Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 28 (2007) 242–248
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

The effect of omitted covariates on confidence interval and study power in binary outcome analysis: A simulation study

Abdissa Negassa [a,*], James A. Hanley [b]

[a] *Department of Epidemiology and Population Health, Division of Biostatistics, Albert Einstein Collage of Medicine, Of Yeshiva University, Bronx, NY, USA*

[b] *Department of Epidemiology, Biostatistics and Occupational Health, Faculty of Medicine, McGill University, Montreal, Quebec, Canada*

Received 4 March 2006; accepted 17 August 2006

Abstract

Background/objectives: The consequence of omitted but balanced covariates on odds ratio point estimation is well-known in the literature. When exposure or intervention has a non-null effect on disease outcome, omitted covariates lead to underestimation of the effect of exposure or intervention. However, the effect of omitted covariates on confidence interval and study power is unknown.

Study design and setting: A simulation study is carried out to assess the effect of omitted covariates on confidence interval and study power for a plausible range of scenarios. Coverage probability and study power are assessed systematically over a range of study size, type of omitted covariate and magnitude of effect. A real-life example using a randomised experiment on flies' sexuality is provided.

Results: When a balanced covariate is omitted, coverage probability was lowered by 2.9–80%. Likewise study power was reduced by as much as 58%. The impact becomes substantial when the covariate is continuous, has large variability and has a larger effect than the effect of exposure or intervention. The result from a real-life example concurs with the simulation finding.

Conclusion: Omitting an important balanced covariate lowers both coverage probability and study power. This implies the need for thoughtful consideration of important covariates at the design as well as the analysis stages of a study.

Keywords: Omitted covariate; Logistic regression; Bias; Confidence interval; Study power

1. Introduction

The consequence of omitting balanced covariates under various non-linear models was first demonstrated by Gail et al. [1]. By balanced covariate we refer to the distribution of the covariate being comparable between the exposure/intervention groups. Asymptotically, i.e., as sample size increases, randomisation ensures well-balanced/comparable intervention groups and hence minimizing the potential for confounding. However, for small-to-moderate studies, chance disparity/imbalance in the distribution of important covariates is still a possibility after randomisation unless stratified block

* Corresponding author. Tel.: +1 718 430 3575; fax: +1 718 430 8780.
E-mail address: anegassa@aecom.yu.edu (A. Negassa).

1551-7144/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2006.08.007

randomisation is employed. The main finding of Gail et al. [1] was a downward bias or underestimation of the effect of the exposure or intervention of interest when important covariates are omitted. Hauck et al. [2] have also discussed the same issue in terms of the different definitions of confounding and illustrated how these definitions at times disagree. In addition, these authors pointed out an important distinction in that the issue of omitted covariates is different than that of classical confounding [3]. The direction of the bias involved in the case of omitted covariates is predictable, i.e., always towards the null, as opposed to classical confounding which could be either way. Particularly, if the variable of main interest such as exposure or intervention does not have effect on outcome, then omitted covariates do not introduce bias while confounders do. Chao et al. [4] extended the general attenuation effect result to the case of correlated binary outcomes.

What is not known in the literature is the impact of omitted covariates on confidence interval and study power. Hauck et al. [2] argue, indirectly, that omitted covariates lead to loss of efficiency since omitting covariates is some form of model misspecification [5]. The goal of this paper is to investigate using a simulation study the effect of omitted but balanced covariates on confidence interval estimation and study power in an uncorrelated binary outcome setting.

2. Example of what is already known

Table 1 illustrates the effect of an omitted covariate on point estimation of odds ratio using a hypothetical study. $P(D|\bar{E})$ is the proportion with the outcome/disease among subjects without exposure/intervention and $P(D|E)$ is the proportion with the outcome/disease among subjects with exposure/intervention. In addition, it is assumed that the probability of assignment to each stratum of the covariate is 0.5 and exposure within each stratum is balanced. Under this configuration, the stratum specific odds ratio associated with exposure/intervention is 9.0 (first two rows of Table 1). However, when we collapse the data over strata, i.e., use a crude analysis, the odds ratio reduces to 5.4 (last row of Table 1). Although this is an extreme example, it is likely to happen in real life when a very strong predictor of outcome is omitted from the analysis; we will provide such an example. As previously pointed out by others [2], here the covariate defining the strata does not satisfy the classical definition of confounding, however, it satisfies the operational definition of confounding, "change-in-estimate" [7]. The effect of omitting such type of covariates is always underestimation of the odds ratio, provided that there is an exposure/intervention effect in the source population. In the subsequent sections, we will refer to exposure/intervention as "exposure" for simplicity of presentation.

3. Simulation study

We simulated data consisting of disease status (D), a binary exposure (E) and a covariate (C), satisfying independence between exposure and covariate. The exposure and the covariate are associated with disease status through the following model:

$$\text{logit}P(D|E,C) = \beta_0 + \beta_1 E + \beta_2 C \tag{1}$$

The coefficients of E and C in the above model were chosen so as to provide a wide range of combinations between exposure and omitted covariate effects. For exposure, an odds ratio of 2.0 was considered throughout while considering an odds ratio ranging between 2.0 and 10.0 for the omitted covariate, the latter specification corresponding to a covariate that is moderate to very strong predictor of disease outcome. Both binary and continuous omitted covariate scenarios were investigated. In the case of a binary omitted covariate, we considered both levels of the covariate to be equally likely, providing the maximum variation. In the case of a continuous omitted covariate, we considered two scenarios in which we kept the mean to be the same but the variance was increased in the second set by 5-fold [i.e., $C \sim N(0,1)$ and $C \sim N(0,5)$, respectively]. In addition, we kept the probability of disease among the non-exposed at the

Table 1
Comparison of crude and adjusted analyses[a]

Levels of omitted covariate	$P(D\|\bar{E})$	$P(D\|E)$	Odds ratio
Stratum 1	10	50	9.0
Stratum 2	50	90	9.0
Crude	30	70	5.4

[a] Adapted from Hanley et al. [6].

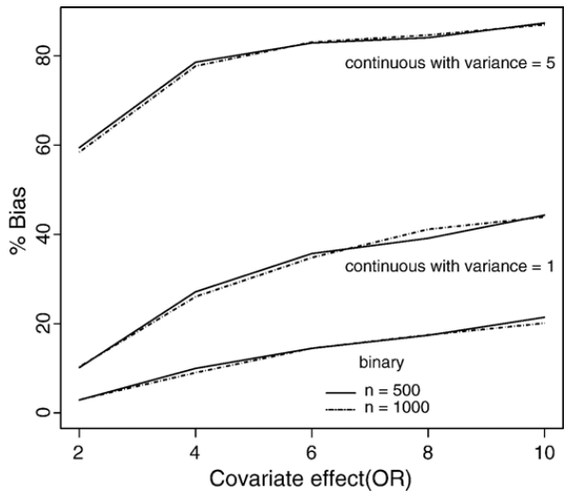

Fig. 1. Percent bias: as a function of the nature (binary vs. continuous), strength and spread of omitted covariates. The exposure effect was fixed at OR=2 (see text for details).

lower level of the binary covariate (or at the mean value of a continuous covariate) at 0.4. The sample sizes were fixed at 500 and 1000. We considered 10,000 simulations per scenario.

We compared the model omitting the covariate:

$$\text{logit}P(D|E) = \beta_0^* + \beta_1^* E \tag{2}$$

with the model including the covariate, i.e., the underlying model generating the data as given by (1) above.

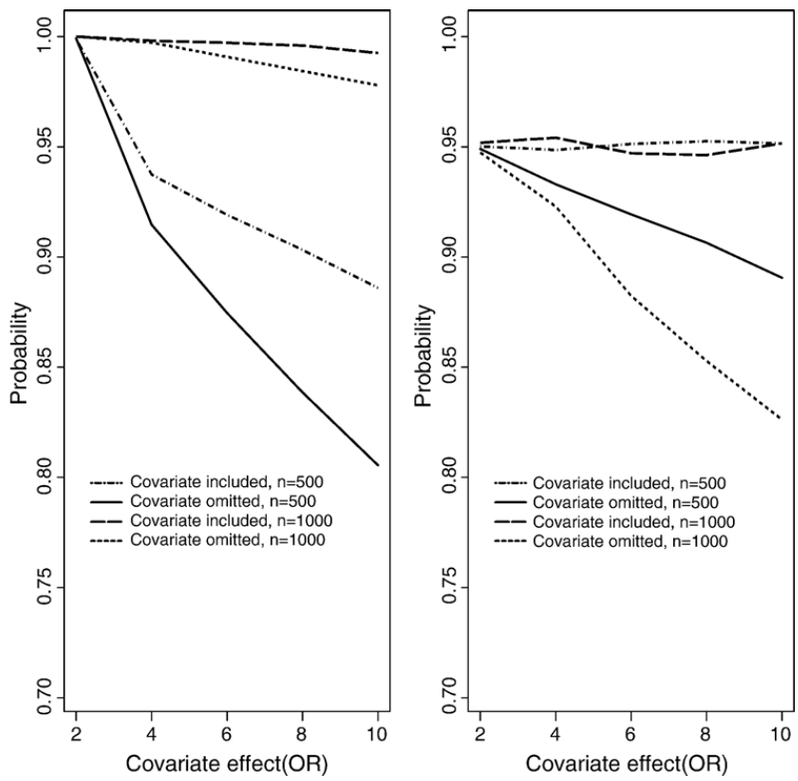

Fig. 2. Power (left panel) and coverage probability (right panel): as a function of the strength of a binary omitted covariate and sample size. The exposure effect was fixed at OR=2 (see text for details).

For each scenario and model investigated, we retained the median of the 10,000 regression coefficients (i.e., log odds ratios). The relative bias was calculated as $|\widetilde{\beta}_1^* - \widetilde{\beta}_1| / \widetilde{\beta}_1$. Where $\widetilde{\beta}_1^*$ is the median of the empirical distribution of estimated coefficients of exposure from the model omitting the covariate and $\widetilde{\beta}_1$ is the corresponding quantity from the model including the covariate. The coverage rate was computed as the percentage of estimated confidence intervals, based on normal approximation, containing the truth. The size of the test/power (based on Wald test) was computed as the proportion of simulated samples in which the hypothesis of no exposure effect was rejected at the 0.05 level.

4. Results

In the absence of an exposure effect, there was no bias involved in the estimation of the odds ratio due to omitted covariate. This has been shown analytically [2]. Moreover, under this scenario, the 95% confidence interval has the correct coverage probability and the size of the test was also correct. The empirical coverage rate was within the range of 0.948–0.953 while the empirical Type I error was within the range of 0.045–0.052 (data not shown).

When the effect of the binary omitted covariate on disease outcome does not exceed that of the exposure, the relative change-in-estimate of effect of exposure is less than 5% and there is practically no difference in the coverage probability and power between the models with and without the covariate (see Figs. 1 and 2). However, as the effect of the covariate increases relative to that of the exposure, the discrepancy between the results obtained from the two models also increases. Specifically, the relative change-in-estimate of effect of exposure increased from 2.9% to more than 20%, corresponding to odds ratios (for the covariate) of 2.0 and 10.0, respectively (see Fig. 1). In the case of a continuous omitted covariate, even when its effect per one unit increase is equal to that of exposure, the bias incurred was more than 10% and it increases dramatically with increasing magnitude of the omitted covariate's effect (see Fig. 1). As expected, the extent of the bias does not depend on sample size; in plotting Fig. 1 we jittered the points to minimize complete overlap between the two scenarios. The coverage rate of the 95% confidence interval derived from the crude analysis [i.e., model (2)] also decreased progressively as compared to the model that included the covariate (see Fig. 2).

Regarding study power, the difference remains relatively modest, within 19% when the covariate is binary; the model including the covariate always afforded better power (see Fig. 2). On the other hand, when a continuous covariate is omitted, study power dropped by as much as 58% (see Figs. 3 and 4). This observation is compatible with the general expectation that keeping a continuous covariate in the model with the correct functional form results in

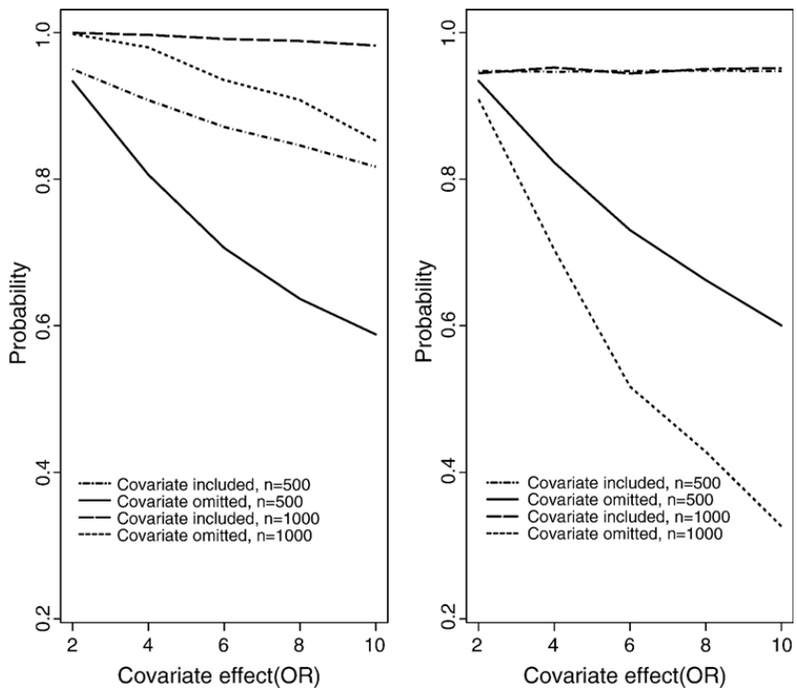

Fig. 3. Power (left panel) and coverage probability (right panel): as a function of the strength of a continuous omitted covariate $\sim N(0,1)$ and sample size. The exposure effect was fixed at OR = 2 (see text for details).

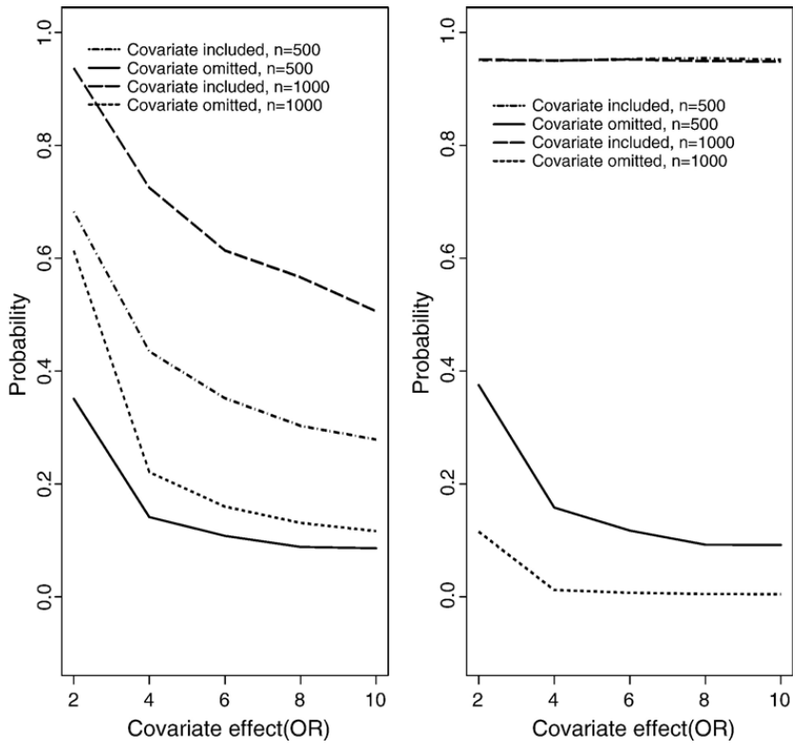

Fig. 4. Power (left panel) and coverage probability (right panel): as a function of the strength of a continuous omitted covariate $\sim N(0,5)$ and sample size. The exposure effect was fixed at OR = 2 (see text for details).

improved efficiency. The coverage of the confidence interval derived from the model with omitted covariate worsens as the sample size increases, while coverage becomes increasingly closer to the nominal level for the model including the covariate (see Figs. 2–4).

This deterioration in coverage, seemingly counterintuitive, could be explained by the fact that the estimate of exposure effect obtained from the model omitting the covariate has a relatively smaller standard error as compared to the one obtained from the extended model. Therefore, the corresponding 95% confidence interval tends to be tighter and shifted to the null (result not shown), i.e., the point estimate is off-target as a consequence the confidence intervals "shoot too low". A similar phenomenon was also previously noted in the context of the proportional hazards model [8]. In general, the difference between the two models becomes more dramatic when a strong continuous covariate is omitted (see Figs. 3 and 4).

For the major part of the simulation study, we considered a background disease risk of 40% among the unexposed at the lower level (mean in the case of a continuous covariate) of the omitted covariate; this is somewhat high compared to the background risk expected in many epidemiological studies. However, when we considered a lower background disease risk combined with a very small exposure probability, both set at 5%, then results from the two models are almost identical (data not shown). Here a possible explanation is that very low background disease risk and exposure probability, in combination with a not so strong omitted covariate, translates to smaller number of events that in turn leads to a rare disease scenario. Generally, when the background disease risk is very small (1%) the effect of a binary omitted covariate on bias, coverage probability and power is negligible (data not shown). However, in the case of a continuous omitted covariate, we observed a similar result when the variability of the omitted covariate is relatively small.

5. Example: sexual activity and longevity of male fruit flies

We give a real-life example using an interesting experimental data set that appeared previously in the literature [10,11]. The design of the experiment has been described in detail elsewhere [10]. In short, 125 fruit flies were randomly divided into five groups of 25 to determine whether increased reproduction reduces the longevity of male flies. This effect is known to occur in female flies. Sexual activity of individual males was manipulated by providing

Table 2
Odds ratios and 95% CI for effect of sexual activity on mortality derived from models omitting and including thorax length

	Thorax	
	Continuous	Binary[a]
Omitted	7.94 (1.88, 33.50)	7.94 (1.88, 33.50)
Included	46.99 (3.50, 631.6)	15.30 (2.45, 95.45)

[a] Thorax is categorized as above/below median.

each male in the first group with eight new receptive females every 2 days and those in another group with only one. These will be considered the experimental arms and hereafter referred to as E8 and E1, respectively. In order to take into account the effect of competition for food or space due to the presence of female flies, two control arms (hereafter referred to as C8 and C1) were created. The control arms were provided with the same number of newly inseminated females every 2 days as the corresponding experimental arms. It is known that newly inseminated females will not mate again for at least 2 days. In this way, any difference in longevity between the experimental and control arms can be attributed to the increased sexual activity not to effects of crowding, competition for food, etc. Each of the male flies in the fifth group was kept alone; these 25 served as control for all other groups. All arms were treated in the same way in terms of provision of fresh food. Just for the sake of brevity, we will restrict ourselves to the analysis of the E8–C8 arms, i.e., the ones with eight partners. The main interest is assessing the effect of sexual activity, coded dichotomously indicating membership in either the E8 or C8 arms. The data set contains another covariate that is strongly associated with longevity, thorax length. To verify that randomisation worked reasonably, we compared the thorax length within the two arms. In the E8 arm thorax length has a mean of 0.80 mm (S.D. = 0.08) while in the C8 arm the corresponding figures are 0.81 mm (S.D. = 0.08). Therefore, the distribution of thorax length between these two arms looks quite balanced. So confounding by thorax length is not expected. All the fruit flies in the E8 and C8 arms were followed to death. However, in order to perform binary data analysis, we dichotomized longevity so as to produce a 30% overall early mortality rate. This was achieved by considering a cut-off point of 40 days. Hence, the outcome of interest is mortality within 40 days. This resulted in 3 deaths out of 25 in the C8 arm and 13 deaths out of 25 in the E8 arm. We carried out a logistic regression, modeling the odds of 40-days mortality. In the crude analysis the regression coefficient of sexual activity was 2.07 with 95% CI (0.63, 3.51), implying that sexual activity is detrimental for the longevity of male fruit flies. Results in terms of odds ratios are presented in Table 2.

In order to see if this estimate changes when thorax length is included in the logistic regression model along with sexual activity, we re-fitted the model including thorax length. The new regression coefficient of sexual activity was 3.85 with 95% CI (1.25, 6.45), i.e., adjusted for thorax length. This constitutes about 46.2% change in the regression coefficient of sexual activity. This is in line with our simulation result, i.e., substantial bias towards the null when a strong continuous variable is omitted from the model. Again, in order to contrast this result with the case of a dichotomous omitted covariate, we dichotomized thorax length at the median of the overall distribution, i.e., E8–C8 arms combined. The re-fitted model, including thorax length as dichotomous, gave a regression coefficient of 2.73 with 95% CI (0.90, 4.56). This constitutes about a 24.2% change in the regression coefficient of sexual activity, and it is less severe than the case of considering thorax as a continuous covariate. Once again, this concurs with our simulation result, i.e., omitting a strong continuous covariate leads to a substantial bias towards the null as compared to omitting a strong dichotomous covariate. In addition, the 95% CI based on the crude analysis is narrower and shifted to the null as compared to results from the adjusted analyses (see Table 2). This is also in agreement with the results of our simulation.

6. Discussion

When the variable of interest, i.e., exposure/intervention, has a non-null effect on disease risk the impact of an omitted but *balanced* covariate is to bias the odds ratio towards the null. This impact also extends to a shift of the corresponding 95% confidence interval to the null with a reduced coverage probability than the nominal level. In addition, study power will be reduced as compared to the model including the covariate. The impact is more dramatic when the effect of the omitted covariate on disease risk exceeds that of exposure/intervention. Moreover, for a given magnitude of effect, large variability of the omitted covariate is associated with increased reduction in coverage rate

and study power. While the extent of bias does not vary with study size, coverage and study power do. Specifically, the impact on coverage seems to increase with increasing sample size while the impact on study power, as would be expected, decreases with increasing sample size. Generally, the impact is more pronounced when the omitted covariate is continuous. Therefore, on the basis of our finding, we advise adjusting for strong prognostic factors during the analysis of binary data where the outcome is not considered rare. Various approaches have been described and employed in choosing covariates for adjustment [12,13]. Selection of covariates on the basis of statistical significance might not be appropriate in this context as pointed out previously in the literature [9,13,14]. However, we recommend covariate selection for adjustment on the basis of *a priori specified* cut-point for change-in-estimate of effect of exposure/intervention. Even though a 10% change-in-estimate is suggested in the literature [14], the cut-point might vary on the basis of the specific subject matter. The substantive investigator(s) might have a good sense as to what constitute a meaningful change-in-estimate. While a 5% change-in-estimate might be considered negligible in most contexts, an excess of 20% change-in-estimate would hardly be considered negligible. The important point is to clearly specify this cut-point in the analysis plan so that covariate adjustment will be carried out in a more principled manner at the data analysis stage. Such *a priori* specification in the analysis plan, at the time of designing the study, also helps to overcome any suspicion that *post hoc* selection of covariates might be based on subjective criteria. However, at the design stage investigators need to consider carefully important prognostic factors based on *a priori* substantive knowledge or relevant previous work. The practical implication of our finding is that identification of important prognostic factors becomes more of a study design issue rather than data analysis issue especially when effect estimation is the primary interest [9], which is usually the case.

References

[1] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regression and omitted covariates. Biometrika 1984;71:431–44.
[2] Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol 1991;44:77–81.
[3] Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 1996;7:498–501.
[4] Chao W-H, Palta M, Young T. Effect of omitted confounders on the analysis of correlated binary data. Biometrics 1997;53:678–89.
[5] Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 1988;7:257–74.
[6] Hanley AJ, Negassa A, Edwards DM, Forrester JE. Statistical analysis using Generalized Estimating Equations (GEE): an orientation. Am J Epidemiol 2003;157:364–75.
[7] Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol 1981;114:593–603.
[8] Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. Biometrika 1986;73:363–9.
[9] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials 1998;19:249–56.
[10] Partridge L, Farquhar M. Sexual activity and the life span of male fruit flies. Nature 1981;294:580–1.
[11] Hanley JA, Shapiro SH. Sexual activity and life span of male fruit flies: a data set that gets attention. J Stat Educ 1994;2(1).
[12] Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. Control Clin Trials 1989;10:161S–75S.
[13] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparison in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917–30.
[14] Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79:340–9.