **57:**325-329.
22. Wickramaratne P, Holford TR: **Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding.** *Biometrics* 1987, **43:**751-765.
23. Mantel N, Haenszel W: **Statistical aspects of the analysis of data from retrospective studies of disease.** *Journal of the National Cancer Institute* 1959, **22:**719-748.
24. Hanley JA, Negassa A, Edwardes MDdeB, Forrester JE: **Statistical analysis of correlated data using generalized estimating equations: an orientation.** *American Journal of Epidemiology* 2003, **157:**364-375.
25. Shapiro S, Slone D, Rosenberg L, Kaufman DW, Stolley PD, Miettinen OS: **Oral- contraceptive use in relation to myocardial infarction.** *Lancet* 1979, **7:**743-746.
26. Greenland S, Maldonado G: **The interpretation of multiplicative-model parameters as standardized parameters.** *Statistics in Medicine* 1994, **13:**989-999.
27. Cytel Software Corporation: *EGRET® for Windows. User Manual* Cambridge, MA: Cytel Software Corporation; 1999.
28. Robins J, Breslow N, Greenland S: **Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models.** *Biometrics* 1986, **42:**311-323.
29. Silcocks P: **An easy approach to the Robins-Breslow-Greenland variance estimator.** *Epidemiologic Perspectives & Innovations* 2005, **2:**9.
30. SAS Institute Inc: *SAS/STAT ® User's Guide: Version 8 Volume 2*. Carey, NC: SAS Institute Inc; 1999.